발 간 등 록 번 호
11-1240245-000057-14

2020년 연구보고서

통계데이터센터 DB를 활용한 재현자료 생성 방법 연구

2021. 4.

http://kostat.go.kr/sri

ISSN 2288-1166(Print)
ISSN 2733-4120(Online)

통계청
통계개발원

연구보고서 2020-13

통계데이터센터 DB를 활용한 재현자료 생성 방법 연구

박민정 · 한정임 · 박노성

발간사

통계개발원(Statistics Research Institute, SRI)은 2006년에 개원한 이래 국내에 유일한 「국가통계 전문연구기관」으로서 국가통계・조사 방법론의 발전과 데이터기반 정책 연구・개발 분야에서 선도적인 역할을 하고 있습니다. 지난 한 해 동안 SRI는 "광출어혁(光出於革)"의 정신, 즉 혁신에(革) 기반한 실용적인 「팀연구」를 통해서 국가통계・데이터과학의 미래를 밝히고자(光) 노력하였습니다.

SRI의 「2020년 연구보고서」는 이러한 혁신연구의 결과물입니다. 특히 코로나19의 경제・사회 영향을 측정하는 지표 설계 연구를 비롯해서 「데이터기반 인구・사회・경제정책」을 뒷받침하는 심도 깊은 연구가 진행되었습니다. 예를 들어서 저혼인・저출산・초고령 사회에 대응하기 위한 데이터 심층 분석, 균형발전 정책수립을 지원하는 지역사회지표 체계구축, 그리고 SDG를 활용한 남북한 통합통계 방안 등입니다. 또한 2020년 2월에 출범한 「SDG 데이터연구센터」는 지속가능발전목표(SDG)의 이행현황 점검과 SDG 대국민 데이터 서비스를 준비하는 실용적인 연구를 진행하였습니다.

국가통계・데이터과학・조사방법론 분야에서는 「데이터 정보보호센터」를 중심으로 개인정보보호 방법의 통계지리정보서비스(SGIS) 적용 등 실용적인 연구가 진행되었습니다. 또한 「조사표연구센터」를 활용하여 빅데이터 활용 홈페이지 개선을 위한 사용성 평가를 비롯해서 「조사표 인지실험」을 적용한 국가통계의 품질개선이 이루어졌습니다. 특히 지난해에는 데이터과학(Data Science)을 국가통계에 활용하고자하는 기초 연구를 추진함으로써 「데이터 혁신 방법론」의 새로운 전기를 준비하였습니다.

금년에 개원 15주년을 맞는 SRI는 본 연구보고서가 증거기반정책 입안자의 데이터 활용에 실제적인 도움이 될 뿐만 아니라 국가통계 생산자의 혁신적인 조사방법론 개발에 유용한 자료로 활용될 수 있기를 기대합니다. SRI가 「국가통계 싱크탱크」로서 국내외적으로 선도적인 역할을 할 수 있도록 독자 여러분의 지속적인 격려를 부탁드립니다. 국가통계의 도약과 혁신적 개발을 위하여 제언이 있다면 언제든지 말씀하여 주십시오. 겸허히 귀 기울이겠습니다.

이번 「2020년 연구보고서」를 위하여 실용적이고 품질 높은 연구 결과를 도출하기 위해 광출어혁(光出於革)의 정신으로 최선을 다한 연구진과 대내외적으로 협력·공동연구에 참여한 민·관의 연구자들에게 따스한 감사를 전합니다. 끝으로 본 연구에서 제시된 내용 및 결과는 저자 개인의 의견이며, 통계청 또는 통계개발원의 공식적인 견해가 아님을 밝혀둡니다.

2021년 4월

통계청 통계개발원장

전 영 일

목 차

요 약

데이터는 제4차 산업혁명 시대의 원유와도 같은 역할을 하고 있다. 기계학습 모델을 학습시키기 위해서도 데이터가 필요하다. 나아가 양질의 데이터가 있으면 양질의 분석 작업과 고도로 훈련된 인공지능 모델을 이용하여 우리의 삶을 훨씬 더 풍요롭게 할 수 있다. 이처럼 데이터를 공유하는 것이 필요하지만 프라이버시 문제로 인하여 실상은 데이터가 활발히 공유되지 못하고 있다. 세계 각국은 이러한 프라이버시 이슈에 대하여 관련 법과 규제를 강화하고 있어, 데이터 공유가 앞으로 더 위축될 우려도 있다.

프라이버시 보호, 즉 개인정보 노출위험 제어를 위하여 전통적으로 데이터 익명화 방법들이 사용되고 있으며, 최근 차등정보보호(differential privacy)를 만족하도록 잡음을 추가하는 기법도 제안된 바 있다. 그러나 실제 자료에 이러한 기법을 적용하면 자료의 유용성이 떨어진다는 문제점이 있다. 때문에 노출위험을 제어하면서 자료의 유용성을 확보하기 위한 또 다른 방법으로 재현자료 생성이 연구되고 있다. 전통적인 익명화 방법들의 한계를 극복하기 위하여 제안된 재현자료는 원본자료를 대신하여 원본과 유사한 분포를 가지도록 생성된 자료를 일컫는다. 재현자료는 통계적 모형을 통해 생성될 수도 있지만 최근 딥러닝 기술을 활용한 방법도 주목받고 있다.

본 보고서에서는 재현자료 생성을 위한 통계적 모형을 활용하는 방법, 재현자료의 노출위험과 정보손실을 측정하는 방법, 국내외 통계기관의 재현자료 생성 사례 및 통계청 통계데이터센터 실제 DB에 대한 재현자료 시범 생성 결과를 설명한다. 나아가 딥러닝 기술을 이용한 재현자료 연구 동향을 상세히 소개한다.

주요용어 : 프라이버시, 익명화, 재현자료, 다중대체, GAN

제 1 장

서 론

데이터 정보보호(data privacy)는 인공지능 시대에서 중요한 문제로 인식되고 있다. 많은 인공지능 모델이 데이터 기반의 학습을 전제로 하고 있으나, 개인정보보호 등 여러 현실적인 이유로 인하여 이해 당사자들 사이의 데이터 공유가 활발하게 이루어지기는 어려운 상황이다. 익명성(anonymity)을 확보하여 프라이버시 보호 문제를 해결하고자 사용되어 온 매스킹(비식별화) 기법은 여러 상황에서 그 한계가 명확하게 지적되고 있으며, 몇몇 실제 사례에서 익명화된 개인 자료들이 재식별(reidentification)되는 것을 발견할 수 있다.

최근에는 재현자료(synthetic data)를 활용하여 기존의 문제점을 극복하려는 일련의 연구들이 수행되고 있다. 원본자료의 분포를 통계적 모형으로 추정한 후 유사한 자료를 생성하는 통계학 분야의 연구뿐만 아니라, 인공지능 모델이 원자료의 통계적 특성을 학습하게 한 후 그와 비슷한 재현자료를 생성하는 연구들도 이루어지고 있다. 후자의 경우 딥러닝 모델을 비롯한 다양한 방법론에 기반을 둔 연구들이 진행되고 있다. 본 보고서의 마지막 제6장에서는 이러한 인공지능 모델 기반의 재현자료 생성 기술에 대한 최근 연구 성과를 다룬다. 여러 인공지능 모델 중에서 적대적 생성망(generative adversarial network) 기반의 기술에 대한 이론 및 여타 기술들과의 비교 평가 결과를 소개한다.

한편 본 보고서의 제2장에서 제5장까지는 재현자료를 다루는 통계적 접근을 전반적으로 소개하고, 통계청 통계데이터센터 실제 DB에 대하여 재현자료를 시범 생성한 결과를 상세히 논의한다. 재현자료가 전통적인 매스킹의 한계를 극복하기 위한 대안으로 다루어지고 있으나, 실제로 재현자료를 생성한 사례는 해외에도 많지 않다. 또한 거의 모든 사례에서 통계적 모형을 이용하여 재현자료를 생성하고 있으므로, 본 보고서에서는 통계적 모형을 활용한 재현자료 생성에 관한 내용을 종합적으로 다루었다. 향후 이러한 통계적 모형을 이용한 접근을 제6장에서 소개하는 딥러닝을 이용한 재현자료 생성과 비교하고, 더욱 효율적인 대안을 논의하는 방향으로 발전되기를 바란다. 한편 이러한 발전의 초석으로서 기능하기 위하여, 본 보고서는 일정 수준 이상의 전문 지식과 경험을 보유한 연구자들을 대상으로 작성되었

음을 밝히는 바이다.

재현자료 생성 연구는 공공데이터의 효과적인 공유를 가능하게 할 수 있는 후보 기술로 여겨지고 있으므로 그 잠재적 파급효과가 크다고 판단된다. 동시에 민간에서의 활발한 데이터 공유를 촉진할 수 있는 촉진제 역할을 할 수 있을 것으로도 기대된다. 또한 많은 이해 당사자들의 개인정보보호 우려를 경감시키면서 활발한 데이터 공유, 나아가서는 활발한 인공지능 기술의 확산으로 이어질 수 있을 것으로 예상한다. 이러한 기대를 받고 있는 재현자료 생성에 관한 내용을 본격적으로 다루기에 앞서, 데이터 정보보호 방법을 간략히 소개하며 서론을 마무리한다.

제1절 데이터 정보보호(Data Privacy) 방법

1. 데이터 익명화(Data Anonymization)

데이터 익명화는 식별자(identifier)를 삭제하고 준식별자(quasi-identifier)를 변형 및 조작한 후에 민감정보를 공유하거나 공개하는 행위를 일컫는 용어이다. 식별자는 특정 개인과 1:1 대응 관계에 있는 정보를 말하는 것으로 주민등록번호가 그 대표적인 예이다. 준식별자는 그 자체로서는 식별자가 아니지만 다른 정보와 결합하면 개인을 특정할 수 있는 속성들을 의미한다. 민감정보는 개인의 사생활 정보를 담고 있는 것으로 인공지능 모델 학습이나 빅데이터 분석에서 주로 사용된다.

k-익명성은 대표적인 프라이버시 모델이며, 익명화 후에는 준식별자를 기준으로 한 개인의 레코드가 다른 $k-1$ 명의 레코드와 구별되지 않도록 하는 목적을 가진다. 예를 들어 아래 그림 <그림 1-1>은 데이터에 우편번호(zip code)와 연령 정보가 있을 때 우편번호의 마지막 자리를 마스킹[1]하고 연령을 연령대(age band)로 바꿀 경우 k 명의 준식별자가 같아지게 되는 것을 보여준다. 이는 익명 처리된 준식별자만을 가지고는 해당 개인의 신분을 밝힐 수 없음을 의미한다.

1) 여기서 마스킹(masking)은 *표시 등으로 원자료의 값을 가리는 한 가지 행위를 나타낸다. 반면 각국 통계기관에서는 매스킹/마스킹(masking)을 비밀보호를 위한 자료 변형 기술 전체를 가리키는 용어로 사용하고 있다. 참고로 국내에서는 통계기관에서 매스킹 처리라고 부르는 것을 비식별화라고 지칭하는 경향이 있다.

No	ZIP	Age	Salary	Disease
1	47677	29	3K	AIDS
2	47672	22	4K	Ebola
3	47678	27	5K	Cancer
4	47905	53	6K	AIDS
5	47909	52	11K	Ebola
6	47906	57	8K	Heart Disease

No	ZIP	Age	Salary	Disease
1	4767*	≤ 40	3K	AIDS
2	4767*	≤ 40	4K	Ebola
3	4767*	≤ 40	5k	Cancer
4	4790*	≥ 50	6K	AIDS
5	4790*	≥ 50	11K	Ebola
6	4790*	≥ 50	8K	Heart Disease

출처: Park et al.(2018)

<그림 1-1> 원본자료(왼쪽)와 익명화 예제(오른쪽)

이러한 k-익명성은 동질성 공격에는 취약하다. 예를 들어 준식별자가 같은 k명의 개인이 모두 같은 속성을 가지고 있을 경우, 해당 개인의 신원을 알 필요도 없이 모두 동일한 속성을 가지고 있으므로 질병의 종류와 같은 민감한 정보가 유출될 수 있다. 이를 방지하기 위한 것이 l-다양성이다. 이는 준식별자가 같은 k명의 개인이 적어도 l개의 다양한 속성을 가지고 있어서 서로 간의 동질성이 존재하지 않음을 의미한다.

하지만 속성들이 좁은 분포에 다르게 분포해 있을 때, l-다양성만으로는 여전히 프라이버시가 침해될 수 있다. 이를 보완하기 위하여 제안된 t-근접성은 준식별자가 같은 k명의 속성 분포 및 전체 데이터에서 해당 속성 분포와의 차이가 t이하이도록 만드는 것을 일컫는다. 분포의 차이는 EMD(Eather Mover's Distance)를 이용해서 측정된다. 한편 정보보호를 위하여 이러한 KLT 기준을 만족하게 하면 자료에서 상당한 정보손실이 발생하는 것을 피할 수 없다.

위와 같이 다양한 개념의 익명화 기술이 개발되었지만, 현실에서는 노출위험 제어 측면에서조차 여전히 취약한 것으로 여겨지고 있다(Ohm, 2010). 따라서 다른 방식의 프라이버시 보호 방법이 절실히 요구되고 있다.

2. 차등정보보호(Differential Privacy)

차등정보보호는 데이터베이스의 집합 정보(aggregate information)를 계산함에 있어서 특정 레코드가 데이터베이스에 포함되어 있음을 숨기는 방식의 프라이버시 보호 개념을 의미한다. 예를 들어 어떤 레코드 하나가 데이터베이스 D_1에는 포함되지만 D_2에는 포함되어 있지 않고 D_1과 D_2가 이 외의 차이점이 없을 경우, D_1과 D_2로부터 계산한 평균 나이를 비교하면 해당 레코드 개인의 나이를 추정할 수 있다.

어떤 알고리즘이 차등정보보호 특성이 있다는 것은 그 알고리즘의 계산 결과를 보고 특성 레코드가 원본자료에 포함되어 있는지 없는지 알 수 없음을 의미한다. 엄

밀한 수학적인 정의는 아래와 같다.

$$\Pr[A(D_1) \in S] \leq \exp(\epsilon) \times \Pr[A(D_2 \in S]$$

예를 들면 S는 알고리즘 A의 이미지(가능한 모든 출력값의 집합)의 임의의 부분집합을 의미한다. 즉, A의 이미지의 모든 부분집합에 대해서 해당 조건을 만족하여야 한다. 위의 조건을 ϵ-차등정보보호라고 부른다.

차등정보보호는 프라이버시 보호 개념으로 그것을 이루는 방법은 다양하게 설계될 수 있다. 라플라스 방법은 알고리즘이 계산된 결과에 라플라스 분포를 따르는 노이즈를 더하여 프라이버시를 보호하는 방법이다. 그 외에도 지수적 방법이나 사후확률 표본 방법 등이 존재한다.

특히 이런 차등정보보호는 딥러닝 모델에서 큰 영향력을 끼치며, 차등정보보호를 만족하는 딥러닝 훈련 알고리즘도 속속 연구되고 있다. 이러한 방법으로 딥러닝 모델을 훈련할 경우, 외부의 쿼리만으로는 원본 학습 데이터의 특징을 추론하기 어렵게 된다. 딥러닝 모델의 경우 원본 학습 데이터의 특징이 유출되면 그 모델의 잠재적인 취약점도 같이 노출될 수 있어서 차등정보보호를 만족하는 학습 알고리즘은 중요한 연구 주제로 인식되고 있다. 참고로 일각에서는 차등정보보호를 차분정보보호 또는 차분프라이버시라고 번역하여 부르기도 한다.

3. 재현자료(Synthetic Data)

재현자료 생성은 데이터베이스 및 정보보안 분야에서도 오랫동안 연구되어온 주제이다. 응결 방법(condensation method, Aggarwal et al., 2004)은 통계적 정보 기반의 샘플링으로 재현자료를 생성하는 전통적인 방법이다. 그 후로도 많은 기술들이 제안되었으며 PrivBayes(Zhang et al., 2014)는 베이지안 네트워크를 이용한 차등정보보호 데이터 공개 방법이다. CTGAN(Xu et al., 2019)은 조건부 적대적 신경망 기술을 활용한 최신 재현자료 생성 방법으로 현시점을 기준으로 데이터 유용성 측면에서 가장 뛰어난 재현 성능을 보이고 있다. 하지만 CTGAN은 프라이버시 보호에 대한 고려를 하지 않고, 원본자료와 비슷한 데이터를 재현하는 것에 중점을 두고 있다.

