The Korean Journal of Applied Statistics (2019) DOI: https://doi.org/10.5351/KJAS.2019.32.1.083
32(1), 83–97

Multiple imputation and synthetic data

Joungyoun Kim[a] · Min-Jeong Park[b,1]

[a]Department of Information & Statistics, Chungbuk National University;
[b]Statistical Research Institute, Statistics Korea

(Received November 26, 2018; Revised January 4, 2019; Accepted January 4, 2019)

Abstract

As society develops, the dissemination of microdata has increased to respond to diverse analytical needs of users. Analysis of microdata for policy making, academic purposes, etc. is highly desirable in terms of value creation. However, the provision of microdata, whose usefulness is guaranteed, has a risk of exposure of personal information. Several methods have been considered to ensure the protection of personal information while ensuring the usefulness of the data. One of these methods has been studied to generate and utilize synthetic data. This paper aims to understand the synthetic data by exploring methodologies and precautions related to synthetic data. To this end, we first explain muptiple imputation, Bayesian predictive model, and Bayesian bootstrap, which are basic foundations for synthetic data. And then, we link these concepts to the construction of fully/partially synthetic data. To understand the creation of synthetic data, we review a real longitudinal synthetic data example which is based on sequential regression multivariate imputation.

Keywords: synthetic data, multiple imputation, Bayesian prediction model, Bayesian bootstrap, microdata

1. 서론

사회가 발전함에 따라 이용자의 심도있는 자료 분석 요구를 뒷받침하기 위해 개인 단위로 구성된 마이크로데이터(microdata) 제공이 증가해왔다. 또한 조사자료를 넘어서, 센서스나 행정자료와 같은 전수자료를 마이크로데이터 형태로 제공받아 연구하고자 하는 수요 역시 늘어나고 있다. 이러한 마이크로데이터 분석은 정책 결정이나 학술 분야의 가치 창출 측면 등에서 대단히 바람직하다. 그러나 이를 지원하기 위해 자료 유용성을 충분히 확보해서 마이크로데이터를 제공하게 되면, 개인정보가 노출될 가능성이라는 위험이 발생하게 된다. 이에, 자료의 유용성은 확보하면서도 개인정보는 보호하기 위해 통계적 노출제어(statistical disclosure control) 분야의 여러 방법들이 고려되어 왔다 (Park, 2016; Park과 Kim, 2016).

본 연구에서는 이러한 방법 중 하나로 재현자료(synthetic data)의 생성 및 활용에 대해 살펴보고자 한다. 최초 제안에 따르면 재현자료란 실제 관측값이 아니라, 관측값을 기반으로 생성된 자료를 의미한다.

Joungyoun Kim's research was supported by the research grant of the Chungbuk National University in 2016 and National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (NRF-2017R1C1B5015192).
[1]Corresponding author: Statistical Center, Daeduckdae-ro 317-gil 9, Seo-gu, Daejeon 35214, Korea.
E-mail: mjstat@korea.kr

또한 재현자료는 원 자료에 있는 개별 자료를 포함하고 있지는 않지만, 원 자료가 보여주는 변수들간의 관계를 온전히 보존하는 새로운 자료를 생성하는 것을 목표로 한다. 따라서 개념적으로 재현자료 관련 연구는 재현자료의 노출위험에 관한 논의는 배제한 상태에서 주로 재현자료의 유용성을 확보하는 문제를 통계적으로 해결하는데 집중하여 왔다고 할 수 있다.

재현자료를 최초로 제안한 것은 Rubin (1993) 및 Little (1993)이다. 먼저 Rubin (1993)은 다중대체(multiple imputation)의 (Rubin, 1978, 1987, 1988) 개념을 노출제어에 적용했다. 주요 아이디어는 모집단에서 관측되지 않은 자료를 결측값으로 해석하고, 다중대체 기법을 적용해 결측값들을 대체해서 몇 개의 재현 모집단을 구축하며, 다시 샘플링을 통해 재현자료 표본 세트들을 만드는 것이다. 한편 같은 저널에서 Little (1993)은 모든 개체에 대해 재현자료를 작성하는 것이 아니라, 일부 민감한 개별값에 대해서만 재현자료를 생성하는 부분(partially) 재현자료를 제안했다. 이후 효율적인 재현자료의 생성 및 다양한 자료형태에 대한 재현자료 작성 방법 개발을 위한 연구들이 활발히 진행되어 왔다 (Reiter, 2002, 2003, 2004; Raghunathan 등, 2001, 2003).

본 논문에서는 Park과 Kim (2017)을 바탕으로 재현자료의 기본이 되는 다중대체에 대해 자세히 살펴본 뒤, 재현자료 생성과 관련된 방법론 및 주의사항을 소개하여 재현자료의 이해를 도모하고자 한다. 특히, 순차회귀 다중대체(sequential regression multivariate imputation; SRMI)를 활용하는 예제를 추가하여 재현자료 작성을 심도 깊이 이해하고자 한다. 이외에도, 비모수적으로 재현자료를 생성할 때 선호되는 베이지안 붓스트랩(Bayesian bootstrap)의 특징 등에 대해서도 자세히 살펴보고자 한다. 구체적인 사례로 경시적(longitudinal) 재현자료 작성에 관한 선행 연구를 다룬다.

본 논문의 구성은 다음과 같다. 2절에서는 다중대체와 베이지안 붓스트랩에 대해 살펴본다. 3절에서는 재현자료 생성 과정에 대해 소개한다. 4절에서는 경시적 자료에 대한 재현자료 작성 사례를 구체적으로 살펴본다. 마지막으로 재현자료를 사용하려고 할 때 추론을 위해 주의할 사항은 5절에서 정리한다.

2. 다중대체 및 베이지안 붓스트랩

재현자료의 작성은 Rubin (1978)이 제안한 다중대체 기법을 기반으로 한다. 한편 비모수적으로 재현자료를 생성할 때는 주로 베이지안 붓스트랩이 고려된다. 따라서 본격적으로 재현자료 생성을 이해하기 위해 다중대체와 베이지안 붓스트랩에 대해 먼저 살펴 볼 필요가 있다.

2.1. 다중대체

결측(missing)값이란 관측되지 않은 자료값으로, 결측값이 생성되는 원인에 따라 세 종류로 나뉜다. (1) 결측이 완전히 우연에 의해 생긴 경우, 완전임의결측(missing completely at random; MCAR), (2) 결측의 원인이 결측 자체와는 상관이 없는 경우, 임의결측(missing at random; MAR), (3) 결측값과 결측 여부 간에 연관이 있는 경우에는 비임의결측(missing not at random; MNAR)이라 한다. 결측값 처리를 위한 가장 손쉬운 방법은 결측이 발생한 행 전체를 삭제(listwise deletion)하는 것이지만, 비임의결측인 경우에 이 방법을 적용하면 분석 결과가 편향될(biased) 가능성이 커진다. 또한 완전임의결측의 경우에도 표본 크기의 감소로 인한 검정력 저하가 우려된다. 이를 보완하고자 결측값을 적절한 값으로 바꾸는 대체 방법들이 고려되어 왔고, 이 방법들은 크게 단순대체와 다중대체로 나눌 수 있다.

단순대체란 각 결측값에 하나의 대체값을 제공하는 것을 말한다. 예를 들어 연속형 변수의 경우 대체값으로 산술평균 또는 회귀 예측값 등을 사용할 수 있다. 한편, 종단자료의 경우라면 마지막 관측값으로(last observation carried forward) 결측값을 대체할 수도 있다. 이 외에도 다른 자료에서 비슷한 값을 가져와(cold deck) 대체값으로 이용하기도 한다. 단순대체의 최대 장점으로는 대체값 결정 후 기존

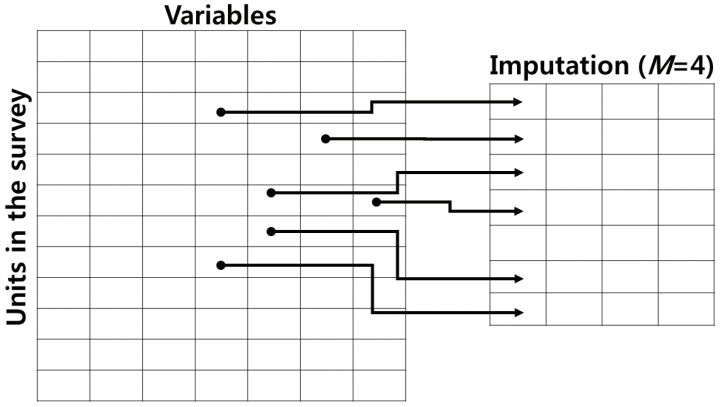

Figure 2.1. An example of multiple imputation.

의 분석 방법을 그대로 적용함으로써 분석의 일관성을 유지할 수 있다는 것을 들 수 있다. 하지만 단순대체는 결측 때문에 발생하는 불확실성을 고려하지는 않으므로 추정량의 분산이 과소 추정되는 문제점을 가지게 된다. 이러한 문제의식에서 제안된 방법이 다중대체이다.

Rubin (1978)이 제안한 다중대체의 핵심은 앞에서 언급한 결측으로 인한 불확실성을 고려하기 위해 하나의 결측값에 대해 여러 개의 대체값을 제공한다는 것이다. 대체값으로 여러 개의 값을 고려하기 때문에, 모수적/비모수적 또는 선형적/비선형적 등과 같은 다양한 시나리오에서 대체값을 생성할 수 있으며, 결측값이 (완전)임의결측에 의한 경우뿐만 아니라 비임의 결측인 경우에도 적용 가능하다. 또한, 단순대체와 마찬가지로 결측값이 대체된 이후에 이 대체 자료에 대한 새로운 분석방법이 요구되는 것이 아니라, 일반적인 통계분석 방법들을 동일하게 적용될 수 있다는 장점이 있다. Figure 2.1은 Rubin (1988)에서 사용된 예로, 각 결측값에 대해 $M(=4)$개의 대체값을 제공하는 것을 묘사하고 있다. Rubin (1987)은 다중대체의 과정을 다음과 같이 세 단계로 제안한다.

(1) 대체 단계: 각 결측값에 대해 적절한 대체 모형을 선택하여 결측된 자료의 예측분포를 구한다. 이 예측분포로부터 각 결측값에 대한 대체값들을 M번 랜덤하게 생성한다. 대체값들로 채워진 자료를 각각 $D^{(1)}, D^{(2)}, \ldots, D^{(M)}$이라 하자.

(2) 분석 단계: 각 자료 $D^{(m)}$에 대해 원하는 통계분석을 시행한다. 각 자료로부터 얻어진 관심 모수 Q에 대한 점추정값을 q_m이라 하고, 분산 추정량을 v_m이라 하자 $(m = 1, \ldots, M)$.

(3) 결합 단계: M개의 추정값들을 다음과 같이 결합하여 최종 추정량(Q_M) 및 분산(T_M)을 구한다.

$$Q_M = \sum_{m=1}^{M} \frac{q_m}{M}, \qquad \bar{v}_M = \sum_{m=1}^{M} \frac{v_m}{M}.$$

$$b_M = \sum_{m=1}^{M} \frac{(q_m - Q_M)^2}{M-1}, \qquad T_M = \frac{b_M}{M} + b_M + \bar{v}_M.$$

Rubin (1987)은 Q_M은 Q의 불편 추정량이며 정규분포로 근사한다는 것을 증명하여, 다중대체된 자료들에 (3)의 결합규칙을 적용하는 방식의 통계분석에 대한 타당성을 확보하였다.

다중대체의 대체값 생성은 사후예측분포(posterior predictive distribution)를 바탕으로 한다. 먼저, 모집단의 특성을 나타내는 모수를 θ, 모집단의 크기를 N, 자료에서 결측값이 없는 변수들을 $X(N \times P)$,

결측값이 있는 변수들을 $Y(N \times K)$라 하자. Y 중에서 결측값이 없는 부분을 Y_{obs}, 결측값이 있는 부분을 Y_{mis}라 하자. 즉, $Y = \{Y_{\text{obs}}, Y_{\text{mis}}\}$가 된다. 또한, X 중에서 Y_{obs}에 대응하는 부분을 X_{obs}, Y_{mis}에 대응하는 부분을 X_{mis}라 하자. 그러면 $X = \{X_{\text{obs}}, X_{\text{mis}}\}$라고 표현된다. 이제 주어진 자료 (X, Y_{obs})를 바탕으로 Y_{mis}를 추정하기 위한 사후예측분포는 다음과 같다.

$$P(Y_{\text{mis}} \,|\, Y_{\text{obs}},\, X) = \int p(Y_{\text{mis}} \,|\, X, Y_{\text{obs}}, \theta) p(\theta \,|\, X, Y_{\text{obs}}) d\theta.$$

이때 $p(\theta \,|\, X, Y_{\text{obs}})$는 (X, Y_{obs})가 주어졌을 때, θ의 사후분포이다. 일반적으로 θ의 사전분포가 정해지면 사후분포 계산이 가능하거나 사후분포로부터의 표본생성이 용이한 편이다. 한편 $p(Y_{\text{mis}} \,|\, X, Y_{\text{obs}}, \theta)$는 $(X, Y_{\text{obs}}, \theta)$가 주어졌을 때 Y_{mis}의 조건부 분포로, 모수 θ가 주어졌을 때 X와 Y간의 관계를 파악하는 것이 대체 모형 구축의 핵심이다. 하지만, (X, Y)가 다차원일 경우 변수들간의 관계를 파악하는 것은 결코 쉽지 않다. 다음절에서 조건부 분포 $p(Y_{\text{mis}} \,|\, X, Y_{\text{obs}}, \theta)$에서의 표본추출 방법을 구체적으로 설명하겠다.

2.1.1. 순차회귀 다중대체 Y_{mis}가 K개의 변수들로 구성되었다고 할 때, $Y_{\text{mis}} = \{y_1, y_2, \ldots, y_K\}$라 하자. 조건부 분포 $p(Y_{\text{mis}} \,|\, X, Y_{\text{obs}}, \theta)$를 K개 변수의 조건부 결합확률분포로 표현하면 다음과 같다.

$$p(Y_{\text{mis}} \,|\, X, Y_{\text{obs}}, \theta) = p(y_1, \ldots, y_K \,|\, X, Y_{\text{obs}}, \theta).$$

다중대체의 핵심은 $(X, Y_{\text{obs}}, \theta)$가 주어졌을 때, $y_1, \ldots, y_K$에 해당하는 값을 생성하는 것이며, 이를 크게 두 가지 방법으로 나눌 수 있다. 첫 번째는 결합확률분포(join probability distribution)를 추정한 뒤 K개의 변수에 대한 대체값을 동시에(jointly) 생성하는 방식으로, 실제로 적용할 때는 식의 유도 및 계산이 불가능한 경우가 대부분이다. 이에 비해, 순차회귀 다중대체는 한 변수씩 차례로 대체값을 생성하는 방법으로, 다음과 같은 조건부 모형을 사용한다.

$$\begin{aligned} p(y_1, \ldots, y_K \,|\, X, Y_{\text{obs}}, \theta) = {} & p_1(y_1 \,|\, X, Y_{\text{obs}}, \theta) \\ & \times p_2(y_2 \,|\, X, Y_{\text{obs}}, \theta, y_1) \\ & \vdots \\ & \times p_K(y_K \,|\, X, Y_{\text{obs}}, \theta, y_1, \ldots, y_{K-1}). \end{aligned}$$

이와 같이 SRMI는 각 조건부 모형으로부터 한 변수씩 순차적으로 대체값을 생성하기 때문에 결합확률분포에 비해 상대적으로 실제 구현이 용이하다.