그 외에도 변분오터인코더(Variational Autoencoder)를 이용한 기술을 비롯한 다양한 형태의 적대적 신경망을 활용한 연구들이 존재한다. PATE-GAN(Jordon et al., 2019)은 대표적인 적대적 신경망을 활용한 차등정보보호 기반의 재현자료 생성 기술이다.

공식 통계 이외의 영역에서 재현자료 생성 기술이 가장 활발히 연구되고 있는 분야는 바로 의료 데이터이다. 의료 데이터는 인공지능을 활용한 의료 기술 발전에 필

수적이지만 공유가 쉽지 않다. 의료 분야에서는 전통적으로 익명화 기술을 계속 사용했지만, 그 한계가 인식되고 있어서 딥러닝 기반의 재현자료 생성 기술이 최근에 많이 연구되고 있다. 대표적으로 medGAN(Choi et al., 2017), ehrGAN(Che et al., 2017) 등이 있다. 특히 medGAN은 오토인코더와 적대적 신경망 기술을 통합한 기술로서 그 구성이 다른 기술들과는 특이한 면이 있다. 한편 Fan et al.(2020)에 따르면 재현성능이 유의미하게 개선되고 있으나, 여전히 더욱 발전해야 할 필요성이 있다.

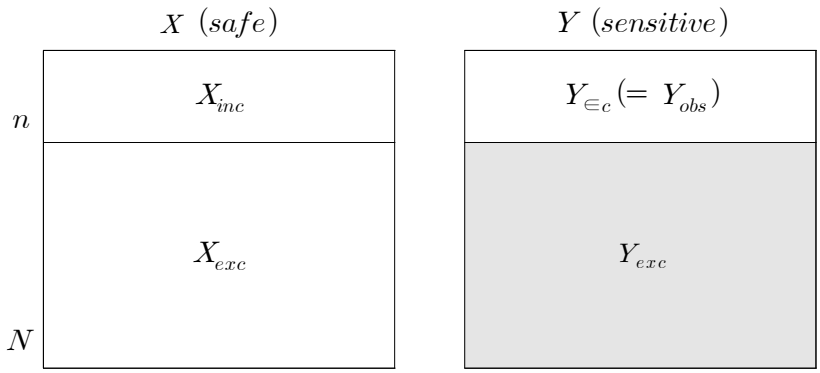
제 2 장

통계적 모형을 이용한 재현자료 생성

이 장에서는 통계적 모형을 이용한 재현자료 생성 방법에 대해 알아본다. 선행연구인 박민정과 김정연(2017)에 기본적인 개념이 상세히 설명되어 있다. 여기서는 제5장의 재현자료 생성 과정을 이해하기 위하여 모수적 재현자료 생성에 관하여 먼저 정리하고, 비모수적 모형을 이용한 재현자료 생성을 간략히 설명한다.

제1절 모수적 다중대체 모형을 이용한 재현자료 생성

재현자료는 다중대체 기법을 활용하여 생성한다. 다중대체는 베이지안 통계학에서 비롯하였으며, 추정된 사후예측분포로부터 여러 세트의 자료를 생성한다. 따라서 이용자는 다중대체된 여러 데이터 세트에 동일한 자료 분석을 수행하고, 그 결과를 결합하여 최종 통계를 생산한다. 재현자료의 생성은 관측되지 않은 자료를 무응답 자료처럼 취급하여 다중대체하고 일부를 추출하여 이루어진다. 재현자료에서 고려하는 기본적인 자료의 구조는 다음 <그림 2-1>과 같이 표현될 수 있다. 외부인이 쉽게 얻을 수 있는 변수들은 X로 표현하고, 통계기관의 데이터베이스가 추가로 제공하는 정보를 가지는 변수들은 Y로 표현되어 있다. 통계기관은 $Y(Y_{obs})$의 자료 유용성을 보존(정보손실을 최소화)하면서 개인정보 노출을 최소화하여 자료를 제공하고자 한다.



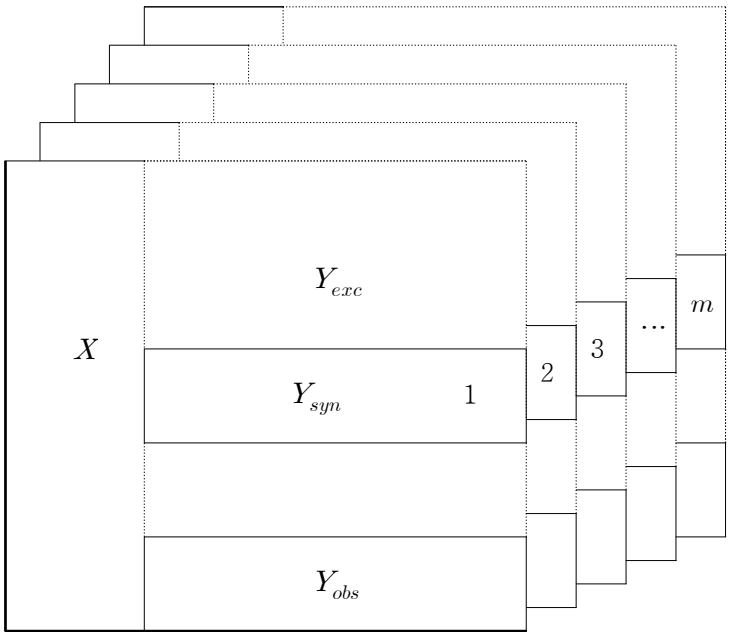
<그림 2-1> 자료구조

재현자료 생성은 완전(fully) 재현자료 생성과 부분(partially) 재현자료 생성으로 나눌 수 있다. 먼저 완전 재현자료 생성은 <그림 2-1>과 같은 자료구조에서 관측된 자료를 기반으로 사후예측분포를 추정하고, 추정된 분포로부터 모집단 전체 자료를 생성한 다음, 생성된 모집단에서 일부 자료를 추출하는 것을 말한다. 완전 재현자료 생성은 다음과 같은 두 단계로 정리할 수 있다.

(1) 결측값 대체: Y_{exc} 전체를 결측값으로 간주하고, (X, Y_{obs})을 바탕으로 추정한 사후예측분포로부터 Y_{exc}의 대체값을 생성한다.

(2) 랜덤 추출: 재현된 모집단, 즉 조사에 참여하지 않은 사람들에 대하여 생성한 자료 (X_{exc}, Y_{exc})로부터 랜덤하게 추출한 표본을 배포한다.

다중대체와 마찬가지로 위의 과정을 m번 반복하여 m개의 데이터 세트를 생성하며, 생성된 재현자료는 <그림 2-2>와 같이 표현할 수 있다. 관측된 자료 Y_{obs}를 대신하여 추출한 m개의 Y_{syn} 데이터 세트를 재현자료로 사용한다.



주: J. Reiter and J. Drechsler, Lecture notes in a professional development continuing education course "Synthetic data sets for statistical disclosure limitation", JSM 2017.

<그림 2-2> 생성된 완전 재현자료의 구조

한편 재현자료를 생성하기 위한 사후예측분포는 다음과 같다.

$$P(Y_{exc}|X, Y_{obs}) = \int P(Y_{exc}|X, Y_{obs}, \theta)\, P(\theta|X, Y_{obs})\, d\theta$$

만약 변수들에 대하여 적절한 분포를 가정할 수 있다면, 해당 분포의 모수를 추정하고 자료를 생성할 수 있다. 예를 들어 변수들이 다변량 정규분포를 따른다고 가정할 수 있으면 사전분포로 Wishart 분포를 사용하여 사후예측분포를 따르는 자료를 생성할 수 있다. 예를 들어 총 p개의 변수 및 n개의 레코드로 구성된 Y_{obs}을 간략히 Y라고 표현하면, 이 과정은 다음과 같이 정리된다.

(1) 주어진 자료에서 평균 $\bar{y}$과 분산 S을 구한다.

(2) Wishart 분포 $W_p(S^{-1}/(n-1), n-1)$으로부터 난수 W 생성 후 $\Sigma^* = W^{-1}$이라 한다.

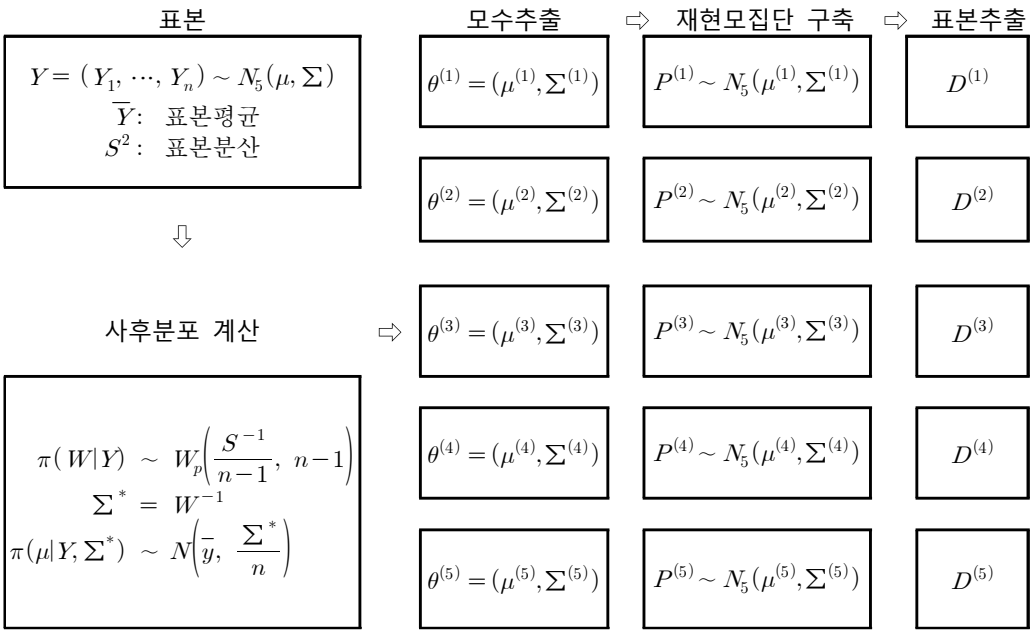
(3) 다변량 정규분포 $N_p(\bar{y}, \Sigma^*/n)$로부터 난수 μ^*를 생성한다.

(4) 다변량 정규분포 $N_p(\mu^*, \Sigma^*)$로부터 크기 N인 재현 모집단을 생성한다.

(5) 재현 모집단으로부터 크기 n^*인 표본을 무작위 추출 후 저장한다.

(6) 주어진 표본에 대하여 (2)~(5)을 $m = 5$회 반복한다.

변수가 5개인 경우를 그림으로 표현하면 다음과 같다.



<그림 2-3> 다변량 정규분포를 이용한 완전 재현자료 생성

다만 실제 자료를 대상으로 재현자료를 생성할 때 모수적으로 분포를 가정하는 것이 적절하지 않은 경우가 많다. 혼합분포(mixture distribution)를 활용하여 모수적 접근 방식의 한계를 어느 정도 극복할 수도 있으나 이러한 시도의 효과는 제한적이다.

한편 다중대체를 실시할 때 사후예측분포를 추정하기 위하여 사용되는 방법을 크게 결합모형(joint modeling) 및 조건부모형(fully conditional specification)으로 구분할 수 있다. 재현자료 생성 역시 이러한 두 접근 방식이 가능하다. 위에서 설명한 예제가 결합모형을 사용한 경우를 나타낸다. 반면 조건부 분포를 사용하여 순차적으로 한 변수씩 표본을 생성하는 방법을 순차회귀 다중대체(Sequential Regression Multiple Imputation, SRMI)라고 하며(Raghunathan et al., 2001), 조건부모형에 속한다. Multiple Imputation by Chained Equations(MICE) 역시 동일한 방식을 지칭한다(van Buuren et al., 2011).

SRMI 방식을 적용하면 $Y=(Y_1,\cdots,Y_p)$의 조건부 결합분포는 다음과 같이 표현된다. 한 변수씩 조건부 분포를 구하여 자료를 생성하고, 생성된 자료를 관측된 자료처럼 사용하여 다음 변수의 조건부 분포를 구한다.

$$P(Y_1,\ Y_2,\ \cdots,\ Y_p|X,\ Y_{obs},\theta)$$

$$=P(Y_1|X,\ Y_{obs},\theta)P(Y_2|X,\ Y_{obs},\theta,\ Y_1)\cdots P(Y_p|X,\ Y_{obs},\theta,\ Y_1,\ Y_2,\ \cdots\ Y_{p-1})$$

위의 조건부 확률을 바탕으로 한 SRMI 알고리즘은 다음과 같다.

(1) $P(\theta|X,\ Y_{obs})$로부터 표본 θ^* 생성

(2-1) $P(Y_1|X,\ Y_{obs},\theta^*)$로부터 표본 Y_1^* 생성

(2-2) $P(Y_2|X,\ Y_{obs},\theta^*,\ Y_1)$로부터 표본 Y_2^* 생성

(2-3) $P(Y_3|X,\ Y_{obs},\theta^*,\ Y_1^*,\ Y_2^*)$로부터 표본 Y_3^* 생성

.....

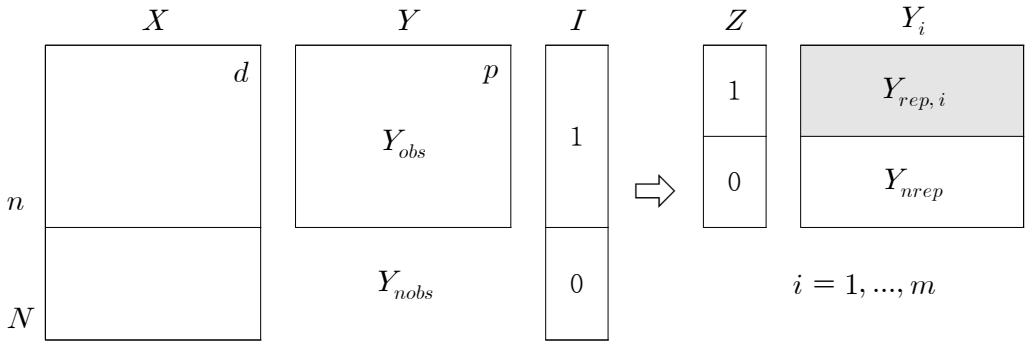
(2-k) $P(Y_p|X,\ Y_{obs},\theta^*,\ Y_1^*,\ Y_2^*,\ \cdots,\ Y_{p-1}^*)$로부터 표본 Y_p^* 생성

(3) 원하는 크기 n인 표본을 얻을 때까지 (1)~(2)의 과정을 반복한다.

SRMI는 한 번에 한 변수씩 표본을 생성하기 때문에 여러 변수를 고려하더라도 각 변수의 특성에 맞는 분포로부터 표본을 생성할 수 있다. 따라서 변수들 사이의 관계를 각 단계에서 충분히 고려할 수 있고, 다양한 모형을 가정하는 것이 가능하다.

마지막으로 부분 재현자료에 관하여 간략히 살펴보자. 부분 재현자료는 모집단 전체를 재현하는 대신 관측된 변수 중에서 노출위험이 높은 일부 레코드 또는 일부 변

수를 재현하는 것을 말한다. 다음 <그림 2-4>는 부분 재현자료를 생성하기 위한 자료구조를 나타며, 일부 레코드를 재현자료로 대체하는 것을 보여준다. 대체된 자료는 Y_{rep}(replaced)로 표현되어 있다.