SRMI의 대체 절차는 L개 라운드로 구성될 수 있다. 먼저 위의 식으로부터 구해진 대체값들을 $y_1^{(1)}$, $\ldots, y_K^{(l)}$이라 하자. 그러면 다음 번인 $(l+1)$번째 $(l = 1, \ldots, L-1)$에서는 대체값을 아래와 같은 조건부 분포로부터 생성한다.

$$f_k \left(y_k^{(l+1)} \,|\, X, Y_{\text{obs}}, \theta, y_1^{l+1}, \ldots, y_{k-1}^{l+1}, y_{k+1}^{l}, \ldots y_K^{l} \right).$$

이는 대체값 생성에 있어 가능한 최신값을 활용한다는 의미를 갖고 있으며 Gibbs sampling의 근사로 간주될 수 있다.

2.1.2. Example 본 절에서는 Raghunathan 등 (2001)에 소개되었던 모의실험을 바탕으로 하여 SRMI에 대한 이해를 높이고 그 성능에 대해 살펴보고자 한다. 데이터가 세 개의 변수 (U, Y_1, Y_2)로 구성되어 있다고 하고, 다음과 같이 모의실험을 위한 데이터를 생성한다.

(1) $U \sim \text{Normal}(0, 1)$

(2) $Y_1 \sim$ Gamma with mean $= \mu_1 = \exp(U - 1)$ and variance $= \mu_1^2/5$

(3) $Y_2 \sim$ Gamma with mean $= \mu_2 = \exp(-1 + 0.5U + 0.5Y_1)$ and variance $= \mu_2^2/5$

이와 같이 생성된 자료의 일부를 결측값으로 만들기 위해 다음을 가정한다.

(1) 변수 U에는 결측값이 전혀 없음

(2) Y_1의 결측 여부는 U에 의존하는 로지스틱 모형 $\text{logit}(\Pr(Y_1 \text{ is observed})) = 1.5 + U$에 따라 랜덤하게 결정됨

(3) Y_2의 결측 여부는 U와 Y_1에 의존하는 로지스틱 모형 $\text{logit}(\Pr(Y_2 \text{ is observed})) = 1.5 - 0.5Y_1 - 0.5U$에 따라 랜덤하게 결정됨

위의 과정을 수행한 결과 Y_1의 결측률은 약 22%, Y_2의 결측률은 약 48%이다.

예측모형으로 선형회귀분석(linear regression model)과 Box-Cox 변환(Box-Cox transformation)만 고려하여 다음과 같이 SRMI를 수행한다.

(1) $Z_1 = (Y_1^{\lambda_1} - 1)/\lambda_1 \sim N(a_0 + a_1 U, \sigma_1^2)$

(2) $Z_2 = (Y_2^{\lambda_1} - 1)/\lambda_1 \sim N(b_0 + b_1 U + b_2 Y_1, \sigma_2^2)$

단, λ_1과 λ_2는 각 단계에서 최대우도추정량으로 추정한다.

크기가 $n(= 100)$인 데이터 세트를 2500개 생산한 뒤, 각 데이터에 대해서 위에서 설명한 바와 같이 결측값을 랜덤하게 정하고 $M(= 5)$개씩 대체 데이터 세트를 생성한다. 각 대체 데이터에 회귀모형 $\log(Y_2) \sim N(\beta_0 + \beta_1 U + \beta_2 Y_1, \sigma^2)$을 적합한 뒤, 얻어진 회귀 계수별로 평균을 구한다. Table 2.1은 2500개 중 처음 10개 데이터에 대해서 생성된 자료에 대한 추정결과이다. q_c는 전체 데이터를 이용했을 때 추정된 회귀 계수를, q_i는 i번째 대체 데이터에서 얻어진 회귀 계수, v_i는 q_i의 표준오차를 나타낸다. Q_M, $\bar{v}_M$은 각각 q_i와 v_i의 평균이다. Figure 2.2는 대체 데이터에서 구한 회귀 계수들의 평균과 원자료 적합으로부터 얻어진 회귀 계수간의 산점도와 상관계수를 보여주고 있다. 각 패널 안의 실선은 기울기가 1이고 원점을 지나는 직선이다. 대체로 점들이 대각선 근처에 몰려 있다. 다중대체 결과 얻어진 추정값과 전체 데이터를 사용하여 얻어진 추정값이 상당하 유사하다고 할 수 있다.

2.2. 베이지안 붓스트랩

베이지안 붓스트랩 (Rubin, 1981)은 비모수적으로 재현자료를 생성할 때 주로 사용하는 방법이다. 일반적으로 붓스트랩 (Efron, 1979)은 크기가 n인 표본 $\{x_1, \ldots, x_n\}$의 각 관측값에 $1/n$의 동일한 가중치(weights)를 부여한 뒤 복원추출을 시행하여 크기가 n인 표본을 생성하는 것을 기본으로 한다. 이에 반해 베이지안 붓스트랩은 각 관측값에 부여되는 가중치 w_i $(i = 1, \ldots, n)$에 대해 사전분포로 Dirichlet $(1, \ldots, 1)$를 가정하는 것이며, 구현 절차는 다음과 같다.

(1) 균일분포 Uniform$(0, 1)$에서 $(n - 1)$개의 난수를 발생 한 뒤, 다음과 같이 오름차순으로 정렬한다.

$$a_1, a_2, \ldots, a_{n-1}$$

(2) 여기에 $a_0 = 0$, $a_n = 1$을 추가한다.

(3) 균일분포 Uniform(0, 1)로부터 n개의 난수 $u_1, \ldots, u_n$을 생성한다.

Table 2.1. Sequential regression multivariate imputation simulation results

	No	q_c	q_1	q_2	q_3	q_4	q_5	Q_M	v_1	v_2	v_3	v_4	v_5	$\bar{v}_M$
β_0	1	−1.20	−1.17	−1.14	−1.20	−1.17	−1.16	−1.17	0.10	0.09	0.09	0.09	0.10	0.09
	2	−1.28	−1.24	−1.34	−1.28	−1.39	−1.21	−1.29	0.11	0.10	0.11	0.10	0.11	0.11
	3	−1.05	−0.92	−1.04	−0.81	−0.98	−0.98	−0.95	0.12	0.13	0.12	0.12	0.13	0.12
	4	−1.31	−1.56	−1.63	−1.58	−1.23	−1.62	−1.52	0.19	0.24	0.20	0.22	0.22	0.21
	5	−1.09	−1.03	−1.07	−1.08	−1.23	−1.07	−1.10	0.13	0.13	0.13	0.14	0.13	0.13
	6	−1.28	−1.23	−1.13	−1.18	−1.23	−1.29	−1.21	0.11	0.11	0.10	0.10	0.11	0.11
	7	−1.16	−1.39	−1.11	−1.19	−1.22	−1.57	−1.30	0.16	0.15	0.16	0.16	0.17	0.16
	8	−1.28	−1.34	−1.06	−1.02	−1.22	−1.12	−1.15	0.11	0.11	0.08	0.12	0.11	0.11
	9	−1.28	−1.10	−1.28	−1.22	−1.21	−1.26	−1.21	0.13	0.13	0.13	0.14	0.14	0.14
	10	−1.34	−1.42	−1.30	−1.31	−1.41	−1.51	−1.39	0.14	0.14	0.14	0.16	0.14	0.15
β_1	1	0.61	0.66	0.66	0.63	0.64	0.65	0.65	0.11	0.10	0.10	0.10	0.11	0.10
	2	0.43	0.51	0.42	0.51	0.43	0.53	0.48	0.10	0.09	0.10	0.09	0.09	0.09
	3	0.56	0.67	0.68	0.79	0.66	0.73	0.70	0.11	0.13	0.12	0.12	0.13	0.12
	4	0.58	0.39	0.29	0.39	0.63	0.29	0.40	0.17	0.20	0.18	0.19	0.19	0.19
	5	0.65	0.77	0.69	0.72	0.59	0.65	0.68	0.12	0.12	0.12	0.13	0.13	0.12
	6	0.54	0.57	0.59	0.49	0.58	0.48	0.54	0.10	0.10	0.09	0.10	0.10	0.10
	7	0.70	0.54	0.71	0.60	0.60	0.52	0.59	0.12	0.12	0.12	0.12	0.13	0.12
	8	0.46	0.42	0.65	0.69	0.58	0.57	0.58	0.11	0.11	0.09	0.12	0.11	0.11
	9	0.66	0.75	0.71	0.62	0.76	0.64	0.70	0.13	0.13	0.13	0.13	0.13	0.13
	10	0.46	0.50	0.79	0.69	1.09	0.39	0.69	0.14	0.14	0.13	0.15	0.13	0.14
β_2	1	0.47	0.53	0.49	0.58	0.61	0.54	0.55	0.07	0.07	0.07	0.07	0.08	0.07
	2	0.53	0.40	0.57	0.55	0.56	0.52	0.52	0.09	0.09	0.10	0.09	0.09	0.09
	3	0.42	0.39	0.40	0.20	0.45	0.43	0.37	0.13	0.14	0.10	0.13	0.14	0.13
	4	0.58	0.85	1.08	1.01	0.40	1.13	0.89	0.30	0.43	0.34	0.39	0.38	0.37
	5	0.29	0.23	0.28	0.40	0.54	0.32	0.35	0.21	0.20	0.20	0.21	0.21	0.20
	6	0.51	0.47	0.42	0.49	0.45	0.52	0.47	0.11	0.11	0.10	0.10	0.11	0.10
	7	0.19	0.48	0.05	0.35	0.32	0.81	0.40	0.25	0.24	0.26	0.27	0.28	0.26
	8	0.47	0.55	0.06	0.03	0.21	0.24	0.22	0.16	0.17	0.06	0.18	0.16	0.14
	9	0.28	0.08	0.21	0.18	0.13	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.15
	10	0.54	0.48	0.19	0.37	0.19	0.60	0.37	0.19	0.20	0.19	0.23	0.20	0.20

(4) 위의 난수들을 바탕으로 다음과 같이 크기 n인 붓스트랩 표본을 구한다

$$\{x_k^* \,|\, x_k^* = x_j, a_{j-1} < u_k \le a_j,\ k = 1, \ldots, n\}$$

(5) (1)–(4)과정을 M번 반복하여 재표집 표본을 M개 세트 얻는다.

관심모수 Q에 대해, 베이지안 붓스트랩을 기반으로 생성된 재표집 표본에서 얻는 점추정량을 q_m $(m = 1, \ldots, M)$이라 하고, 이들이 결합된 추정량을 $Q_{\mathrm{BB}} = \sum_{m=1}^{M} q_m/M$이라 하자.

Rubin (1981)은 이 때 사용되는 가중치 $f_i = u_{i+1} - u_i$의 기대값이 $1/n$이고 일반적인 Bootstrap에서 사용되는 가중치 g_i에 비해 분산이 작거나 같음을 증명하였다.

$$\mathrm{Var}(g_i) = \mathrm{Var}(f_i)\frac{n+1}{n} \ge \mathrm{Var}(f_i).$$

따라서 Q_{BB}의 변동이 일반적인 Bootsrap에서 얻어진 추정량보다 더 작아지는 경향이 있다 (Clyde와 Lee, 2001).

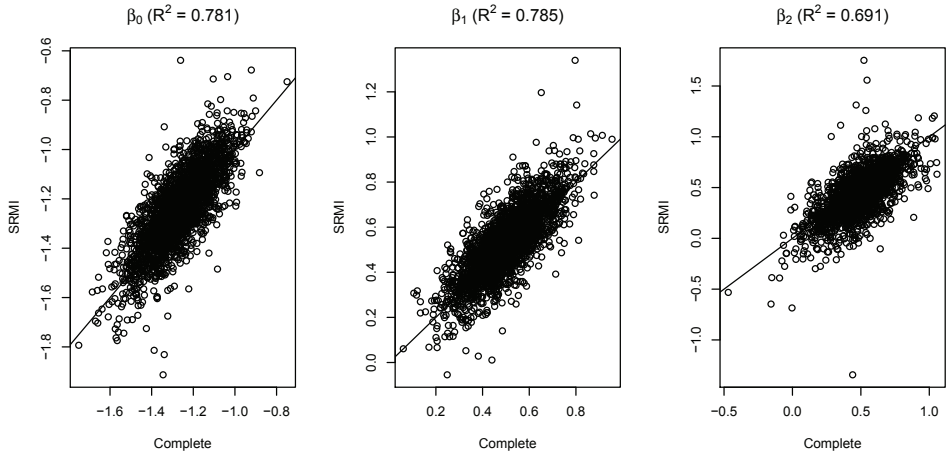

Figure 2.2. Scatter plots of the coefficient estimates from the analysis of the complete data (x-axis) and the mean of the coefficient estimates from the analysis of multiple imputation data (y-axis).

3. 재현자료 작성

3.1. 완전 재현자료

완전(fully) 재현자료란 Rubin (1993)이 정보보호를 위해 최초로 정의한 개념으로 다중대체 기법을 기반으로 하고 있다. 즉, (1) 표본틀(모집단)에서 조사되지 않거나 민감정보여서 공개될 수 없는 값들을 결측값으로 취급하여 다중대체하고, (2) 대체되어 채워진 재현 모집단에서 단순랜덤추출(simple random sampling)로 표본을 추출하여 제공하는 두 단계로 이루어진다. Rubin (1993)에 따르면, 제공되는 재현자료는 실제 데이터를 포함하지 않으므로 민감한 정보가 노출되기 대단히 어렵다. 또한 대체 모형을 잘 선택하여 예측력을 높인다면 원래 자료에 담긴 정보 대부분이 재현자료에 보존될 수 있다고 보았다. 이를 통해 기존 매스킹 기법들이 가지는 한계를 극복하고, 다음과 같은 자료제공에 있어 바람직한 세 조건을 만족시킬 수 있는 방향으로서 재현자료를 제시했다. 즉 (1) 개체별 정보 노출 방지 등 안전성 확보, (2) 타당한 추론을 가능하게 하는 적은 정보손실, (3) 일반적인 통계기법을 쉽게 적용할 수 있다는 의미에서 이용자의 편의성이 그 세 조건에 해당한다.

Rubin (1987)의 다중대체와 유사하게 완전 재현자료 작성을 위한 수식 표현을 정리하면 다음과 같다.

(1) $X(N \times P)$ = 전수조사가 가능한 행정자료 등으로 노출되어도 문제가 없다고 판단한 안전한 자료

(2) $Y(N \times K)$ = 표본에 대한 설문조사로 얻어지는 자료들 중에서 개별정보가 노출되어서는 안되는 민감자료

(3) $Y_{\text{obs}} = Y$ 중 설문조사에서 포함된(included) 개체의 관찰 자료

(4) $Y_{\text{mis}} = Y$ 중 설문조사에서 제외된(excluded) 개체의 미관찰 자료

(5) $X_{\text{obs}} = X$ 중 Y_{obs}에 대응되는 자료

(6) $X_{\text{mis}} = X$ 중 Y_{mis}에 대응되는 자료.

이후부터 재현자료 작성에 초점을 맞추어 수식 표현들로 $Y_{\text{inc}}(= Y_{\text{obs}})$, $Y_{\text{exc}}(= Y_{\text{mis}})$, $X_{\text{inc}}(= X_{\text{obs}})$, $X_{\text{exc}}(= X_{\text{mis}})$을 사용하도록 하겠다.

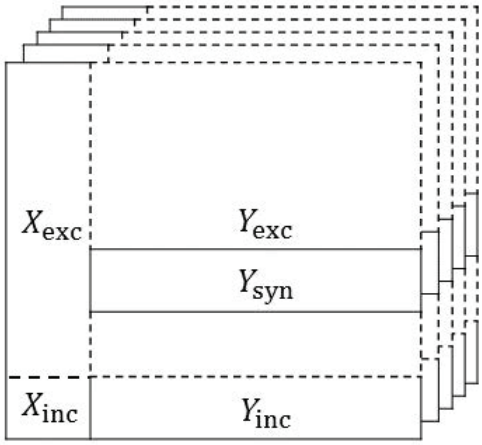

Figure 3.1. An example of generating fully synthetic data.