<그림 2-4> 부분 재현자료 생성을 이해하기 위한 자료구조

참고로 재현자료와 다중대체와의 차이점은 다음과 같다. 다중대체는 데이터가 결측된 상황에서 사후예측분포를 추정하지만, 재현자료 생성은 관측된 자료로부터 사후예측분포를 추정한다. 때문에 불확실성 추정에 대한 가정이 달라지며, 결합규칙도 다르다. 다중대체, 완전 재현자료, 부분 재현자료를 사용하여 각 데이터 세트에서 이용자가 원하는 분석을 수행하고, 그 결과를 결합하는 규칙을 정리하면 다음 <표 2-1>과 같다. 통계량 혹은 분석결과는 각 세트에서 얻은 결과의 평균으로 동일하지만, 분산 추정식이 조금씩 다르다. <표 2-1>에서 사용한 b_m 및 v_m은 다음과 같다.

$$b_m = \sum_{i=1}^{m} (q_i - \bar{q}_m)^2/(m-1),\ \ \bar{v}_m = \sum_{i=1}^{m} v_i/m$$

<표 2-1> 결합 규칙(combining rule) 비교

	통계 Q에 대한 추정값	분산 추정량
다중대체	$\bar{q}_m = \sum_{i=1}^{m} q_i/m$	$T_m = b_m/m + b_m + \bar{v}_m$
완전 재현자료		$T_f = b_m/m + b_m - \bar{v}_m$
부분 재현자료		$T_p = b_m/m + \bar{v}_m$

제2절 비모수적 다중대체 모형을 이용한 재현자료 생성

사후예측분포를 모수적으로 구하기 어려운 경우에 비모수적 접근이 이루어질 수 있다. 비모수적으로 완전 재현자료를 생성할 때는 보통 베이지안 붓스트랩으로 랜덤 추출된 자료를 사용한다. 앞 절 <그림 2-3>의 다변량 정규분포를 따르는 예제에서 베이지안 붓스트랩을 이용한 재현자료 생성과정을 설명하면 다음과 같다. <그림 2-5>은 이를 도식화하였다.

(1) 균일분포 $U(0,1)$로부터 $(n-1)$개의 난수를 생성하고 오름차순으로 정렬하면 다음과 같다.

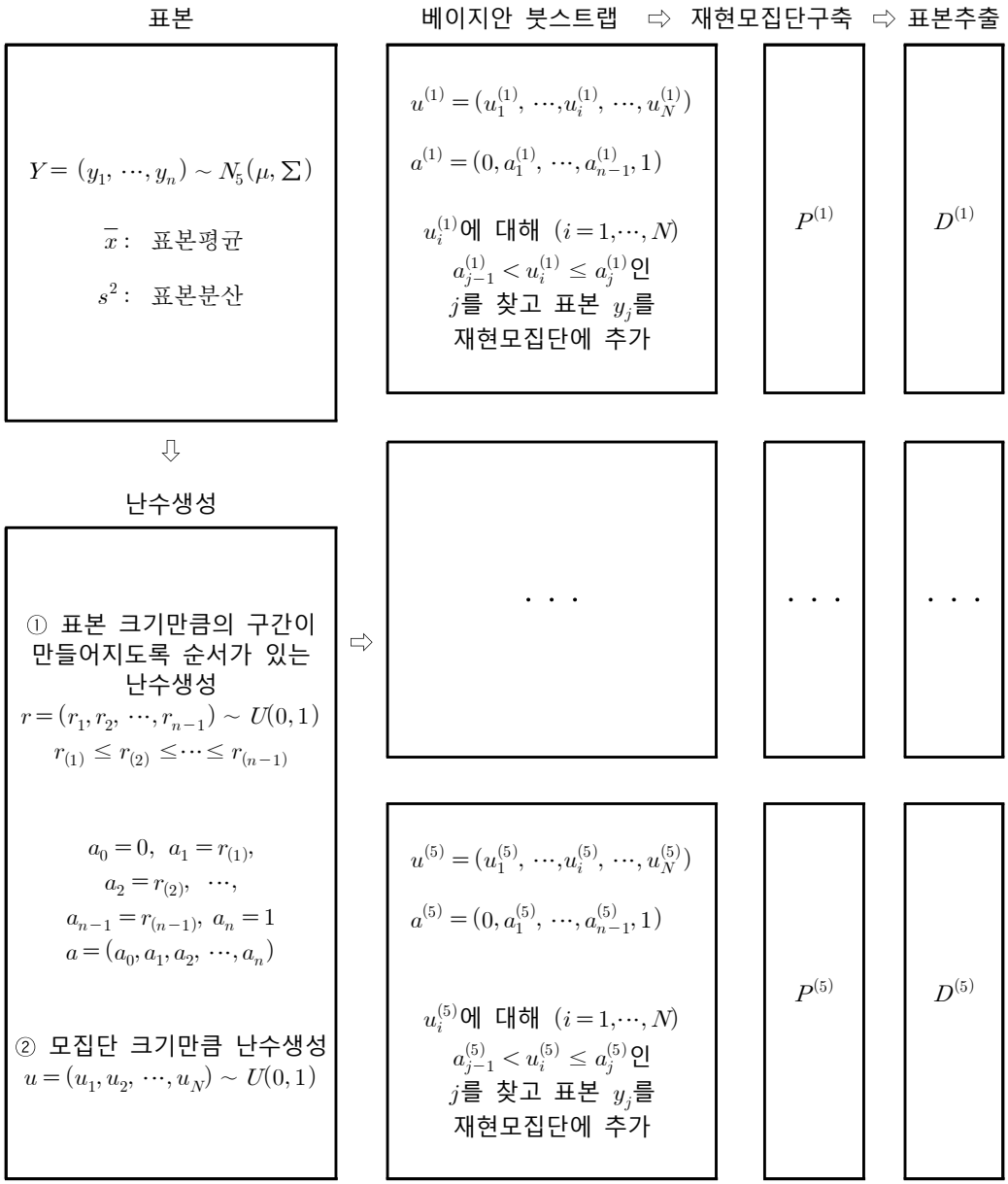
$$a_0 = 0,\ a_1,\ a_2,\ \cdots,\ a_{n-1},\ a_n = 1$$

(2) 균일분포 $U(0,1)$로부터 새로운 N개의 난수 $u_1, u_2, \cdots, u_N$을 생성한다. 각 u_i $(i = 1, \cdots, N)$에 대하여, $a_{j-1} < u_i \leq a_j$를 만족하는 j를 찾은 뒤, 이에 대응하는 표본 y_j를 재현 모집단에 포함시킨다.

(3) (2)에서 구한 재현 모집단으로부터 크기 n^*인 표본을 무작위 추출 후 저장한다.

(4) 주어진 표본에 대하여 (1)~(3)의 과정을 m회 반복한다.

한편 CART는 주어진 여러 설명변수들을 이용해 한 변수의 조건부 분포를 추정하기에 적절한 비모수적 방법으로 알려져 있다. Reiter(2005b)는 CART를 이용하여 재현자료를 생성하는 방법을 제안하였다. 이는 SRMI 방식을 따라가면서 조건부 분포를 비모수적으로 추정하기에 적절한 모형이다. 국내에서는 주로 CART를 사용하여 재현자료를 생성하는 사례가 있었으며, 이와 관련하여 자세한 내용을 5장의 통계데이터센터 DB의 재현자료 시범 생성에서 다룬다.



<그림 2-5> 베이지안 붓스트랩을 이용한 완전 재현자료 생성

제 3 장

재현자료의 노출위험과 정보손실

이 장에서는 재현자료의 노출위험과 정보손실을 측정하는 방법을 논의한다. 공표하는 자료에 대하여 노출위험 및 정보손실을 측정하는 방법은 매스킹 처리 연구와 더불어 발전해 왔으며, 매스킹 처리된 자료에 대한 측도 관련 내용은 박민정(2014)에 자세히 기술되어 있다. 이 장에서는 재현자료를 위하여 노출위험과 정보손실을 측정하는 방법을 고찰하고자 한다.

다중대체 기법을 이용한 재현자료 생성 및 활용이 매스킹 처리의 한계를 극복하기 위하여 제안되었음에도 불구하고(Little, 1993; Rubin, 1993), 노출위험 및 정보손실 측도는 기존의 매스킹을 위한 측도들을 그대로 사용하는 경우가 많다. 새로 제안된 측도들 역시 기존의 측도를 기반으로 하는 경우가 대부분이다. 따라서 이 장에서는 기존의 측도를 정리하고, 몇 가지 새로 제안된 측도들을 소개하도록 한다.

참고로 재현자료를 활용하는 전략은 국가마다 다르다. 미국에서는 주로 원자료를 대체하여 제공하는 용도로, 캐나다에서는 RDC(Research Data Center)를 방문하기 이전에 사전 분석용으로 제공한다. 영국에서는 개인정보 노출위험 때문에 잘 활용되지 않는 자료를 배포하기 위한 목적으로 재현자료 연구가 시작되기도 하였다(Nowok et al., 2017). 따라서 재현자료의 이용 목적이나 활용 전략에 따라 노출위험 측도 적용의 엄격함이나 배포하는 데이터 세트의 개수가 다를 수 있다. 그러므로 이 장에서 정리하는 다양한 노출위험 및 정보손실 측도들은 상황에 맞게 활용될 필요가 있다. 2020년 수행되고 있는 통계청 통계데이터센터 DB의 재현자료 시범 구축 사업의 경우에는 캐나다와 유사하게 이용자가 RDC를 방문하기 이전에 사전 분석용 자료 제공을 목적으로 용역 사업이 진행되었다.

제1절 노출위험 측정

고전적인 매스킹 처리를 평가하기 위한 노출위험 측정은 실무적으로는 키변수 조합의 유일성을 근거로 이루어진다. 먼저 이와 관련된 내용을 선행연구(박민정, 2014)를 참조하여 간략히 정리하자.

마이크로데이터에서 노출위험을 직접 측정하기는 어려우므로 키변수 조합에 대하여 빈도표를 작성하고 노출위험을 측정할 수 있다. 이를 이해하기 위하여 예를 들어 <표 3-1>과 같이 4개의 변수 S(성별, 2개 범주), H(가구구성, 10개 범주), A(나이, 18개 범주), M(혼인상태, 4개 범주)로 이루어진 마이크로데이터를 가정하자.

<표 3-1> 마이크로데이터 예제

ID	S	H	A	M
1	2	10	8	3
2	1	3	5	2
...	...	...	...	...
n = 8,399	2	9	17	1

이 마이크로데이터에 대하여 4개 키변수를 이용해 빈도표를 작성하면 다음 <표 3-2>와 같다. 빈도표에서 k번째 셀의 값을 f_k라고 하자. 또한 동일한 셀에 대하여 모집단에서 빈도를 F_k라고 하자. 노출위험 측도 중에서 직관적으로 이해하기 가장 쉽고 널리 활용되고 있는 k-익명성을 기준으로는 빈도가 1 또는 2인 경우 노출위험이 높다고 판단할 수 있다. 한편 통계적으로는 조사된 마이크로데이터와 함께 제공하는 가중값을 감안한 적절한 모형을 적용하여 노출위험을 추정할 수도 있다. 예를 들어 표본 마이크로데이터에서 유일한 개체가 모집단에서도 유일할 확률이나 모집단 개체수 역수의 기댓값을 이용해 다음과 같이 노출위험 측도를 정의하기도 한다 (Skinner & Shlomo, 2008).

$$r_{1k} = P(F_k = 1 | f_k = 1), \quad r_{2k} = E[1/F_k | f_k = 1]$$

<표 3-2> 빈도표 예제

ID	S	H	A	M	빈도
1	1	1	1	1	f_1
2	1	1	1	2	f_2
...	...	...	...	...	...
K = 1,440	2	10	18	4	f_K

위와 같이 정의된 노출위험을 추정하기 위하여 음이항 모형, 로그 선형 모형 등이 제안되어 있으며, 마이크로데이터 매스킹 처리를 위한 프로그램인 μ-Argus 및 R패키

지인 sdcMicro에 일부 모형이 구현되어 있기도 하다. 각 개체(record)는 자신이 속하는 빈도표의 셀에 대한 위의 추정값을 노출위험으로 가지며, 이로써 각 개체의 노출위험을 나타낼 수 있다. 파일 전체에 대한 노출위험은 개체별 노출위험의 평균을 활용하는 경우가 많다.

예를 들어 가계금융·복지조사 마이크로데이터의 노출위험을 평가한 결과를 살펴보자(박민정 외, 2013). 총 12개의 키변수를 제공하고자 할 때 제공하는 변수의 개수에 따른 노출위험을 살펴보아 제공범위를 결정하고자 하는 경우를 가정하였다. 다음 <표 3-3>은 그 결과를 보여준다. 변수의 이름은 각각 시도(V1), 성별(V2), 연령(V3), 교육정도(V4), 종사상지위(V5), 동거여부(V6), 혼인상태(V7), 직업(V8), 가구원수(V9), 주택유형(V10), 주거면적(V11), 주택소유형태(V12)이며, 각 변수를 포함할 때는 1, 포함하지 않을 때는 0으로 표시하였다. 처음 5개 변수는 제공하는 것으로 가정하고, 나머지 7개 변수 중에서 제공 여부에 따른 시나리오는 $2^7 = 128$ 가지 존재한다. 각각 시나리오에 대하여 U1은 유일한 개체수를 U2는 2개씩 존재하는 레코드의 수를 나타낸다. 마지막 열의 risk는 앞에서 설명한 개체별 유일성에 근거한 노출위험 추정값을 네덜란드 통계청이 제안한 음이항 모형을 이용해 추정하고 평균값(%)으로 나타낸 것이다. 총 $n = 19{,}744$ 개 레코드 중에서 5개 변수만을 제공할 때 유일한 개체는 1,938개에 불과하였으나, 7개 변수를 모두 제공할 때는 17,138개의 레코드가 유일한 것으로 나타난다. 다만 이 마이크로데이터는 조사 자료이므로 가중값을 반영하면 파일 전체에 대하여 노출위험의 평균은 총 12개의 모든 변수를 제공하더라도 0.42%에 불과하다. 전수자료의 경우 통계적 모형을 이용한 노출위험 측정하기보다 <표 3-3>의 U1 또는 U2와 같은 k-익명성 기준을 직접 적용할 수 있다.

<표 3-3> 유일성에 근거한 노출위험 측정 예제

	키변수 포함 여부												노출위험 평가		
	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	U1	U2	risk
S1	1	1	1	1	1	0	0	0	0	0	0	0	1938	1740	0.07
S2	1	1	1	1	1	0	0	0	0	0	0	1	4854	2826	0.14
S3	1	1	1	1	1	0	0	0	0	0	1	0	5746	3178	0.17
S4	1	1	1	1	1	0	0	0	0	0	1	1	9001	3626	0.25
S5	1	1	1	1	1	0	0	0	0	1	0	0	4369	2748	0.13
...	...												...		
S127	1	1	1	1	1	1	1	1	1	1	1	0	15837	2246	0.39
S128	1	1	1	1	1	1	1	1	1	1	1	1	17138	1647	0.42

지금까지 살펴본 키변수 조합 유일성에 근거한 노출위험 측정은 민감변수에 대한 매스킹 처리의 효과를 반영하지 못하는 단점이 있다. 더욱 폭넓은 범위의 매스킹 처리에 대해서 효율적으로 노출위험을 측정하기 위하여 외부 침입자에 대하여 의사결정 이론을 적용한 노출위험 측정 방식이 제안된 바 있다(Reiter, 2005a). 이는 키변수뿐만 아니라 민감변수에 대한 매스킹 처리의 효과를 평가할 수 있는 장점이 있으나 실무적으로 활용하기에는 어렵다는 단점이 있다(박민정, 2014).

외부 침입자에 대하여 의사결정 이론을 적용한 노출위험 측정 방식을 정리하면 다음과 같다. 총 p개 변수를 가지는 행렬 형태의 마이크로데이터를 Y라고 하면, i번째 레코드는 $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{ip})$라고 표현할 수 있다. 또한 외부인이 정보를 가지고 있는 변수들로 이루어진 부분을 A(available), 외부인이 가지고 있는 정보는 없으며 오직 공표자료를 통해서만 정보를 얻을 수 있는 변수들로 이루어진 부분을 U(unavailable)라고 하면, 자료의 참값 정보를 $\boldsymbol{y}_i = (\boldsymbol{y}_i^A, \boldsymbol{y}_i^U)$라고 표현할 수 있다. A와 U를 구분하는 외부인의 정보는 모든 개체에 대해서 동일하다고 가정한다. 통계기관이 공표를 위해 Y를 매스킹 처리한 결과를 Z라고 하자. 공표하는 마이크로데이터 Z에서 j번째 레코드는 $\boldsymbol{z}_j = (\boldsymbol{z}_j^A, \boldsymbol{z}_j^U)$이다.