Figure 3.1은 완전 재현자료의 구조를 보여주는 예시이다. 완전 재현자료는 설문에 관찰되지 않는 자료 Y_{exc}를 결측값으로 간주하고, 다음의 사후예측분포로부터 Y_{exc}를 생성한다.

$$P(Y_{\text{exc}} \,|\, Y_{\text{inc}}, X) = \int p(Y_{\text{exc}} \,|\, X, Y_{\text{inc}}, \theta) p(\theta \,|\, X, Y_{\text{inc}}) d\theta.$$

대체값을 생성한 뒤 X_{exc}와 결합한 자료 $(X_{\text{exc}}, Y_{\text{exc}})$에서 랜덤추출한 세트를 제공한다. 단, 다중대체의 이론과 일관되게 하나의 세트를 제공하는 것이 아니라 다음과 같이 여러개의 데이터 세트를 제공한다.

$$\left(X^1_{\text{exc}}, Y^1_{\text{exc}}\right), \ldots, \left(X^M_{\text{exc}}, Y^M_{\text{exc}}\right).$$

이때 관심모수 Q에 대한 추론은 아래와 같은 불편추정량 Q_M과 분산 추정량 T_f을 이용한 정규근사를 바탕으로 한다 (Reiter, 2002; Raghunathan, 2003).

$$Q_M = \sum_{m=1}^{M} \frac{q_m}{M}, \qquad \bar{v}_M = \sum_{m=1}^{M} \frac{v_m}{M}.$$
$$b_M = \sum_{m=1}^{M} \frac{(q_m - Q_M)^2}{M-1}, \qquad T_f = b_M \left(1 + \frac{1}{M}\right) - \bar{v}_M.$$

단, 이때 분산 추정량 T_f이 음(negative)의 값을 가질 수 있다는 단점이 있다.

3.2. 부분 재현자료

앞에서 살펴본 완전 재현자료는 민감변수의 모든 값에 대해 재현자료를 작성하기 때문에, 모집단에 가정하는 모형에 매우 의존적이고, 모형이 잘못 가정된 경우 분석결과 얻어진 통계량의 편의(bias)가 매우 클 수 있다는 문제를 가진다. 이에 Little (1993)은 노출제어가 필요한 일부 변수, 더 나아가 그 중에서도 일부 개체에 대해서만 다중대체 할 것을 제안하고 이는 부분 재현자료 작성 방법론으로 발전한다. 부분 재현자료는 노출위험이 높은 변수 및 개체에 대해서만 값을 대체하기 때문에 대체되지 않은 자료에 대해서는 자료 고유의 성질이 대부분 유지되어 정보손실을 줄일 수 있다는 장점이 있다. 부분 재현자료

의 작성방법은 재현자료의 작성과 비슷하지만, 그 결합 규칙은 아래와 같다. 관심모수 Q에 대한 추론은 아래와 같은 불편추정량 Q_M과 분산 추정량 T_p을 이용한 정규근사를 바탕으로 한다 (Reiter, 2004).

$$Q_M = \sum_{m=1}^{M} \frac{q_m}{M}, \qquad \bar{v}_M = \sum_{m=1}^{M} \frac{v_m}{M}.$$
$$b_M = \sum_{m=1}^{M} \frac{(q_m - Q_M)^2}{M-1}, \qquad T_p = \frac{b_M}{M} + \bar{v}_M.$$

완전 재현자료와 달리, 분산 추정량 T_p가 항상 양(positive)의 값을 가지게 된다.

3.3. 결측값이 존재하는 경우 재현자료 작성

결측값이 있는 경우 재현자료의 작성은 기본적으로 다음과 같다.

(1) 먼저, 다중대체를 통해 R개의 데이터 세트 생성한다.

(2) (1)에서 생성된 각 대체 데이터세트에 대한 M개의 재현자료를 생성한다. 이 때, $D_m^{(r)}$ $(r = 1, \ldots, R,\ m = 1, \ldots, M)$은 r번째 대체 데이터에서 생성된 m번째 재현자료라고 하자.

(3) 이와 같이 생성된 총 $L = R \times M$개의 세트에 대해 분석을 실시한다. 각 분석에서 얻어진 관심 모수 Q에 대한 점 추정량과 분산 추정량을 q_m^r, v_m^r이라 할 때, 결합규칙은 다음과 같다.

$$Q_{MR} = \sum_{r=1}^{R}\sum_{m=1}^{M} \frac{q_m^{(r)}}{(MR)} = \sum_{r=1}^{R} \frac{\bar{q}^{(r)}}{R},$$
$$\bar{v}_{MR} = \sum_{r=1}^{R}\sum_{m=1}^{M} \frac{v_m^{(r)}}{(MR)},$$
$$b_{MR} = \sum_{r=1}^{R}\sum_{m=1}^{M} \frac{\left(q_m^{(r)} - \bar{q}^{(r)}\right)^2}{(R(M-1))},$$
$$B_{MR} = \sum_{r=1}^{R} \frac{\left(\bar{q}^{(r)} - Q_{MR}\right)^2}{(R-1)},$$
$$T_{MR} = \left(1 + \frac{1}{R}\right) B_{MR} - \frac{\bar{b}_{MR}}{M} + \bar{v}_{MR},$$

관심모수 Q에 대한 최종 추정량 Q_{MR}은 평균이 0이고 분산이 T_{MR}인 정규분포로 근사한다 (Reiter, 2003).

4. 재현자료 작성의 실제 사례

본 절에서는 재현자료 작성의 실제 구현에 대해 살펴보고자 한다. 사례로 살펴보는 프랑스 고용자료는 경시적 자료로, 개인 또는 기업에 대한 사항들을 여러 차례 측정하여 구축된다. 일반적으로 이러한 자료는 여러 기관에서 수집되어 연계되어 있으므로, 가능한 한 변수들간의 관계를 잘 보존하는 재현자료를 작성하는 것이 중요하다. 때문에 재현자료 작성의 대상으로 우선적으로 고려되었을 것으로 생각되나, 경시적 성질을 가지는 자료의 구조를 보존하여 재현자료를 생성하는 것이 대단히 어려움을 보여주기도 한다. 이 사례를 통해 재현자료 작성을 심도 깊게 이해할 수 있어 본 논문에서는 예제로 설명하도록 한다.

4.1. 프랑스 고용 자료의 구조

Abowd와 Woodcock (2001)에서 소개한 프랑스 고용 자료는 프랑스의 통계청(Instut National de la Statistique et des Etudes Economiques: INSEE)에서 구축한 개인의 고용기록과 해당 기업의 기록 등을 결합한 경시적 자료로 유용성이 높다고 판단된다. 때문에 Abowd 등 (1999)는 mixed effect model을 이용해 이 자료를 분석하기도 했다. 자료의 구성은 다음과 같다. 첫번째 자료는 인구통계자료(Echantillon Demographique Permanent; EDP)로 교육수준 등 개인에 대한 자세한 정보들을 포함하고 있다. 두번째 자료는 연간소득자료(Déclarations Annuelles des Données Sociales; DADS)로 매년 세금정산에 보고되는 수입을 기반으로 한 근무 기록 자료이다. 근로자 및 자영업자를 포함 세금 납세자 전체에 대한 정보를 포함하고 있다. 세번째 자료는 연간기업조사자료(Eneuête Annuelle d'Enterprises; EAE)로 기업에 대한 경시적 자료이다. 관측변수로는 산업 종류, 연 매출, 연 평균 고용수, 자본금등이 있으며, 이 중 연 매출, 연 평균 고용수, 자본금을 민감변수로 간주한다.