Z	
A	U

외부인이 가지고 있는 표적은 $\boldsymbol{t} = \boldsymbol{t}_i^A$이며 이는 원래 마이크로데이터 Y에서 i번째 개체 $\boldsymbol{y}_i^A$와 같은 것이라 가정할 수 있다. 외부인은 이제 n개의 개체를 가지는 매스킹 처리된 공표자료 Z에서 어떤 $\boldsymbol{z}_j^A$을 찾아 $\boldsymbol{t}$라고 판단하고 $\boldsymbol{z}_j^U$를 $\boldsymbol{t}$와 연결하려는 목적을 가지게 된다. 외부인은 공표자료 Z의 모든 개체에 대하여 표적과 동일할 확률을 다음과 같은 조건부 확률을 이용하여 계산하고, 그 확률이 가장 높은 개체를 표적이라 결정할 것이다.

$$\Pr(J = j \mid \boldsymbol{t}, Z)$$

Rieter(2005)는 의사결정론에 기반한 노출위험 측정에 관한 이론을 제시하고, 실제 자료를 가지고 개체 유형별 노출위험 측정을 통하여 해당 자료의 다양한 매스킹 처리 효과를 평가하였다.

지금까지 매스킹 처리된 자료의 노출위험을 측정하는 두 접근 방식을 간략히 살펴보았다. 첫 번째 방식은 키변수의 유일성에 기반한 노출위험 측정이며, 두 번째 방

식은 침입자의 정보를 조건부 매칭 확률에 반영한 노출위험 측정이다. 두 번째 방식을 침입자 의사결정 기반 노출위험 측정이라 부르도록 하겠다.

재현자료의 노출위험을 측정할 때는, 첫 번째 유일성 기반 노출위험 측정 방식을 사용하는 것은 적절하지 않다. 원자료의 분포를 따르도록 새로 생성한 자료에서 유일한 개체수를 논하는 것은 의미가 없기 때문이다. 다만 재현자료를 생성하는 과정에서 원자료에서 발생한 유일한 레코드들이 동일하게 생성되지 않도록 적절한 조치를 취하여 노출위험을 제어할 수도 있다(유성준과 박나리, 2020). 반면 매스킹 처리된 자료 평가에 잘 사용되지 않던 두 번째의 침입자 의사결정 기반 노출위험 측정 방식이 재현자료의 노출위험 측정을 위한 주요 아이디어로 활용되고 있다.

모수적 모형을 통해 생성한 완전 재현자료의 경우 어떠한 의미에서든 노출위험을 측정하는 것이 의미가 없다고 판단할 수도 있다. 그럼에도 불구하고 우연히 동일한 값들이 생성되고 식별될 위험성을 배제할 수는 없으므로 노출위험을 사후적으로라도 점검할 필요는 있다. 한편 부분 재현자료의 경우에는 노출위험을 반드시 점검해야 한다. 생성된 값들이 우연히 참값과 동일하여 다른 값들과 연결될 수 있기 때문이다. 이러한 상황을 감안하면 외부인의 정보를 고려하여 노출위험을 측정하고자 할 때, 의사결정론 기반 노출위험 측도를 고려하는 것은 자연스럽다. 다만 매칭된 개체가 여러 개 있을 수 있으므로 기존 방식에 더하여 추가적인 측도가 필요하다. 재현자료의 노출위험 측정을 이해하기 위하여 침입자 의사결정론 기반 노출위험 측정과 관련된 다음 예를 살펴보도록 한다.

다음 <표 3-4>와 같은 변수로 구성된 데이터를 제공하는 상황을 생각해 보자 (Drechsler & Reiter, 2010). 이 자료는 미국의 경제활동인구조사(US Current Population Survey, 2000.3) 공공이용파일의 주요 항목이며, 처음 4개 항목(X, A, M, R)은 센서스 자료에서도 제공하는 변수들이다. 따라서 정보 노출을 가정하는 상황은 외부인이 특정 개체에 대하여 처음 4개 변수의 참값을 센서스 자료에서 확인하는 경우이다. 이를 기반으로 공표된 CPS 재현자료에서 해당 개체에 대한 다른 정보를 노출시키고자 할 때 노출위험 측정을 고려하자. 참고로 이 예에서 언급하고 있는 CPS 재현자료는 성별(S)은 그대로 두고, 나이(A), 혼인상태(M), 인종(R)의 순서로 자료를 생성하였다. 노출위험 측정은 이 4개 변수만을 고려하도록 한다.

<표 3-4> Drechsler와 Reiter(2010)에 나타난 미국 CPS 자료의 변수 예제

변수	라벨	범위
Sex	X	Male, Female
Age (years)	A	0-90
Marital status	M	7 categories, coded 1-7
Race	R	White, Black, American Indian, Asian
Highest attained education level	E	16 categories, coded 31-46
Number of people in household	H	1-16
Number of people in household under age 18	Y	0, 1-11
Household property taxes ($)	P	0, 1-99997
Social security payments ($)	S	0, 1-50000
Household income ($)	I	-21011-768742

표적 $\boldsymbol{t}$에 대하여 노출위험은 다음과 같이 정의한다. 이때 M은 모형에 관한 모든 정보를 나타내며 D_{syn}은 m개 재현자료 세트를 표현한다. 이 확률을 계산하는 문제는 박민정(2014) 및 Reiter와 Mitra(2009)를 참조할 수 있고, 변수 유형별로 확률 추정을 위한 모형은 Hu et al.(2014) 등을 통하여 확인할 수 있다.

$$\Pr(J=j \mid \boldsymbol{t}, D_{syn}, M)$$

이제 재현자료의 노출위험 측정 방법으로 매칭 비율(true matching rate) 및 오매칭 비율(false matching rate)을 이용할 수 있다. 먼저 주어진 표적 $\boldsymbol{t}$에 대하여 재현자료 n개의 개체마다 위의 매칭 확률을 계산하고 그 확률이 가장 큰 개체를 찾는다. 이때 동일한 확률을 가지는 개체가 여러 개 나타날 수 있고 그 개수를 c_j라 하자. c_j의 값이 1이면 재현자료에서 단 하나의 개체가 표적 $\boldsymbol{t}$와 동일하여 바로 매칭이 되는 것을 의미하며, 1보다 큰 값이면 표적 $\boldsymbol{t}$의 후보가 재현자료에서 여러 개 존재하는 것을 의미한다. 이 c_j개의 개체 중에서 $\boldsymbol{t}$가 실제로 있으면 1, 없으면 0의 값을 가지는 변수 I_j를 고려하자. 또한 $c_j I_j = 1$이면 1의 값을, 아니면 0의 값을 가지는 변수 K_j, 반대로 $c_j(1-I_j) = 1$이면 1의 값을, 아니면 0의 값을 가지는 변수 F_j를 도입하자. K_j가 1이면 표적을 재현자료에서 식별해 내는데 성공한 것을 나타내고, F_j가 1이면 표적을 찾았으나 잘못 매칭한 것이다. 마지막으로 $c_j = 1$인 개체수를 s라고 하자. 이는 외부인이 재현자료에서 표적과 매칭에 성공한 레코드 개수를 의미한다. 이제 매칭 비율은 $\sum K_j/n$로, 오매칭 비율은 $\sum F_j/s$로 나타낼 수 있다.

Drechsler와 Reiter(2010)는 $n = 51{,}016$개의 레코드로 구성된 CPS 자료를 재현할

때, <표 3-4>의 처음 4개 변수를 고려하여 익명성이 1 또는 2인 개체 1,089개를 선택하였다. 이처럼 노출위험이 높다고 판단한 1,089개 개체들에 대하여 나이(A), 혼인상태(M), 인종(R) 변수들의 값을 생성하여 재현자료를 만들고, 매칭 비율과 오매칭 비율을 계산하였다. 매칭 비율은 5.7%인 반면, 오매칭 비율은 88%였다. 원본 센서스 자료에 대하여 동일한 방식으로 매칭 비율을 계산하면 50%이므로 재현자료를 이용하면 노출위험이 현저히 낮아진다는 것을 알 수 있다. 한편 원본 자료에 대해서는 오매칭이 발생하지 않는다. 참고로 McClure와 Retier(2016)는 평균제곱오차(MSE)의 형태로 노출위험을 계산할 수도 있음을 제시하였다.

한편 이론적 지식을 크게 사용하지 않고 실무적으로 용이하게 노출위험을 측정하는 방식도 제시된 바 있다. Loong et al.(2013)은 다음과 같은 측도를 제안하였다.

$$MXM = \sum_{i=1}^{n}\sum_{k=1}^{m} C_{ik}$$

$$EMR_i = \sum_{k=1}^{m} \frac{1}{F_{ik}} C_{ik}, \qquad EMR = \sum_{i=1}^{n} EMR_i$$

$$TMR_i = \sum_{k=1}^{m} W_{ik}, \qquad TMR = \sum_{i=1}^{n} TMR_i$$

첨자 i는 원자료의 레코드를 나타내며, j는 재현자료의 레코드를 나타낸다. 재현자료는 총 m개 세트를 생성하고, 첨자 k는 재현자료 세트를 나타낸다. 이제 표적 $\boldsymbol{t}_i$에 대하여 k번째 재현자료 세트에서 j번째 레코드가 같으면 1, 아니면 0의 값을 가지는 변수를 R_{ikj}라고 하자. 표적이 재현자료의 j번째 레코드와 일치하면 이를 $C_{ik} = R_{ikj}$라고 하자. 또한 k번째 재현자료 세트에서 표적과 일치하는 레코드의 개수를 F_{ik}라고 하면 $F_{ik} = \sum_{j=1}^{n} R_{ikj}$이다. 노출위험 측도 MXM(maximum number of matches)은 F_{ik}개의 레코드 중에서 올바르게 일치하는 개수를 모두 합한 것을 나타내며, EMR_i(expected match risk)는 F_{ik}개의 레코드 중에서 랜덤하게 선택하여 맞출 기대 리스크를 표현한다. W_{ik}는 $F_{ik} = 1$이고 $C_{ik} = 1$일 때 1, 아니면 0을 가지는 변수를 표현하여, TMR_i(true match risk)는 정확하고 유일하게 노출될 확률을 의미한다.

Loong et al.(2013)은 원자료에 비하여 재현자료에서 각종 노출위험 측도가 현저히 감소함을 나타내었다. 참고로 이러한 방식의 노출위험 측정은 생성하는 변수의 개수가 많을수록 낮아지는 특징을 가진다. 여러 변수가 매칭되는 경우를 고려해야 하므로 이는 당연한 현상이다.

제2절 정보손실 측정

정보손실 혹은 자료 유용성 측정에 관하여 박민정(2014)에서 자세히 다루었다. 이 절에서는 그 내용을 정리하고 최근 재현자료의 정보손실을 측정하기 위하여 소개된 몇몇 측도들을 설명한다.

정보손실은 매스킹 처리된 자료이든 재현자료이든 기본적으로 원자료와 달라진 정도를 일컫는다. 단순히 자료의 값이 얼마나 바뀌었는지를 측정할 수도 있고, 자료의 구조가 얼마나 변화되었는지를 측정할 수도 있다. 나아가 추론의 측면에서 정보손실을 측정할 수도 있다. 가장 단순한 정보손실 측도는 거리에 기반한 것이며, 흔히 사용되는 거리 측도로 평균제곱오차(mean square error), 평균절대오차(mean absolute error), 평균변동(mean variation) 등을 사용할 수 있다. 이러한 측도를 사용하여 자료의 값이 얼마나 바뀌었는지 측정하기 위한 식은 다음 <표 3-5>와 같이 정리될 수 있다. X는 자료값 행렬을 나타낸다. X는 원자료를 X'은 변형된 자료를 나타낸다. 매스킹 처리된 자료에 대하여 이러한 정보손실 측도를 사용하는 것은 적절할 수 있으나, 완전 재현자료에 대하여 이처럼 레코드를 일대일로 연결시켜 정보손실을 측정하는 것은 적절하지 않다.

<표 3-5> 자료 값의 변화에 대한 정보손실 측정

	평균제곱오차(MSE)	평균절대오차(MAE)	평균변동
$X-X'$	$\frac{\sum_{j=1}^{p}\sum_{i=1}^{n}(x_{ij}-x'_{ij})^2}{np}$	$\frac{\sum_{j=1}^{p}\sum_{i=1}^{n}\lvert x_{ij}-x'_{ij}\rvert}{np}$	$\frac{\sum_{j=1}^{p}\sum_{i=1}^{n}\frac{\lvert x_{ij}-x'_{ij}\rvert}{\lvert x_{ij}\rvert}}{np}$

한편 자료가 가지는 구조 혹은 분포의 변화를 측정하여 정보손실을 정의할 수도 있다. 거리 측도를 이용하여 이러한 정보손실을 측정하는 방식은 다음 <표 3-6>에 정리하였다. V는 공분산 행렬, R은 상관계수 행렬, RF는 각 변수와 주성분 사이의 상관계수 행렬, C는 각 변수별 첫 번째 주성분의 설명 비율 벡터, F는 주성분 행렬을 나타낸다. 이처럼 자료의 구조를 나타내는 다양한 성분을 거리를 이용하여 정보손실 측도로 사용할 수 있다.

<표 3-6> 자료 분포의 변화에 대한 정보손실 측정

	평균제곱오차(MSE)	평균절대오차(MAE)	평균변동
$V-V'$	$\frac{\sum_{j=1}^{p}\sum_{1\le i\le j}(v_{ij}-v'_{ij})^2}{\frac{p(p+1)}{2}}$	$\frac{\sum_{j=1}^{p}\sum_{1\le i\le j}\lvert v_{ij}-v'_{ij}\rvert}{\frac{p(p+1)}{2}}$	$\frac{\sum_{j=1}^{p}\sum_{1\le i\le j}\frac{\lvert v_{ij}-v'_{ij}\rvert}{\lvert v_{ij}\rvert}}{\frac{p(p+1)}{2}}$
$R-R'$	$\frac{\sum_{j=1}^{p}\sum_{1\le i\le j}(r_{ij}-r'_{ij})^2}{\frac{p(p-1)}{2}}$	$\frac{\sum_{j=1}^{p}\sum_{1\le i\le j}\lvert r_{ij}-r'_{ij}\rvert}{\frac{p(p-1)}{2}}$	$\frac{\sum_{j=1}^{p}\sum_{1\le i\le j}\frac{\lvert r_{ij}-r'_{ij}\rvert}{\lvert r_{ij}\rvert}}{\frac{p(p-1)}{2}}$
$RF-RF'$	$\frac{\sum_{j=1}^{p}w_j\sum_{i=1}^{p}(rf_{ij}-rf'_{ij})^2}{p^2}$	$\frac{\sum_{j=1}^{p}w_j\sum_{i=1}^{p}\lvert rf_{ij}-rf'_{ij}\rvert}{p^2}$	$\frac{\sum_{j=1}^{p}w_j\sum_{i=1}^{p}\frac{\lvert rf_{ij}-rf'_{ij}\rvert}{\lvert rf_{ij}\rvert}}{p^2}$
$C-C'$	$\frac{\sum_{i=1}^{p}(c_i-c'_i)^2}{p}$	$\frac{\sum_{i=1}^{p}\lvert c_i-c'_i\rvert}{p}$	$\frac{\sum_{i=1}^{p}\frac{\lvert c_i-c'_i\rvert}{\lvert c_i\rvert}}{p}$
$F-F'$	$\frac{\sum_{j=1}^{p}w_j\sum_{i=1}^{p}(f_{ij}-f'_{ij})^2}{p^2}$	$\frac{\sum_{j=1}^{p}w_j\sum_{i=1}^{p}\lvert f_{ij}-f'_{ij}\rvert}{p^2}$	$\frac{\sum_{j=1}^{p}w_j\sum_{i=1}^{p}\frac{\lvert f_{ij}-f'_{ij}\rvert}{\lvert f_{ij}\rvert}}{p^2}$

한편 확률밀도 함수를 추정하고 쿨백-라이블러 거리(Kullback-Liebler divergence)를 측정하여 자료유용성 측도로 삼을 수 있다. 그러나 이러한 KL측도는 자료가 다변량 정규분포를 따르지 않고, 변수의 개수가 많으면 계산하기 어려운 단점을 가진다(Karr et al., 2006).

$$d_{KL}(X', X) = \int log\left[\hat{f}_{X'} / \hat{f}_{X}\right]\hat{f}_{X'}$$

또한 경험적 분포를 이용하여 정보손실을 다음과 같이 정의할 수 있다. N개 레코드를 가지는 자료 X의 p개 변수들에 대한 경험적 분포가 아래와 같고

$$S_X(x_1, \dots, x_p) = \frac{1}{N}\sum_{i=1}^{N} I(x_{i1} \le x_1, \dots, x_{ip} \le x_p),$$

변형 전후의 자료를 합한 자료를 $T = (X', X)$라고 할 때 다음과 같은 측도들을 생각해 볼 수 있다(Woo et al., 2009). 이는 두 경험적 분포 사이의 절대 차이의 최대값 혹은 차이 제곱의 평균을 측정하는 것이나, 분포의 차이를 감지하는데 변별력이

약한 것(low power)으로 알려져 있다.

$$U_m = max_{1 \le i \le N_T} |S_X(\vec{t}_i) - S_{X'}(\vec{t}_i)|$$

$$U_s = \frac{1}{N_T} \sum_{i=1}^{N_T} |S_X(\vec{t}_i) - S_{X'}(\vec{t}_i)|^2$$

정보손실을 측정할 때 자료나 자료 구조의 변화를 측정하는 대신에 자료를 분석한 결과를 비교하는 방식도 있으며, 이러한 유형의 측도가 오히려 널리 활용되고 있다. 이에 관하여 주요 통계량에 대하여 신뢰구간의 중복을 측정하는 방식을 소개한다. 한편 성향점수를 비교하거나 군집분석을 활용하여 정보손실을 측정하는 방식도 참고할 수 있다.