Abowd와 Woodcock (2001)에서 사용하는 자료는 다음과 같이 구축되었다. 먼저 1912년–1980년 사이의 짝수년 10월에 출생한 사람들에 대해 1976년부터 1996년까지의 DADS 자료를 수집하였다(단, 1981년, 1983년 및 1990년도는 행정적인 이유로 제외되었음). 다음으로 수집된 자료에서 20%를 랜덤 추출하였다. 단, 이들 중 절반은 EDP 자료와 결합하여 개인수준의 자세한 정보(예: 교육수준)를 사용하는 것이 가능한 반면, 나머지 반은 EDP의 개인 정보와 결합하지 않았다. 변수의 종류는 성별, 출생연도와 같은 변수(time-invariant)도 있고, 직업, 직장 위치, 연간 임금, full/part time 상태, 근무일수(paid days)와 같이 조사시점마다 관측되는 변수(time-variant)들도 있다. 이들 중 교육수준, 연간 임금, full/part time 상태, 근무일수를 민감변수로 간주한다. 최종적으로 개인 362,913명에 대한 3,213,374개의 근무 기록이 포함되어 있다. DAD의 최종 20% 표본과 일치(matching)되는 EAE 자료만 추출한 결과 총 105,813개 기업에 대한 470,812개의 기록을 사용한다. 연 고용 20인 이상 기업에 대해서만 설문이 이루어지기 때문에 제외된 기업에서 근무한 사람들의 근무 기록은 결측으로 간주한다.

4.2. 다중대체의 과정

프랑스 고용자료와 같은 경시적 자료의 재현자료 작성에 대한 효율적인 설명을 위해 다음과 같은 수식 표현을 사용하자. 먼저 EDP와 같은 개인에 대한 자세한 정보 자료를 U라 하고, EAE와 같은 기업의 경시적 자료는 Z, DADS와 같은 개인의 근무기록에 대한 경시적 자료는 W라 하자. 이때 포함된 총 개인의 수를 I, 기업의 수를 J라 하고, $j = J(i,t)$는 t 시점에 개인 i가 근무했던 기업을 가리킨다. 단, 개인과 시점의 조합에 해당하는 기업은 유일하다고 가정한다. 자료 U에 있는 변수 중 결측값의 개수가 k개이면서 대체값을 생성할 변수를 u_k라 하자. $U_{<k}$는 U에 있는 변수 중 결측값이 k개 미만인 변수들을, $U_{>k}$는 U에 있는 변수 중 결측 값이 k개 초과인 변수를 표현한다. 마찬가지로 W와 Z에 대해 $W_{<k}$, $W_{>k}$, $Z_{<k}$, $Z_{>k}$를 정의한다. i-번째 개인이 t-시점에 근무했던 기업 j에 대한 결측값 u_k, w_k, z_k에 대한 다중대체를 위해, 다음과 같은 조건부분포를 바탕으로 한 SRMI방법을 제안한다.

$$\int f_{u_k}\left(\begin{array}{c} u_k \,|\, U^{l+1}_{<k,i}, U^{l}_{>k,i}, U_{obs,i}, g_k\left(\left\{Z^{l+1}_{<k,J(i,t)}, Z^{l}_{>k,J(i,t)}, Z_{\mathrm{obs},J(i,t)}\right\}_{t=1}^{t=T_i}\right), \\ h_k\left(\left\{W^{l+1}_{<k,it}, W^{l}_{>k,it}, W_{\mathrm{obs},it}\right\}_{t=1}^{t=T_i}\right), \lambda_i, \theta_k \end{array}\right) p_k(\theta_k \,|\, \cdot) d\theta_k,$$

$$\int f_{w_k}\left(\begin{array}{c} w_k \,|\, U^{l+1}_{<k,i}, U^{l}_{>k,i}, U_{obs,i}, \left\{Z^{l+1}_{<k,J(i,\tau)}, Z^{l}_{>k,J(i,\tau)}, Z_{\mathrm{obs},J(i,\tau)}\right\}_{\tau=t-s}^{\tau=t+s}, \\ \left\{w^{l}_{k,i\tau}\right\}_{\tau=t-s,\tau\neq t}^{\tau=t+s}, \left\{W^{l+1}_{<k,i\tau}, W^{l}_{>k,i\tau}, W_{\mathrm{obs},i\tau}\right\}_{\tau=t-s}^{\tau=t+s}, \kappa_{it}, \mu_{it}, \theta_k \end{array}\right) p_k(\theta_k \,|\, \cdot) d\theta_k,$$

$$\int f_{z_k}\left(\begin{array}{c} z_k \mid m_k\left(U^{l+1}_{<k,J^{-1}(i,t)}, U^{l}_{>k,J^{-1}(i,t)}, U_{\mathrm{obs},J^{-1}(i,t)}\right), \\ \left\{z^{l}_{k,j\tau}\right\}^{\tau=t+s}_{\tau=t-s,\tau\neq t}, \left\{Z^{l+1}_{<k,j\tau}, Z^{l}_{>k,j\tau}, Z_{\mathrm{obs},j\tau}\right\}^{\tau=t+s}_{\tau=t-s}, \\ n_k\left(\left\{W^{l+1}_{<k,J^{-1}(i,t)\tau}, W^{l}_{>k,J^{-1}(i,t)\tau}, W_{\mathrm{obs},J^{-1}(i,t)\tau}\right\}^{\tau=t+s}_{\tau=t-s}\right), \gamma_{jt}, \nu_{jt}, \theta_k \end{array}\right) p_k(\theta_k \mid \cdot) d\theta_k.$$

이때 g, h, m, n은 각 대체값과 나머지 자료간의 연관성을 나타내는 것으로 각 개인별로 특화될 수 있다. λ_i, θ_k는 성별 및 full-time/part-time 상태의 조합으로 자료를 나누었을 때 각 subgroup을 나타내는 parameter들이다. $\{(U^m, W^m, Z^m) : \ m = 1, \ldots, M\}$는 결측값이 채워진 최종 M개의 완전 자료를 나타낸다.

4.2.1. 재현자료 작성 결과 본 절에서는 Abowd와 Woodcock (2001)에서 다중대체 및 재현자료 생성을 위해 실제 구현된 결과를 정리한다. 기본적으로 재현자료 작성은 다중대체가 완성된 이후에 시행되며, 특히 Abowd와 Woodcock (2001)는 다중대체나 재현자료 작성 모두 동일한 SRMI 과정으로 분석하였다. 따라서 본 절에서는 다중대체를 중심으로 설명한다.

먼저 결측값의 대체 순서는 결측값의 개수를 바탕으로 정한다. 즉, 결측값의 수가 적은 변수를 먼저 대체한다. 예를 들어 자본금, 연 매출액, 연평균 고용 건수 및 교육수준의 결측값 개수는 각각 (1) 35,989, (2) 47,796, (3) 150,833, (4) 약 180,000이다. 따라서 자본금이 가장 먼저 대체되고 교육수준이 가장 늦게 대체된다.

SRMI 예측모형으로 자본금, 연 매출액, 연평균 고용 건수와 같은 연속형 변수에 대해서는 선형회귀모형을 적용한다. 단, 이 값들이 모두 양수이고 분포가 치우쳐 있으므로 로그변환 후 분석한다. 반면 교육수준(범주 개수 8개)과 같은 범주형 변수의 경우 다항로짓 모형을 적용한다. 또한 예측모형으로 가능한 간단한 모형을 선호하여 변수선택 방법을 통해 각 변수별 가장 간단한 모형을 구축한다. time-variant 변수인 경우 mixed-effect model을 고려할 수 있다.

한편 이상값(outlier)의 생성원인으로는 (1) 관측오차(measurement error) 및 (2) 모집단의 이질성(heterogeneity)을 생각할 수 있다. 이때, 관측값이 예측값 ±5×SD 범위를 벗어난다면 이를 관측오차에 의한 outlier로 간주한다. 관측오차에 의한 이상값이라면 결측으로 간주하고 대체값으로 바꾼다. 그러면 노출위험에 대비하여 분석의 유용성을 크게 향상시킬 수 있다.

사후예측분포는 모수에도 의존하는 조건부 분포이므로, 모수를 랜덤추출로 고정한 뒤 대체값 (또는 재현값)을 생성한다. 따라서, 효율적인 모수의 표본추출은 재현자료 작성의 효율성에 영향을 미치는 중요한 요소이다. 가능한 모수가 사후분포의 모드(mode)에서 추출되는 것이 재현자료 작성에 효율적이기 때문에 importance sampling과 같은 방법을 적용해서 모수를 추출하도록 한다. 예를 들어 사후분포의 모드를 기준으로 ±3SD 범위 내에서 모수를 추출한다.