먼저 주요 통계량에 대하여 신뢰구간의 중복(Confidence Interval Overlap, IO)을 측정하여 정보손실 측도로 사용하는 경우를 살펴보겠다. 어떤 분석을 수행하고 얻고자 하는 통계량이 θ라고 하자. 원자료 Y를 사용하여 얻은 통계값은 $\hat{\theta}^Y$이고, 재현자료 Z를 사용하여 얻은 통계값은 $\hat{\theta}^Z$, 각각의 신뢰구간은 (L^Y, U^Y) 및 (L^Z, U^Z)라고 하자. 이제 확률적인 신뢰구간 중복 IO를 다음과 같이 정의할 수 있다. 첫째 항은 재현자료 Z로부터 얻어지는 θ^Z의 사후분포에 원자료 Y에서 얻은 신뢰구간 (L^Y, U^Y)를 적용할 때의 누적확률을, 두 번째 항은 그 반대를 의미한다. 원자료와 재현자료에서 추정량 및 추정량의 사후분포들이 서로 유사하다면, 두 항 모두 의도한 신뢰구간의 길이와 유사하게 된다. 만약 신뢰구간의 길이를 0.95라고 한다면, 두 신뢰구간이 완전히 겹치면 I는 0.95의 값을, 전혀 겹치지 않으면 0의 값을 가지게 된다. 신뢰구간 중복을 이용한 정보손실 측정은 다수의 문헌에서 활용되고 있다.

$$I = \frac{1}{2}\left[\int_{L^Y}^{U^Y} f^Z(t)\,dt + \int_{L^Z}^{L^Z} f^Y(t)dt\right]$$

이상으로 노출위험 및 정보손실 관련 측도를 살펴보았다. 각종 문헌에서 제안되어 있는 노출위험 측도나 정보손실 측도는 종류도 다양하고 그 숫자도 상당히 많다. 노출위험과 정보손실 측도에 관한 연구는 지속적으로 이루어지고 있지만, 명확한 해법이 주어져 있지는 않은 상황이다. 때문에 현재로서는 주어진 자료마다 적절한 측도를 사용하여 생성한 재현자료를 평가할 필요가 있다. 현재 재현자료와 관련된 대다수 문헌에서는 노출위험이 높은 레코드가 생성되지 않도록 재현과정을 제어하면서, 정보손실을 측정하여 재현자료의 유용성을 설명하는 경향이 있다.

제 4 장

재현자료 생성 국내외 사례

이 장에서는 실제 재현자료를 생성하여 활용하고 있는 국내외의 사례를 살펴본다. 이를 통해 재현자료 생성 방법과 더불어 재현자료의 유용성 및 노출위험 측정, 평가 현황을 살펴보고자 한다.

해외에서는 개인 또는 사업체의 민감한 정보가 포함된 자료의 경우 정보의 노출 위험을 제어한 상태에서 실제 분석에 활용할 수 있도록 재현자료 생성하여 제공하고 있다. 또한 재현자료를 활용하여 분석한 결과를 검증해 주는 분석 체계도 함께 운영하고 있다. 구체적인 사례로 국가 차원에서 재현자료를 생성하여 제공하고 있는 미국의 SIPP Synthetic Beta(SSB), Synthetic Longitudinal Business Database(SynLBD), 영국의 SYnthetic Data Estimation for the UK LongitudinaL Studies(SYLLS)를 먼저 살펴본 다음에, 독일의 IAB Establishment Panel을 살펴본다.

다음으로 국내 사례를 살펴본다. 우리나라는 재현자료를 도입하는 초기 단계이며, 통계청을 중심으로 재현자료 방법론에 관한 연구가 진행되어왔다(박민정과 김정연, 2017). 이후 신용정보원의 개인신용정보(대출, 연체, 신용카드 개설 등) 재현자료 시범 구축사업(유성준과 박나리, 2020), 한국정보화진흥원의 코리아크레딧뷰로(KCB)를 이용한 제주도 주민정보(직업, 소득, 소비 등) 재현자료 시범 생성 사업이 이루어진 바 있다. 이 중 신용정보원 재현자료는 개인신용정보 표본DB, 개인신용정보 교육용 DB로 재현자료를 제공하고 있다. 또한 올해 통계청 통계데이터센터 일부 DB를 재현자료로 구현하는 프로젝트가 진행되고 있다. 이 장에서는 기존의 국내 사례로서 그 결과가 발표되어 있는 한국신용정보원의 개인신용정보 재현자료에 대해서 살펴본다.

제1절 미국 SSB[2)]

미국 Census Bureau의 SIPP Synthetic Beta(SSB)는 Survey of Income and Program Participation(SIPP) 조사자료와 Social Security Administration(SSA) 및 Internal Revenue

2) Benedetto et al.(2018), The Creation and Use of the SIPP Synthetic Beta v7.0 재정리 및 인용

Service(IRS) 행정자료를 통합한 자료에 대한 재현자료를 제공한다.

SSB는 현재까지 공개된 재현자료 중에서 가장 광범위하게 검증된 자료로 원자료와 재현자료의 주요 변수 통계치가 매우 유사한 것으로 알려져 있다. SSB는 현재 7.0 버전(2018)까지 공개되어 있으며, 7.0 버전은 기존 6.0 버전(2015)에서 몇 가지 사항이 변경되었다. 7.0 버전에서는 결측치의 패턴이 함께 재현자료로 생성되며, 공개자료의 범위가 2014년까지로 확대되었다. 또한 18세 미만 자녀와 모친을 연계하고 산업과 직업 범주가 세분화되었다.

SSB는 완전 재현자료(fully synthetic data)를 제공하고 있으며, 재현자료는 141개 변수로 이루어져 있다.

Variable Name	Label	Codebook
afdc_Y_M	Indicator for receipt of AFDC or TANF benefits	SIPP Synthetic Beta v7
afdcamt_Y_M	Amount of AFDC received	SIPP Synthetic Beta v7
birthdate	Date of Birth	SIPP Synthetic Beta v7
brthmn_sipp	Month of Birth - SIPP	SIPP Synthetic Beta v7
brthyr_sipp	Year of Birth - SIPP	SIPP Synthetic Beta v7
current_enroll_coll	Currently Enrolled in College	SIPP Synthetic Beta v7
current_enroll_hs	Currently Enrolled in HS (or less)	SIPP Synthetic Beta v7
date_enter_sipp	Wave Entered SIPP	SIPP Synthetic Beta v7
db_pension	Defined Benefit Pension Plan	SIPP Synthetic Beta v7
dc_pension	Defined Contribution Pension Plan	SIPP Synthetic Beta v7

자료: https://www2.ncrn.cornell.edu/ced2ar-web/all

<그림 4-1> SSB 변수 일부

1. 재현자료 생성

SSB를 생성하기 위해서는 재현 대상이 되는 각 변수들에 대하여 사후예측분포(posterior prediction distribution)를 계산한다. 이때 순차적 회귀(sequential regression) 방식을 사용한다. 이러한 순차적 회귀모형을 활용하기 위해서는 1) 분석 모델(OLS, logistic, Bayesian bootstrap), 2) 변수들 간의 인과관계(parent-child relationship), 3) 변수들의 제약사항, 4) 모델에 사용할 변수의 4가지 사항을 결정해야 한다.

참고로 재현자료 생성과 관련하여 SSB 7.0 버전에서 변경된 사항은 결측치와 같은 불완전 데이터도 함께 재현한다는 것이다. 기존에는 4개 다중대체 데이터 셋

(multiply imputed dataset)을 생성하고, 각 다중 대체 데이터 셋에 대하여 4개 재현자료를 생성하여 총 16개 재현자료를 제공하였다. 7.0 버전에서는 결측치를 포함하는 snapshot을 생성하고 이에 대한 4개 재현자료를 생성하여 제공한다.

2. 재현자료 평가

SSB의 유용성 측정에는 로지스틱 회귀(logistic regression)를 이용한다. 원자료와 재현자료를 합친 후 원자료는 0, 재현자료는 1로 하는 종속변수를 생성하고 예측모형의 mean squared error(MSE)를 계산하여 활용하고 있다고 설명한다. 측정한 값은 0에서 1 사이의 값으로 재현자료의 유용성이 가장 좋지 않은 경우에 0, 가장 좋은 경우에 1의 값을 가진다. 하지만 SSB에 대한 측정 결과를 별도로 제시하고 있지는 않다.

SSB의 노출위험 측정은 유클리디안 거리(Euclidean distance)를 이용한다. 유클리디안 거리를 최소로 하는 원자료와 재현자료의 가장 근사(closet)한 재현자료 개체를 찾는다. 그리고 root mean squared error(RMSE), sample standard deviation(STDDEV)을 계산한 후 최종적으로 RMSE/STDDEV를 계산하여 측정한다. 측정한 값은 0에서 1 사이의 값으로 노출위험이 가장 큰 경우에 0, 가장 작은 경우에 1의 값을 가진다. 아래 결과를 살펴보면 노출위험이 높지 않게 재현자료가 생성된 것을 확인할 수 있다.

Select Quantiles						
p10	p25	p50	p75	p90	p95	p99
$ -	$ -	$ 17,000.00	$ 111,000.00	$ 300,000.00	$ 499,000.00	$ 1,134,000.00

Wealth Category	n	RMSE
<0	18000	$ 35,600.00
0	240000	$ 15,320.00
<p50	131000	$ 75,560.00
<p75	203000	$ 113,300.00
<p90	118000	$ 173,400.00
<p95	38000	$ 353,600.00
<p99	28500	$ 463,100.00
>=p99	6500	$ 5,586,000.00

Overall		
RMSE	Std. Dev.	Ratio
$ 521,000.00	$ 548,400.00	0.95

자료: Benedetto et al.(2018), The Creation and Use of the SIPP Synthetic Beta v7.0

<그림 4-2> SSB: 총 순자산 노출위험 측정 결과 예

제2절 미국 SynLBD

미국 Census Bureau의 Synthetic Longitudinal Business Database(SynLBD)는 미국의 사업체(business establishments)를 대상으로 하는 센서스에 대한 재현자료를 제공한다. 원자료는 1976년부터 2009년까지의 약 2,400만 개 사업체에 대한 센서스 자료이며, 샘플링에 기반을 둔 일반 서베이 자료보다 노출위험이 크다고 할 수 있다. SynLBD는 이러한 원자료에서 산업 코드(industry code) 등을 제외한 변수에 대하여 생성한 재현자료이다.

<표 4-1> SynLBD 변수

Name	Notation	Description	Action
ID	-	Unique Random Number for Establishment	created
County	x_1	Geographic Location	not released
SIC	x_2	Industry Code	unmodied
NAICS	x_2	Industry Code	unmodied
Firstyear	y_1	First Year Establishment is Observed	synthesized
Lastyear	y_2	Last Year Establishment is Observed	synthesized
Year	-	Year dating of annual variables	created
Multiunit	$y_3^{(t)}$	Multiunit Status(annual)	synthesized
Employment	$y_4^{(t)}$	Number Employees on March 12(annual)	synthesized
Payroll	$y_5^{(t)}$	Total Payroll(annual)	synthesized
Firm ID	$y_6^{(t)}$	Unique firm identier(annual)	synthesized

자료: Kinney et al.(2014), Improving the Synthetic Longitudinal Business Database
주: t는 연도를 의미함

1. 재현자료 생성[3)]

SynLBD 생성 시에는 각 변수들의 특성 및 변수들 간의 인과관계 등을 고려한다. 예를 들어, Firstyear와 Lastyear는 각각 왼쪽과 오른쪽으로 절단된 변수로 재현자료에도 이와 같은 특성이 반영되어야 한다. 그리고 employment는 Firstyear 이전과 Lastyear

3) Kinney et al.(2014), Improving the Synthetic Longitudinal Business Database 재정리 및 인용

이후에는 값들이 존재하지 않아야 한다.

SynLBD는 종단(longitudinal)자료의 특성을 고려하여 변수들을 정해진 순서대로 생성한다. 변수 생성에 이용하는 모형은 Dirichlet-multinomial과 CART(Classification And Regression Trees)이다.

[1단계] Firstyear(y_1) : Dirichlet-multinomial 모형을 이용하여 $f(y_1|x_1, x_2)$ 생성

[2단계] Lastyear(y_2) : Dirichlet-multinomial 모형을 이용하여 $f(y_2|y_1, x_1, x_2)$ 생성

[3단계] Multiunit status($y^{(t)_3}$) : Dirichlet-multinomial 모형을 이용하여
$f(y_3^{(t)}|y_2, y_1, x_1, x_2)$ 생성

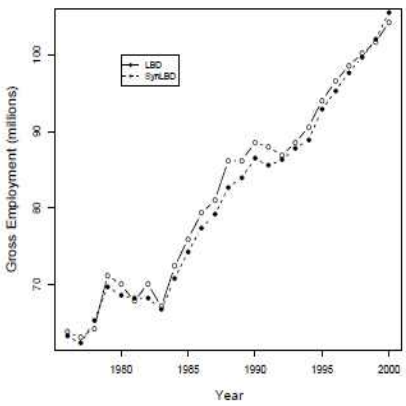
[4단계] Employment($y^{(t)_4}$) and Payroll($y^{(t)_5}$) : CART 모형을 이용하여 생성

Employment와 Payroll은 1.0 버전에서는 normal liner regression 모형, 2.0 버전에서는 CART 모형을 사용하여 생성한다.

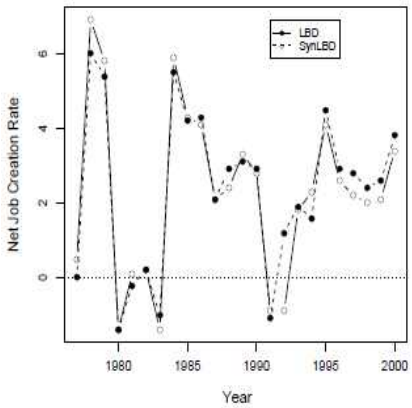
2. 재현자료 평가[4)]

유용성 측면에서는 SynLBD가 원자료와 재현자료의 분포가 얼마나 유사한지로 측정하며, 평균과 상관관계를 함께 살펴본다.

[총 종사자 수]



[순일자리 창출률]



자료: Kinney et al.(2011), Towards Unrestricted Public Use Business Microdata: The Synthetic Longitudinal Business Database

<그림 4-3> SynLBD: 원자료와 재현자료의 총 종사자 수, 순일자리 창출률 분포 비교

4) Kinney et al.(2011), Towards Unrestricted Public Use Business Microdata: The Synthetic Longitudinal Business Database 재정리 및 인용

<그림 4-3>에서 원자료와 재현자료의 총 종사자 수 분포는 유사한 형태를 보인다. 1976년 ~ 2000년 기간의 평균 불일치율은 1.3%이다. 또한 1976년 ~ 2000년 기간의 사업체 설립률은 원자료 11.14%, 재현자료 11.13%이며, 가중 평균 입직률(average employment weighted entry rate)은 원자료 3.16%, 재현자료 2.85%이다. 이상의 결과는 원자료와 재현자료가 유사함을 나타낸다.

한편 종단(longitudinal)자료의 특성을 고려하여 dynamics에 대한 분포도 함께 확인할 수 있다. LBD에서 중요한 통계량인 순일자리 창출률(net job creation rate)[5]을 살펴보면, 원자료와 재현자료의 연도별 분포가 유사하게 나타난다는 것을 알 수 있다.

LBD는 County Business Patterns(CBP) 프로그램[6]에 의해 지역과 산업별 통계표를 제공하고 있다. 따라서 SynLBD는 식별(identification) 노출위험보다 속성(attribute) 노출위험을 더 중요하게 고려할 필요가 있다.

이산형 변수인 Firstyear의 노출위험은 각 산업에 대해서 아래와 같이 계산하며, Lastyear도 동일하다. Firstyear과 Lastyear은 모든 산업에서 노출위험이 매우 낮은 것으로 나타난다.

$$P(LBD\ \ Firstyear = yy \mid \ synLBD\ \ Firstyear = yy)$$

<표 4-2> SynLBD: $P(LBD\ \ Firstyear = yy \mid \ synLBD\ \ Firstyear = yy)$ 결과 일부

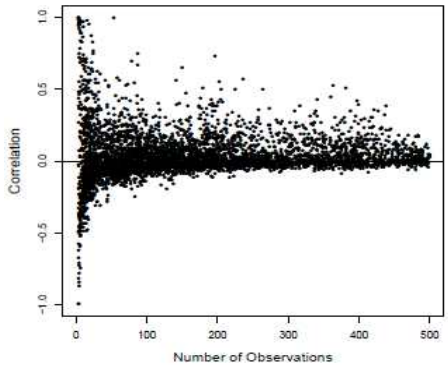
Firstyear		$P(LBD\,Firstyear = yy \mid synLBD\,Firstyear = yy)$		
재현자료	원자료	최소값	평균	최대값
1975	1975	1.52	25.41	88.89
1976	1976	0.12	5.12	75.00
1977	1977	0.43	5.09	71.43
1978	1978	0.46	3.65	16.22
1979	1979	0.27	3.90	50.00
1980	1980	0.39	3.46	25.00
1981	1981	0.26	3.91	50.00
1982	1985	0.36	3.69	50.00
1983	1983	0.39	4.10	50.00
1984	1984	0.69	3.79	19.30
1985	1985	0.15	3.75	23.73

자료: Kinney et al.(2011), Towards Unrestricted Public Use Business Microdata: The Synthetic Longitudinal Business Database

5) 순일자리 창출률(net job creation rate)은 일자리 창출률에서 일자리 소멸률을 뺀 증가분을 의미한다.

6) County Business Patterns(CBP) 프로그램은 매년 상세한 지리 및 산업 수준에서 U.S., Puerto RICO and Island 지역 사업체의 사업체 수, 고용, 임금 등의 통계를 매년 제공한다.

연속형 변수 Employment, Payroll은 연도×산업별로 원자료와 재현자료의 피어슨 상관관계(Pearson correlation)를 측정한다. 각 변수별 원자료와 재현자료의 상관관계가 높다는 것은 자료의 노출위험이 크다는 것을 의미한다. 측정 결과 상관계수가 균일하게(uniformly) 작은 것으로 나타난다. 여기서 연도×산업별 사업체 수가 적은 경우에는 높은 상관관계로 나타나므로 주의가 필요할 것이다.



자료: Kinney et al.(2011)

<그림 4-4> SynLBD: 원자료와 재현자료의 Payroll 변수 피어슨 상관관계

제3절 독일 IAB Establishment Panel[7)]

독일 Institute of Employment Research(IAB)의 IAB Establishment Panel은 매년 6월 30일 기준으로 German Social Security Data(GSSD)에 등록되어 있는 근로자가 있는 약 16,000개 사업체를 대상으로 하고 있다. IAB Establishment Panel에는 건강보험, 연금, 실업보험 등의 민감한 정보가 포함되어 있어 IAB Establishment Panel의 부분 재현자료로 IAB Establishment Panel Scientific Use File(synthetic dataset)을 제공하고 있다.

1. 재현자료 생성

IAB Establishment Panel Scientific Use File은 미국의 SIPP와 SynLBD 사례를 참고하여 생성하고 있다. 재현자료 생성에 순차회귀 다중대체(SRMI) 방법을 이용한다. 재현할 변수가 연속형 변수인 경우에는 linear regression model, 이항 변수인 경우에는 logistic model, 범주형 변수인 경우에는 Dirichlet/multinomal model을 사용한다.

재현자료 생성 시에 어떠한 변수를 재현할지가 중요한데, IAB Establishment Panel

7) Drechsler, J.(2009), Synthetic Datasets for the German IAB Establishment Panel 재정리 및 인용

Scientific Use File에서는 사업체 규모, 지역, 산업코드와 같은 주요 변수와 총 매출액, 정부로부터 받는 지원금 등 민감한 변수를 재현한다. 재현자료는 5개의 다중 대체 데이터 셋(MID)을 생성하고, 각 다중 대체 데이터 셋에 대하여 5개의 재현자료를 생성하여 총 25개(5×5)의 재현자료 셋을 제공한다.

2. 재현자료 평가

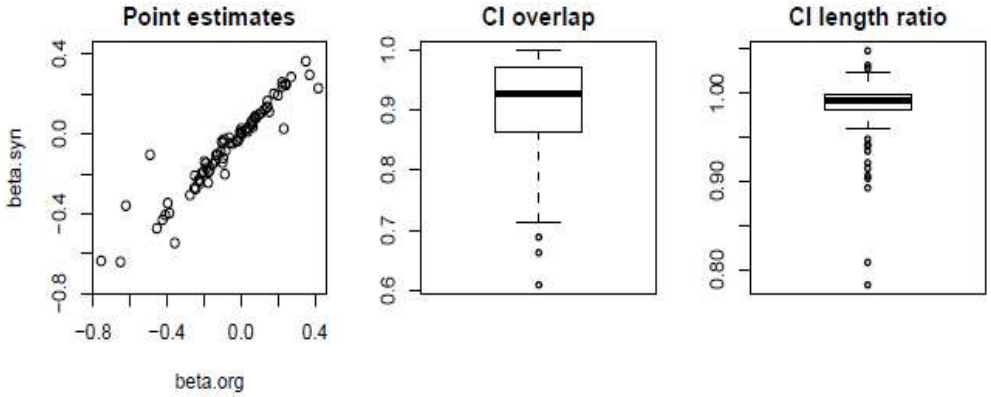
IAB Establishment Panel Scientific Use File의 유용성 평가에는 2개의 회귀모형을 사용한다. 먼저 사업체 규모, 인사 구조 정보 등을 설명변수, 시간제 종사자 고용 여부(예/아니오)를 종속변수로 하는 프로빗 모형(probit regression)을 활용하여 원자료와 재현자료의 점 추정량과 z-scores를 비교한다. 원자료와 재현자료의 점 추정량은 유사하며, 점 추정량의 신뢰구간의 중복(confidence interval overlap)도 90% 이상으로 나타난다. 그리고 원자료와 재현자료의 z-scores도 유사하며, 원자료의 95% 신뢰구간과 재현자료의 95% 신뢰구간을 비율로 나타낸 confidence interval length ratio도 95% 이상이다.

<표 4-3> IAB Establishment Panel Scientific Use File: 원자료와 재현자료 비교1

	original data	synth. data	CI over-lap	z-score org.	z-score syn	CI length ratio
Intercept	-0.809	-0.752	0.87	-7.23	-6.85	0.99
5-10 employees	0.443	0.437	0.97	8.52	7.99	1.06
10-20 employees	0.658	0.636	0.90	11.03	10.88	0.98
20-50 employees	0.797	0.785	0.95	13.02	12.36	1.04
100-200 employees	0.892	0.908	0.96	9.23	9.48	0.99
200-500 employees	1.131	1.125	0.99	9.99	9.87	1.01
>500 employees	1.668	1.641	0.97	8.22	8.33	0.97
growth in employment exp.	0.010	0.006	0.98	0.18	0.12	0.99
decrease in emp. expected	0.087	0.100	0.96	1.11	1.27	1.00
share of female workers	1.449	1.366	0.73	17.63	18.71	0.89
sh. of emp. with uni. degree	0.319	0.368	0.91	2.18	2.59	0.97
sh. of low qualified workers	1.123	1.148	0.93	12.17	11.87	1.05
sh. of temporary employees	-0.327	-0.138	0.75	-1.74	-0.71	1.05
share of agency workers	-0.746	-0.856	0.88	-3.09	-4.24	0.84
empl. in the last 6 mths	0.394	0.369	0.87	8.33	7.82	1.00
dismissal in the last 6 mths	0.294	0.279	0.92	6.38	6.03	1.00
foreign ownership	-0.113	-0.117	0.99	-1.33	-1.38	0.99
good/very good profitability	0.029	0.033	0.98	0.72	0.82	0.99
salary above coll. wage agr.	0.020	0.031	0.95	0.35	0.54	0.99
collective wage agreement	0.016	0.007	0.95	0.31	0.13	0.97

자료: Drechsler, J.(2009), Synthetic Datasets for the German IAB Establishment Panel

두 번째는 39개의 설명변수와 향후 예상되는 종사자 수 변화(증가, 변화 없음, 감소)를 종속변수로 하는 ordered probit regression을 이용한다. <그림 4-5>의 원자료와 재현자료의 boxplot이 45도 대각선에 가까워 원자료와 재현자료의 점 추정량이 매우 유사한 것으로 나타난다. confidence interval overlap, confidence interval length ratio도 높은 수치를 보인다.



자료: Drechsler, J.(2009), Synthetic Datasets for the German IAB Establishment Panel

<그림 4-5> IAB Establishment Panel Scientific Use File: 원자료와 재현자료 비교2

IAB Establishment Panel Scientific Use File의 노출위험은 Drechsler와 Reiter(2008) 방법을 적용하여 true match risk을 계산하여 측정한다. 이를 간단히 설명하면 다음과 같다. $D = (D^{(1)}, \cdots, D^{(m)})$는 m개가 공개된 재현자료이며, M은 재현자료로부터 공개된 정보이다. t라는 개체를 $j(j = 1, \cdots, s)$개의 개체를 가지는 D부터 찾는다고 하면, $P(J = j \mid t, D, M)$을 최대로 하는 $j(\hat{J}$라고 정의한다)를 t에 대응하는 실제 개체로 선택하게 된다. 이때, c_j는 $\hat{J}$에 해당하는 개체의 수, I_j은 $\hat{J}$에서 실제 포함되어 있는 경우에 1, 그렇지 않은 경우에 0으로 정의한다. true match risk는 K_j의 합으로 정의하며, 여기서 $c_j I_j$= 1인 경우 $K_j = 1$, 그렇지 않은 경우 $K_j = 0$이다.

제4절 영국 SYLLS

UK Economic and Social Research Council의 UK Longitudinal Studies는 The England and Wales Longitudinal Study(ONS LS), Scottish Longitudinal Study(SLS), Northern Ireland Longitudinal Study(NILS)를 통합한 자료로 인구조사와 출생, 사망, 결혼, 암 등록 등의 행정자료를 포함하고 있다.

ONS LS, SLS, NILS는 자료의 활용 가치는 높지만 개별 정보 노출위험 또한 높아 센서스법에 의하여 100년 후에 공공 이용이 가능하다. Synthetic Data Estimation for the UK LongitudinaL Studies(SYLLS)는 ONS LS, SLS, NILS의 자료 활용을 위해 시작되었다.

SYLLS는 사용자의 자료 이용 목적에 따라 개별 맞춤형 재현자료(bespoke synthetic data)를 제공하고 있다. 그리고 개별 맞춤형 재현자료를 제공하기 위해서 R 패키지인 'synthpop'을 개발하여 활용하고 있다. SYLLS는 개별 맞춤형 재현자료를 제공하고 있으므로, 재현자료의 생성방법과 생성한 재현자료의 평가에 관한 내용보다는 synthpop을 이용하여 재현자료를 어떻게 생성하고 평가하는지를 중심으로 살펴본다.

1. 재현자료 생성[8)]

SYLLS는 재현자료 생성에 순차회귀 다중대체(SRMI, sequential regression and multiple imputation) 방법을 이용한다. 각 변수들에 대한 회귀모형으로 classification and regression trees(CART)를 기본 모형으로 사용하고 있다.

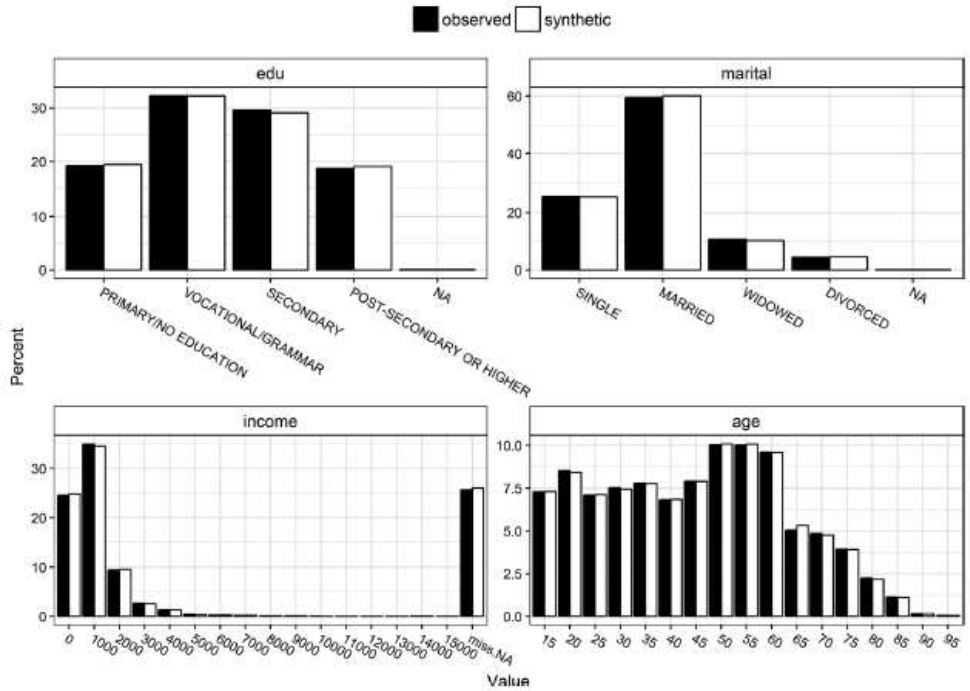
synthpop에서는 syn() 함수를 이용하여 재현자료를 생성한다. method argument를 이용하여 모형을 지정할 수 있으며, CART 모형은 디폴트이다. 재현자료는 자료의 왼쪽에서 오른쪽 변수 순으로 생성되는데, visit.sequence argument를 이용하여 재현되는 변수의 순서를 정하거나 변수 중에서 일부만 생성할 수도 있다. drop.not.used argument를 이용하면 재현하지 않는 변수를 재현자료에 포함할지 여부를 결정할 수 있는데 재현하지 않는 변수를 재현자료에서 제외할 경우에 drop.not.used = TRUE를 이용한다.