변수 근무일수의 경우 관측값의 범위가 $(1, 360)$이므로 다음과 같은 변환을 고려한다.

$$\log\left(\frac{\text{days paid}}{\text{365-days paid}}\right).$$

재현자료 작성의 대부분 과정은 위에서 설명한 다중대체 과정과 유사하다. 단, Abowd와 Woodcock (2001)는 결측값을 생성할 때는 L-라운드 ($L > 1$)를 거쳐 대체값을 생성하는데 반해, 재현자료 작성에서는 1-라운드로 값을 생성한다. 또한, 재현자료가 실제값에 가까운 값이 생성되도록 구간제약을 추가할 수 있다. 예를 들어 연속형 변수인 경우 재현된 값이 (실제값 ± 실제값의 20%) 구간 안에 포함되도록 한다. 100번의 추출 시도에도 구간에 포함되지 않을 경우 가까운 end-point값으로 결정한다.

5. 재현자료를 이용한 추론

완전 재현자료라는 용어를 사용할 때는 보통은 재현자료 내에 관측값이 단 하나도 존재하지 않는다고 생각하게 된다. 마찬가지로 부분 재현자료는 관측된 값에서 일부만이 재현된 자료라고 생각하게 된다. 이러한 해석은 각 방법이 만들어진 의도에 부합하는 것이지만, 반드시 그렇다고 단정해서는 안된다. 예를 들면 부분 재현자료 방법을 이용하지만 실제로는 완전 재현자료를 만들 수도 있기 때문이다. 그런 경우 어느 분산 추정식을 이용해서 추론하는 것이 정확한가 확인해 볼 필요가 있다. Drechsler (2018)에서는 이러한 문제를 포함해 재현자료를 사용할 때 혼란스러운 점들을 정리하고 모의실험을 통해 재현자료의 분산 추정식을 어떻게 사용할지 정리했다. 이 절에서는 Drechsler (2018)를 기반으로 완전 재현자료를 만들고 통계적 추론을 시행할 때 주의해야 할 사항을 소개한다.

먼저 완전 재현자료의 생성 및 분산 추정식을 다시 한번 간단히 정리해 보자. Drechsler (2018)에 설명된 완전 재현자료를 생성하는 Rubin의 방식은 다음과 같은 두 단계로 이루어진다. 단계1에서는 사후예측분포, $f(Y_{\text{exc}}|X, Y_{\text{inc}})$에서 Y_{exc}를 생성해 재현 모집단을 완성한다. 단계2에서는 재현 모집단 $(Y_{\text{inc}}, Y_{\text{exc}})$에서 단순랜덤추출 방식으로 원하는 크기의 표본 Y_{syn}을 추출해서 제공한다. 즉, 자료제공자는 사후예측분포를 구한 후, 재현 모집단 구축 및 샘플링 과정을 M회 반복하여 M개 표본을 제공한다. 이용자는 각 표본 세트에서 원하는 통계의 추정량 q_m 및 분산 추정량 v_m을 계산하고, 결합규칙을 통해 최종 추정량 q_M 및 T_f를 얻는다. 이제 이용자는 원자료를 계산해 얻는 통계 추정량을 대신해 q_M 및 T_f를 사용할 수 있다. Rubin 방식의 완전 재현자료 생성에서 가장 중요한 특징은 Y_{exc}만 재현되며 원래 관측값 Y_{inc}에는 변화가 없다는 것이다.

그러나 이러한 방식에 의해서는 완전 재현자료라는 이름과 달리 제공되는 자료에 원래 자료의 값이 들어 있을 수 있다. 즉, 완전 재현자료에 원래 관측값이 없다는 것은 용어가 초래하는 혼란인 셈이다. 하지만 완전이라는 단어를 사용하면서 얻고자 하는 것은 원자료가 전혀 들어있지 않아서 완전히 안전한 자료이다. 그러한 목적을 달성해서 원래 관측값이 없는 완전 재현자료를 만들기 위해, Raghunathan 등 (2003)에서는 이론적으로 초(super, future) 모집단을 근거로 사후예측분포를 얻고 모집단을 생성할 수 있다고 제안했다. 따라서 원래 관측값 부분 Y_{inc}도 재현자료로 대체한 후 Y_{syn}을 샘플링한다. 또한 문헌에 따라 Y_{exc}에서만 샘플링해서 자료를 제공하기도 한다. 이런 방식들을 통해 원래 관측값은 전혀 포함되지 않은 완전 재현자료를 제공할 수 있다. 보통 이렇게 생성한 재현자료를 이용해 추론을 하고자 할 때도 T_f를 사용했었다. Rubin의 방식을 따라 자료를 생성했기 때문이다. 그러나 분산 추정량 T_f를 사용해서 추론의 타당성이 확보되는 것은 $(Y_{\text{inc}}, Y_{\text{exc}})$에서 샘플링된 자료를 사용하는 경우에만이라는 것이 최근에 알려졌다 (Drechsler, 2018).

한편, 부분 재현자료는 Little의 방식으로 Y_{inc}의 일부를 재현값으로 대체해서 제공하는 것을 말한다. 이용자에게 있어 완전 재현자료와 다른 것은 추론을 위해 T_f 대신에 T_p를 사용한다는 것이다. 만약 재현자료에 원래 관측값이 없기를 원한다면, Little의 방식을 따르면서도 완전 재현자료를 만들고자 하게 된다. 이때는 Y_{inc}의 일부 대신에, 모든 값을 재현값으로 대체해서 완전 재현자료를 만들 수 있다. 그러면 원래 부분 재현자료 생성과 달리, 표본의 크기 n_{sys}도 완전 재현자료와 같이 자유롭게 조절할 수 있다. 이러한 흐름 속에 최근 개발된 보다 일반적인 분산 추정량은 다음과 같다 (Raab 등, 2017).

$$T_s = \left(\frac{n_{\text{syn}}}{n} + \frac{1}{M}\right)\bar{v}_M.$$

T_s의 장점은 T_f와 달리 음의 값을 가지지 않으며, 변동성이 더 작고, T_f 및 T_p와 달리 $M = 1$일 때도 계산된다는 것이다.

이제 Little의 방식으로 얻은 완전 재현자료를 이용해 통계적 추론을 할 때, T_s를 이용하면 된다고 생각

할 수 있지만 언제나 그렇지는 않다. 우선 모집단과 같은 크기의 표본을 만들려는 특이한 경우를 생각해보자. Drechsler (2018)에 따르면, $n < N$이고 $n_{\text{syn}} = N$이면 T_f를 사용해야 불편 추정량이 되며, $n = n_{\text{syn}} = N$이면 T_p가 불편 추정량이 된다. 따라서 그러한 경우에서는 T_s를 사용하는 것이 적절하지 않다.

정리하면, 많은 경우에 Rubin의 방식으로 완전 재현자료를 생성했다면 T_f를, Little의 방식으로 생성된 완전 재현자료라면 T_s를 사용하는 것을 자연스럽게 여겨왔다고 할 수 있다. 또한 위에서 T_f를 사용할 때, 그리고 T_s를 사용할 때 주의 사항을 각각 설명했다. 이제 T_f와 T_s 중 무엇을 사용하는 것이 좋은지에 관한 모의실험 결과를 소개하겠다. Drechsler (2018)는 각 재현자료 생성 방식에 대해 T_f 및 T_s를 모두 이용해 분산을 계산하고 비교하는 모의실험을 수행했다.