예시: syn(data = ods, visit.sequence = c(1, 2, 6, 4, 3), drop.not.used = TRUE)

8) Nowok et al.(2016), synthpop: Bespoke Creation of Synthetic Data in R 재정리 및 인용

2. 재현자료 평가[9)]

synthpop에서 재현자료의 유용성 평가에는 compare(), glm,synds() 함수를 이용한다. compare() 함수는 원자료와 재현자료의 분포를 비교하며, glm,synds() 함수는 logistic regression 모형을 원자료와 재현자료에 각각 적용한 결과를 비교한다.



자료: Nowok et al.(2017), Providing bespoke synthetic data for the UK Longitudinal Studies and other sensitive data with the synthpop package for R

<그림 4-6> SYLLS: 원자료와 재현자료 분포 비교

SYLLS는 자료의 노출위험에 대해서 크게 언급하고 있지 않다. synthpop에서도 노출위험을 측정할 수 있는 별도의 기능을 제공하고 있지 않다. 이는 SYLLS이 개별 맞춤형 재현자료를 제공한다는 점이 하나의 이유일 것이다.

지금까지 논의한 각 해외 사례의 특징을 정리하면 다음 <표 4-4>와 같다.

9) Nowok et al.(2017), Providing bespoke synthetic data for the UK Longitudinal Studies and other sensitive data with the synthpop package for R 재정리 및 인용

<표 4-4> 재현자료 생성 해외 사례 특징 비교

사례	자료의 특징	세트의 개수	재현 모형	평가		기타
				노출위험 측정	정보손실 측정	
미국 SSB ver.7.0(2018)	141개 변수 자녀와 모친 연계 산업, 직업 세분화	다중대체 4개 세트 작성 후 각 세트별 4개 재현자료 생성, 총 16개 세트 ↓ 결측치 패턴 제공 및 총 4개 세트	순차회귀 (logistic, Bayesian bootstrap 등) 변수별 인과관계 및 제약사항 반영	가장 가까운 원자료와의 거리를 이용	로지스틱 회귀를 이용한 예측모형의 MSE	-
미국 SynLBD	다년간의 사업체 전수조사 자료 (종단 자료) 산업 코드 등은 제외하고 생성	-	순차회귀 (Dirichlet-multinomial, CART 등) 변수별 인과관계 및 제약사항 반영	동일한 Firstyear에 대한 조건부 확률, 상관관계 등 이용	분포의 유사성 평균 및 상관관계	-
독일 IAB (노동 패널 자료)	민감한 항목을 재현	다중대체 5개 세트 작성 후 각 세트별 5개 재현자료 생성, 총 25개 세트	순차회귀 (Dirichlet-multinomial, logistic 등)	매칭 위험 등 이용	프로빗 회귀를 이용한 추정량 비교, 신뢰구간 중복 등	미국 연구진과 밀접한 네트워크
영국 SYLLS	법적으로 100년 후 이용 가능한 자료	-	순차회귀 (CART 등)	-	분포의 유사성	이용 목적에 따른 맞춤형 자료 생성 R패키지 개발

이 절에서는 해외사례로 미국의 SIPP Synthetic Beta(SSB), Synthetic Longitudinal Business Database(SynLBD), 독일의 IAB Establishment Panel, 영국의 SYnthetic Data Estimation for the UK LongitudinaL Studies(SYLLS)를 다루었다. 구체적으로는 각 사례에서 재현자료의 생성방법과 재현자료의 유용성 및 노출위험을 어떻게 측정, 평가하고 있는지를 살펴보았다.

재현자료 생성에는 대부분의 경우 순차회귀 다중대체(SRMI) 방식이 사용되고 있으며, 재현할 변수의 형태에 따른 회귀모형으로 선형 회귀 모형, 로지스틱 모형, 디리슈렛-다항분포 모형, CART 등이 활용되고 있다. 생성된 재현자료는 원자료에서의 변수들 간의 관계를 잘 보존하는 것이 중요하므로 원자료 특성 및 변수들의 제약조건 등을 이해하여 이를 재현자료에 반영할 필요가 있다. 미국의 SSB는 7.0 버전에서 결측치의 패턴도 함께 재현하고 있으며, 미국의 SynLB에서는 원자료의 왼쪽/오른쪽으로 절단된 변수, employment는 Firstyear 이전과 Lastyear 이후에는 값들이 존재하지 않아야 한다는 등의 변수가 가지는 제약 조건을 재현자료에도 반영하고 있다. 이러한 원자료 특성 및 변수의 제약조건 등은 재현자료의 유용성 평가에도 반영되어야 한다. 미국의 SynLB에서는 유용성 평가를 위해 당해 연도의 원자료와 재현자료의 분포 비교와 함께 종단(longitudinal)자료의 특성을 고려하여 dynamics에 대한 분포도 함께 확인하고 있다.

한편 해외사례에서는 재현자료의 노출위험을 다양한 방법으로 측정하고 있다. 미국의 SSB는 유클리디안 거리(Euclidean distance)로 거리 매칭(distance-matching)을, 미국의 SynLB는 피어슨 상관관계(Pearson correlation)를, 독일의 IAB Panel은 매칭 위험(true match risk)을 활용하고 있다. 재현자료의 노출위험은 다양한 방법으로 측정되고 있으나, 기본적으로는 재식별(re-identification) 위험과 함께 속성(attribute) 노출위험이 고려되어야 할 것이다.

제5절 한국신용정보원 개인신용정보 재현자료[10)]

신용정보원은 개인신용정보를 재현자료로 생성하여 개인신용정보 표본 DB와 교육용 DB를 제공한다. 이는 금융 빅데이터 개방시스템(CreDB)의 서비스 이용자 분석 및 교육 자료 지원 등을 목적으로 한다. 개인신용정보 표본 DB는 일반신용정보 DB에 등록된 약 4,000만 명의 약 5%에 해당하는 200만 명의 2015년 12월부터 2019년 6월까지 관찰된 카드, 대출, 연체정보 등의 개인신용정로 이루어져 있고, 이에 대한 재현자료를 구축하였다. 이 중에서 5만 명을 샘플링하여 교육용 DB로 제공한다.

---

10) 유성준, 박나리(2020), CART 기법을 이용한 개인신용정보 재현자료 생성 기법 재정리 및 인용

<표 4-5> 개인신용정보 원자료의 전처리 후 생성 변수 및 설명

	컬럼명	코드	설명
차주 정보	차주일련번호	JOIN_KEY	차주를 구별하기 위한 비식별 조치된 숫자
	차주생년	BTH_YR	10개 카테고리
	차주성별	GENDER	2개 카테고리
대출 정보	업권코드	SCTR_CD	8개 카테고리
	기관일련번호	COM_KEY	카드, 대출, 연체가 발생한 금융기관 종류
	대출상품코드 1	LN_CD_1	3개 카테고리
	대출상품코드 2	LN_CD_2	19개 카테고리
	대출시작년월	LN_FROM	1～283(월) 단, 283은 현재 대출 유지로 간주
	대출유지기간	LN_DUR	
	대출기간 내 평균대출금	LN_AMT	기간 내 평균대출금, 앞의 두 자리까지 반올림하여 범주화
연체 정보	연체유형코드	DLQ_TYPE	1개 카테고리
	연체사유코드	DLQ_CD_1	8개 카테고리
	연체등록사유코드	DLQ_CD_2	2개 카테고리
	연체시작년월	DLQ_FROM	1～283(월) 단, 283은 현재 대출 유지로 간주
	연체지속기간	DLQ_DUR	
	연체기간 내 평균연체금	DLQ_AMT	앞의 두 자리까지 반올림하여 범주화
카드 정보	카드개설사유코드	CD_OPN_CD_1	2개 카테고리
	카드유형코드	CD_OPN_CD_2	2개 카테고리
	카드개설년월	CD_OPN_FROM	1～283(월) 단, 283은 현재 대출 유지로 간주
	카드개설기간	CD_OPN_DUR	

자료: 유성준, 박나리(2020), CART 기법을 이용한 개인신용정보 재현자료 생성 기법

1. 재현자료 생성

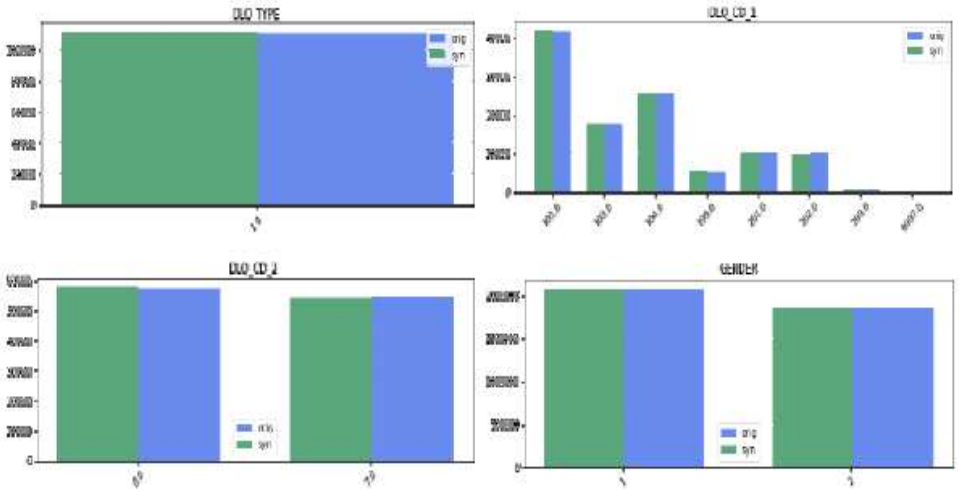
개인신용정보 재현자료는 CART 기법을 이용하여 생성한다. CART를 선택한 이유는 범주형 자료와 연속형 자료가 섞여 있는 개인신용정보의 경우에는 결합분포를 기반으로 모든 변수를 한 번에 고려하여 여러 개의 값들을 생성하는 모수적 방식을 적용하기 어렵기 때문이다. 또한 현실적으로 이러한 모수적 방식의 접근으로는 많은 변수를 대체할 수 있는 모형을 만들기도 쉽지 않다. 그리고 모집단을 잘못 가정할 경우에 정보손실이 커진다는 단점이 있어 CART와 같은 비모수적 방식이 정보손실을 최소화할 수 있을 것으로 판단하였다(유성준, 박나리, 2020).

원자료에서 먼저 차주의 카드, 대출, 연체발생 여부 및 조합의 종류에 따라 차주를 그룹화한다. 그리고 각 그룹 내 계좌 생성순서 및 기관종류에 따라 계좌를 그룹화 한 뒤, R 패키지 synthpop을 활용하여 차주 그룹 내 계좌 그룹별로 재현자료를 생성한다. 사후확률분포(PPD)에서 샘플링하는 비모수적 기법은 이전에 생성된 변수의

분포에 의존하여 다음 변수를 생성하여 변수 생성 순서에 따른 오차가 매우 크다. 따라서 범주의 수가 적은 변수를 먼저 생성한다.

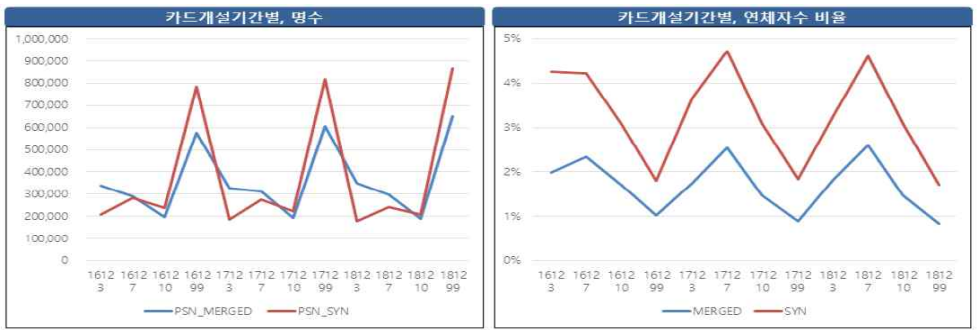
2. 재현자료 평가

재현자료의 유용성 평가에는 synthpop의 compare() 함수를 활용한다. 단변수 분포 비교 결과, 비교적 원자료와 재현자료가 유사한 분포를 보인다. 하지만 신용정보데이터의 특성상 전체 데이터의 변수별 분포보다는 특정 시점에서의 대출과 연체현황을 분석하고 각 대출과 카드개설사유별 연체자 비율을 비교하여 정보손실을 측정하는 것이 적절할 것이다. <그림 4-8>은 카드개설기간별 차주 수 및 연체자 비율을 비교한 결과이다. 연체자 비율은 다소 과대계상 되기는 하였으나 유사한 패턴을 보인다.



자료: 유성준, 박나리(2020), CART 기법을 이용한 개인신용정보 재현자료 생성 기법

<그림 4-7> 개인신용정보 재현자료: 원자료와 재현자료의 범주형 단변수 분포 비교



자료: 유성준, 박나리(2020), CART 기법을 이용한 개인신용정보 재현자료 생성 기법

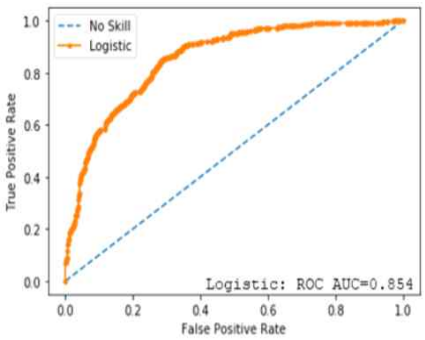
<그림 4-8> 개인신용정보 재현자료: 원자료와 재현자료의 차주 수 및 연체자 비율

재현자료의 유용성은 추가로 머신러닝을 활용하여 평가하였다. 2018년 1월부터 12월의 전체 대출 수 및 총 금액, 비은행권 대출 수 및 총 금액, 비담보 대출 수 및 총 금액, 최장기 대출기간, 최단기 대출기간 등 대출 정보를 활용하여 2019년 연체발생 여부를 예측한다. 이때 로지스틱 회귀(logistic regression), 의사결정나무(decision tree), 선형판별분석(linear discriminant analysis), 서포트벡터머신[11](support vector machine)의 4가지 머신러닝 알고리즘을 활용한다.

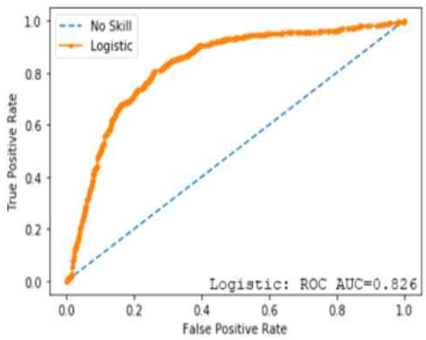
4가지 머신러닝 알고리즘을 이용한 연체발생 여부 예측 결과를 살펴보면, 원자료와 재현자료가 매우 유사한 양상으로 교육용 DB로 활용하기에 무리가 없는 것으로 판단하고 있다.