먼저 X에 대한 모집단 정보가 없을 때를 살펴 보자. 모의실험에 따르면 한 변수와 관련된 통계량, 예를 들어 Y_1의 평균에 관해 추론할 때는 일반적으로 T_f 보다는 T_s를 사용하는 것이 바람직하다고 판단된다. 다른 변수와 관련된 통계량, 예를 들어 $(Y_1|Y_2 > 0)$의 평균에 관해 추론할 때는, M이 크지 않다면 T_s보다는 T_f를 사용하는 것이 바람직하고, M이 큰 경우에는 바람직한 추정량이 없다. 한편, X에 대한 모집단 정보가 있는 경우 T_s는 T_f에 비해 편의가 매우 큰 경향을 가지므로 T_f를 사용하는 것이 바람직하다. 즉, 정보가 좀 더 많은 상황이라면 완전 재현자료의 분산 추정량으로는 T_f를 사용하는 것이 T_s보다 바람직하다고 볼 수 있다. 참고로, 모의실험 결과에 따르면 대부분의 경우 완전 재현자료를 생성하는 방식에 따라 분산 추정량의 분포가 다르다고 하기는 어려우므로, 자료 생성 방식보다는 어느 분산 추정량을 사용하는가가 추론에 있어 중요하다고 할 수 있다.

6. Conclusion

본 연구에서는 재현자료를 이해하기 위해 먼저 다중대체와 베이지안 붓스트랩에 대해 2절에서 설명하였다. 3절에서 재현자료 작성의 종류와 결합규칙에 대해 다룬 뒤, 4절에서 프랑스고용자료를 활용한 재현자료 작성 사례를 살펴보았다. 5절에서는 재현자료를 제공하고 통계적 추론을 시행할 때 발생할 수 있는 혼란의 문제를 다루었다. 특히 완전 재현자료를 사용할 때 분산 추정량으로 무엇을 사용하는 것이 좋은지를 정리했다.

재현자료를 제공하면 원래 자료를 제공하지 않아 보호 효과가 크고, 결합규칙을 사용해 분석 결과의 정확성 보장을 높일 수 있다고 여겨지므로 관련 연구가 활발하다. 하지만 다른 통계적 비밀보호 기법들이 그렇듯이, 정말 노출위험이 전혀 없는지 혹은 차등정보보호를 만족시키는지 (Machanavajjhala, 2008), 재현자료가 가지는 노출위험에 대한 연구도 필요하다고 보인다. 또한 원래 관측된 표본과 동일한 수준의 분석 유용성을 확보하기 위한 모형 연구도 지속되어야 한다.

4차 산업혁명 시대에는 데이터의 공유가 필수적이나 개인정보보호 또한 지켜져야 할 중요한 가치이다. 재현자료는 프라이버시를 보호하면서 자료를 제공할 수 있는 좋은 방법일 수 있다. 실무적 활용이 가능한 수준까지 재현자료에 관해 보다 광범위한 연구가 활성화되기를 기대한다. 이를 위해 본 논문이 좋은 길라잡이가 되기를 바란다.

References

Abowd, J. M., Kramarz, F., and Margolis, D. N. (1999). High wage workers and high wage firms, *Econometrica*, **67**, 251–333.

Abowd, J. M. and Woodcock, S. D. (2001). Disclosure limitation in longitudinal linked data. In P. Doyle, J. Lane, J. Theeuwes, L. Zayatz (Eds.) *Confidentiality, Disclosure, and Data Access: Theory and*

Practical Applications for Statistical Agencies (pp. 215–277), Amsterdam, North Holland.

Clyde, M. A. and Lee, H. K. H. (2001). Bagging and the Bayesian bootstrap. In T. Richardson and T. Jaakkola (Eds) *Artificial Intelligence and Statistics* (pp. 169–174), Morgan Kaufmann, Burlington.

Drechsler, J. (2018). Some clarifications regrading fully synthetic data. In Domingo-Ferrer, J., Montes, F. (eds.) *LNCS*, (Vol. 11126, pp. 109–121), Springer, Heidelberg.

Efron, B. (1979). Bootstrap methods: another look at the jackknife, *Annals of Statistics*, **7**, 1–26.

Little, R. J. A. (1993). Statistical analysis of masked data, *Journal of Official Statistics*, **9**, 407–426.

Machanavajjhala, A., Kifer, D., Abowd, J., Gehrke, J., and Vilhuber, L. (2008). Privacy: theory meets practice on the map. In *Proceedings of the 24th International Conference on Data Engineering*, 277–286.

Park, M. J. (2016). Comparative study on the recent SDC methods. Statistical Research Institute.

Park, M. J. and Kim, H. (2016). Statistical disclosure control for public microdata: present and future, *Korean Journal of Applied Statistics*, **39**, 1041–1059.

Park, M. J. and Kim, J. (2017). Reveiw on the synthetic data generation methodologies. Statistical Research Institute.

Raab, G. M., Nowork, B., and Dibben, C. (2017). Practical data synthesis for large samples. *Journal of Privacy and Confidentiality*, **7**, 67–97.

Raghunathan, T. E., Lepkowski, J. M., Hoewyk, J. V., and Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models, *Statistics Canada*, **27**, 85–95.

Raghunathan, T. E., Reiter, J. P., and Rubin, D. B. (2003). Multiple imputation for statistical disclosure limitation. *Journal of Official Statistics*, **19**, 1–16.

Reiter, J. P. (2002). Satisfying disclosure restrictions with synthetic data sets, *Journal of Official Statistics*, **18**, 531–543.

Reiter, J. P. (2003). Inference for partially synthetic, public use microdata sets, *Survey Methodology*, **29**, 181–188.

Reiter, J. P. (2004). Significance tests for multi-component estimands from multiply imputed, synthetic microdata, *Journal of Statistical Planning and Inference*, **131**, 365–377.

Rubin, D. B. (1978). Multiple imputations in sample surveys - a phenomenological Bayesian approach to nonresponse. In *Proceedings of the Survey Research Methods Section, American Statistical Association*, 20–34.

Rubin, D. B. (1981). The Bayesian bootstrap, *Annals of Statistics*, **9**, 130–134.

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

Rubin, D. B. (1988). An overview of multiple imputation. In *Proceedings of the Survey Research Section, American Statistical Association*, 79–84.

Rubin, D. B. (1993). Discussion statistical disclosure limitation, *Journal of Official Statistics*, **9**, 461–468.

다중대체와 재현자료 작성

김정연[a] · 박민정[b,1]

[a]충북대학교 정보통계학과, [b]통계청 통계개발원

(2018년 11월 26일 접수, 2019년 1월 4일 수정, 2019년 1월 4일 채택)

요 약

사회가 발전함에 따라 이용자의 다양한 분석 요구에 대응하기 위해 개인 단위로 구성된 마이크로데이터 제공이 증가했다. 나아가 센서스, 행정자료와 같은 전수자료를 마이크로데이터 형태로 제공받아 연구하고자 하는 요구 역시 커지고 있다. 정책결정, 학술목적 등을 위한 마이크로데이터 분석은 가치 창출 측면에서 대단히 바람직하다. 하지만 자료 유용성이 확보된 마이크로데이터 제공은 개인정보가 노출될 가능성이라는 위험을 가질 수 밖에 없다. 이에, 자료의 유용성을 확보하면서 개인정보보호를 보장할 수 있는 여러 방법들이 고려되어 왔다. 이러한 방법 중 하나로 재현자료(synthetic data)를 생성해서 활용하는 방법이 연구되어 왔다. 본 논문은 재현자료 생성과 관련된 방법론 및 주의사항을 소개하여, 재현자료의 이해를 도모하고자 한다. 이를 위해 재현자료 작성에 필수적인 다중대체, 베이지안 예측 모형 및 베이지안 붓스트랩 등의 개념들을 먼저 설명하고, 완전 재현자료 및 부분 재현자료에 대해 살펴본다. 특히, 재현자료 작성을 심도 깊이 이해하기 위해 순차회귀 다중대체(sequential regression multivariate imputation)를 이용해 경시적(longitudinal) 자료를 재현자료로 작성하는 구체적 사례를 살펴본다.

주요용어: 재현자료, 다중대체, 베이지안 예측모형, 베이지안 붓스트랩, 마이크로데이터

이 논문은 2016학년도 충북대학교 학술연구지원사업의 연구비 지원에 의하여 수행된 연구이며, 2017년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (NRF-2017R1C1B5015192).

[1]교신저자: (35214) 대전광역시 서구 대덕대로 317번길 9, 통계센터. E-mail: mjstat@korea.kr