* 로지스틱 회귀(logistic regression)

[원자료]

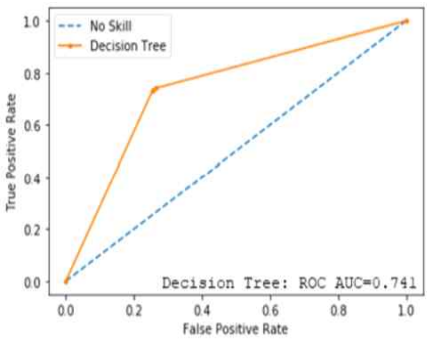


[재현자료]

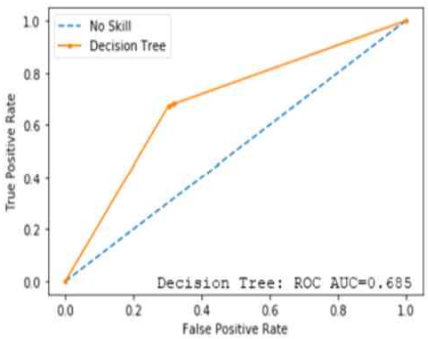


* 의사결정나무(decision tree)

[원자료]



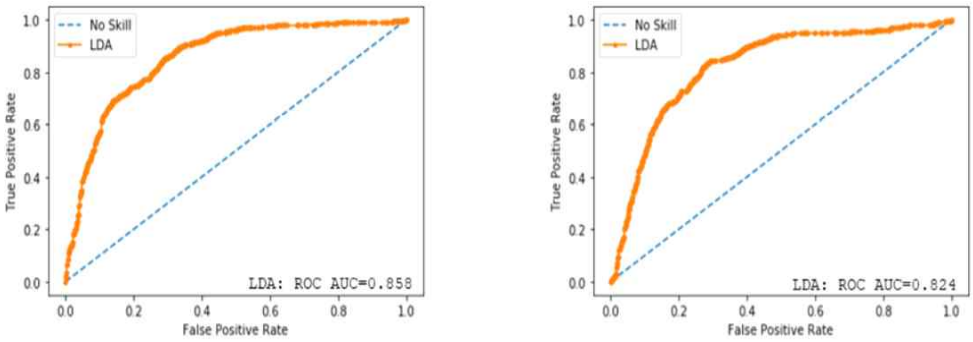
[재현자료]



11) 서포트벡터머신(support vector machine)은 기계학습 분야의 하나로, 패턴인식, 자료 분석을 위한 지도학습 모델로 주로 분류와 회귀분석을 위해 사용한다. 주어진 데이터의 집합을 바탕으로 하여 새로운 데이터가 어느 범주에 속할 것인지 판단하는 비확률적 이진 선형분류 모델을 만든다.

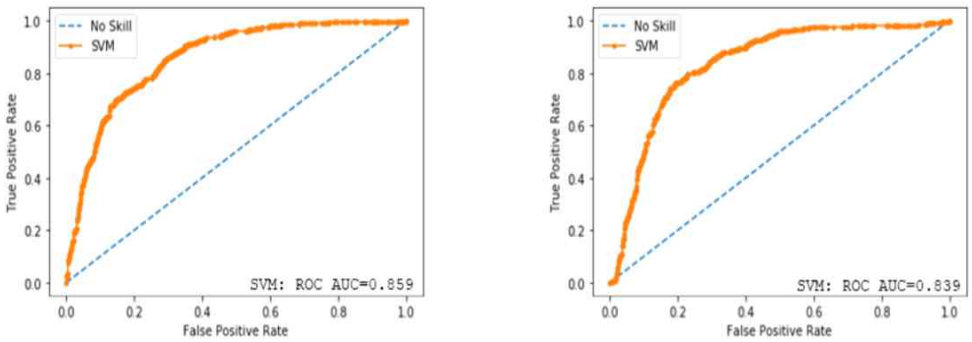
* 선형판별분석(linear discriminant analysis)

[원자료] [재현자료]



* 서포트벡터머신(support vector machine)

[원자료] [재현자료]



자료: 유성준, 박나리(2020), CART 기법을 이용한 개인신용정보 재현자료 생성 기법

<그림 4-9> 개인신용정보 재현자료: 원자료와 재현자료의 머신러닝 알고리즘별 예측결과 비교

재현자료의 노출위험을 평가하기 위해 교육용 DB 내에 원자료와 정확하게 일치하는 정보가 있는지를 살펴본다. 교육용 DB 내에 원자료와 정확하게 일치하는 데이터가 있는지를 확인하기 위해 개인의 월별 차주, 대출, 연체 및 카드개설정보를 1개 행(row) 형태로 재구성하고, 원자료와 교육용 DB 간에 정확하게 일치하는 개체의 빈도를 확인한다. 교육용 DB와 원자료가 정확하게 일치하는 비율은 원자료의 0.049%로 재식별 가능성은 낮은 것으로 평가하고 있다.

재현자료는 자료를 분석하고 활용하려는 연구자들에게 탐색과 학습의 좋은 수단이다. 해외에서는 재현자료와 관련하여 다양한 연구들이 진행되고 이에 따라 재현자료를 제공하고 있으나, 우리나라는 신용정보원의 개인신용정보 재현자료가 재현자료를 제공하는 첫 사례이다. 신용정보원의 개인신용정보 재현자료는 원자료와 재현자료에서의 개체 일치 비율을 이용해 노출위험을 측정하며, 정보손실 측면은 신용정보

데이터의 특성상 전체 데이터의 변수별 분포보다 특정 시점에서의 대출과 연체현황을 분석하고 각 대출과 카드개설사유별 연체자 비율을 비교하고 있다.

한편 한국신용정보원 외에도 한국정보화진흥원의 코리아크레딧뷰로(KCB)를 이용한 제주도 주민정보(직업, 소득, 소비 등)에 대하여 재현자료가 시범적으로 생성되어 분석 결과를 KCB 데이터 스토어에서 제공하고 있다. 올해 통계청에서도 통계데이터센터 재현 DB 구축 사업을 진행하였으며, 이는 국내에서 실제 자료를 활용한 세 번째 재현자료 구축 사업으로 파악된다.

제 5 장

통계데이터센터 DB의 재현자료 시범 생성

이 장에서는 통계청이 올해 실시한 통계데이터센터 DB의 재현자료 시범 생성 관련 용역 사업의 결과를 정리한다. 재현자료 생성 과정과 관련된 설명 및 주요 결과는 용역 사업단에서 제출한 사업 결과 보고서를 인용하였으며, 본 보고서의 연구자는 사업단에서 수행한 재현자료 생성 과정에 자문 등으로 참여하였다.

이 프로젝트의 목적은 이용자가 통계청 통계데이터센터를 직접 방문하여 실제 대용량 DB를 분석하기 이전에 사용할 수 있는 사전분석용 자료를 생산하는 것이다. 나아가 동일한 자료가 해마다 수집될 때 반복적으로 재현자료를 생성하도록 관련 모형을 시스템으로 구축하는 것이다. 센터를 방문하는 이용자에게는 시공간의 제약이 발생하지만, 원자료와 비교적 유사한 사전분석용 자료를 제공한다면 이용자의 센터 방문 시간을 감소시킬 수 있으므로 이는 이용자 편의성 향상의 의의를 가진다. 나아가 추가적인 노출위험 검토 과정을 거친 이후에, 이번 사업을 통하여 생성한 재현자료를 빅데이터 분석 교육을 위한 교육용 자료로 활용할 수도 있다.

재현자료 시범 생성은 개인정보 노출위험의 제어, 제공하는 자료의 유용성 확보 및 대용량 마이크로데이터의 분석을 원하는 이용자의 편의성 도모라는 자료제공 3대 목적 달성을 위한 새로운 시도이다. 이 장에서는 현재까지 실무적으로 활용 가능한 기술을 이용하여 통계청 실제 DB로 재현자료를 생성한 사례를 자세히 논의하고, 다음 장에서는 활발한 연구가 진행되고 있는 새로운 기술을 소개하여 향후 발전 방향을 논의할 수 있도록 한다.

제1절 통계기업등록부(SBR)의 재현자료 생성

1. 통계기업등록부(SBR) 개요

통계기업등록부(이하 SBR)는 국세청 자료, 사업자 등록부 등의 다양한 행정 자료 및 전국사업체조사, 경제총조사 등의 각종 조사 자료를 연계하여 최근에 통계청에서 구축한 경제통계 모집단이다. SBR은 수집 가능한 모든 범위의 자료를 활용하여 구축되었으며, 산업별 기업체 수, 산업분류, 종사자 및 매출액 자료 등을 제공한다. 더불

<표 5-1> 통계기업등록부(2018) – 동일대표자 기준(총 9,491,060건)

no	항목 한글명	항목 영문명	속성	자료 길이	Null 건수	유일 건수 (NA포함)
1	기준연도	BASE_YEAR	문자	4	-	1
2	사업체 고유번호	ESTM_NTE_NUM	문자	10	5,491,245	3,999,806
3	대표자 성별	RPRS_SEXDSTN	문자	1	-	3
4	개인대체식별번호	PERS_EXC_IDNTFC_NUM	문자	10	1,434,091	6,204,218
5	대표자 출생년도	RPRS_BRYY	문자	4	293,567	217
6	법정동코드	LGLD_CD	문자	10	19,133	18,781
7	행정구역코드	AD_CLS_CD	문자	7	176,002	3,521
8	시도코드	CP_CD	문자	2	9,601	18
9	시군구코드	CDW_CD	문자	5	11,673	251
10	사업장 도로명코드	BZNT_ROD_NM_CD	문자	12	28,360	153,575
11	등록일자	RGST_DATE	문자	8	284,666	19,047
12	개업일자	OPBIZ_DATE	문자	8	12,110	27,774
13	폐업일자	CLSB_DATE	문자	8	8,553,176	366
14	폐업여부	CLSB_YN	문자	1	-	2
15	자료구분	DATA_DIV_CD	문자	1	-	4
16	조직구분	ORG_DIV_CD	문자	1	-	3
17	활동구분	BR_ACT_CD	문자	1	-	3
18	매칭구분	MACH_DIV_CD	문자	1~2	-	6
19	조직형태코드	ORG_FORM_CD	문자	1	-	5
20	BR사업체 고유번호	BR_BZNT_NTE_NUM_TMPR	문자	10~11	937,874	8,553,177

no	항목 한글명	항목 영문명	속성	자료 길이	Null 건수	유일 건수 (NA포함)
21	최종 BR사업체 고유번호	FNL_BR_BZNT_NTE_NUM	문자	10	9,491,050	1
22	공공기관 유형코드	PBLINSTT_TYPE_CD	문자	1	9,437,403	9
23	비영리사업자_산업중분류	NFCR_INDST_MCLS_CD	문자	2	9,491,050	1
24	BR사업체_매출액	BR_AMTC	숫자	12	2,224,549	43,845
25	BR사업체_종사자_총계	BR_EMP_MF_T	숫자	12	-	1,649
26	BR사업체_상용근로자_합계	BR_EMP_T	숫자	12	-	1,386
27	BR사업체_상용근로자_남	BR_EMP_M	숫자	12	997,213	1,081
28	BR사업체_상용근로자_여	BR_EMP_F	숫자	12	997,097	805
29	BR사업체_임시및일용근로자_합계	BR_IMSI_T	숫자	12	-	811
30	BR사업체_임시및일용근로자_남	BR_IMSI_M	숫자	12	998,302	637
31	BR사업체_임시및일용근로자_여	BR_IMSI_F	숫자	12	998,033	464
32	BR사업체_산업분류_대	BR_SANUP	문자	1	-	22
33	BR사업체_산업분류코드_세분류	BR_SNB_CD	문자	1~5	192,972	1,480
34	BR기업체 구분	BR_ETPR_DIV_CD	문자	1	394,721	2
35	BR기업체 고유번호	BR_ETPR_NTE_NUM	문자	10~11	913,508	7,240,932
36	BR기업체_활동여부	BR_ETPR_ATV_YN	문자	1	394,721	4
37	기업규모	COM_SCL_NFCR_COR_CD	문자	2	8,782,431	6
38	중소기업 구분코드	SMLPZ_DIV_CD	문자	1~2	394,721	7
39	BR기업체_3개년 평균 매출액	BR_ETPR_3YRS_AVG_SLS	숫자	12	2,174,419	72,327
40	BR기업체_매출액	BR_ETPR_SLE_AMTC	숫자	12	3,220,425	38,091
41	BR기업체_종사자_총계	BR_ETPR_WOKE_TOTL	숫자	12	1,601,373	2,050

no	항목 한글명	항목 영문명	속성	자료 길이	Null 건수	유일 건수 (NA포함)
42	BR기업체_상용근로자_합계	BR_ETPR_CE_LARR_SUM	숫자	12	1,601,373	1,719
43	BR기업체_상용근로자_남	BR_ETPR_CE_LARR_M	숫자	12	2,250,124	1,335
44	BR기업체_상용근로자_여	BR_ETPR_CE_LARR_F	숫자	12	2,250,124	1,075
45	BR기업체_임시및일용근로자_합계	BR_ETPR_TEMR_DALY_LARR_SUM	숫자	12	1,601,373	1,045
46	BR기업체_임시및일용근로자_남	BR_ETPR_TEMR_DALY_LARR_M	숫자	12	2,250,124	797
47	BR기업체_임시및일용근로자_여	BR_ETPR_TEMR_DALY_LARR_F	숫자	12	2,250,124	664
48	BR기업체_산업분류_대	BR_ETPR_INDST_LCLS	문자	1	1,601,373	23
49	BR기업체_산업분류_대_매출액	BR_ETPR_INDST_LCLS_SLS	숫자	12	3,557,888	35,828
50	BR기업체_산업분류_중	BR_ETPR_INDST_MCL	문자	2	1,723,905	77
51	BR기업체_산업분류_중_매출액	BR_ETPR_INDST_MCL_SLS	숫자	12	3,795,097	33,898
52	BR기업체_산업분류_소	BR_ETPR_INDST_SCL	문자	3	1,739,204	231
53	BR기업체_산업분류_소_매출액	BR_ETPR_INDST_SCL_SLS	숫자	12	4,033,244	32,815
54	기업집단명	COM_GROP_NM	문자	4~12	9,450,890	35
55	기업집단_금융여부	COM_GROP_FNC_YN	문자	1	9,450,890	3

어 다른 표본 조사를 위한 모집단의 기능을 수행하고, 행정 자료를 활용한 또 다른 신규 통계 작성 등의 목적으로도 활용할 수 있다.

SBR DB에는 두 가지 버전이 존재한다. 하나는 1명의 개인이 N개 사업자로 등록될 때 1개의 기업체로 처리된 동일인(대표자) 기준을, 다른 하나는 N개의 기업체로 처리된 사업자등록번호 기준을 따른다. 각 DB는 크게 대표자, 대표자의 기업체 그리고 각 기업체를 이루는 사업체의 정보가 각 레코드/행으로 구성되어 있다. 대표자 정보로는 개인대체식별번호, 성별 및 출생년도 등이 있다. 기업체 및 사업체 정보로는 행정구역, 해당연도 내 폐업여부, 등록일자, 폐업일자, 산업분류, 조직 관련 정보, 매출액 및 종사자 수 등이 있다. 앞의 <표 5-1>에서는 재현자료 생성에서 고려한 변수를 정리하였다. 참고로 SBR이 제공하는 정보의 통계단위, 기준 및 참고자료는 '통계기업등록부 이용자 설명자료'에 자세히 제시되어 있다.

2. 재현자료 생성

SBR은 정규성이 보장되기 어려운, 치우친 분포를 가진 자료이다. 따라서 재현자료를 생성하기 위하여 모수적 모형을 적용하기는 어렵다. 비모수적 모형을 적용하기 위하여 이번 용역 사업에서는 R패키지 synthpop이 제공하는 CART 모형 등을 주로 사용하였다. R패키지는 자료 및 설정된 매개 변수를 입력받아 비모수 모형을 구축하고, 최종 노드에서 원자료를 베이지안 붓스트랩하여 재현자료를 생성한다.

CART를 적용할 때는 자료를 하위 그룹으로 나누어 각 그룹별로 재현자료를 생성한다. 이는 지역별로 분포의 특징을 유지하거나 종사자 근무활동 정보를 유의미하게 보존하기 위하여 필요하다. 또한 하위 그룹 설정은 재현자료를 효율적으로 생성하기 위해서이기도 한데, R패키지 synthpop 가이드라인[12]에서는 데이터에 12개 이상의 변수가 있거나 변수 범주의 개수가 20개 이상인 경우에는 작은 데이터프레임을 만들면서 작업할 것을 권장하고 있다.[13] 또한 R패키지가 메모리를 많이 사용하므로 계산 자원을 적절히 확보하기 위해서도 이러한 방식이 요구된다. 실제로 SBR의 경우 재현자료를 생성하는데 약 2~4일의 시간이 소요되었다.

재현자료를 생성하는 모형을 만들 때는 먼저 원자료의 변수별 특징과 구조를 상세하게 파악해야 한다. 자료의 통계적 특성에 따라 순차회귀의 순서를 정할 수 있다면 바람직하겠으나, 현실적으로는 변수들 사이에 성립하는 물리적 의미에 의해서 결정한 순서나, 각 변수가 가지는 범주의 개수가 적은 순서로 재현자료를 생성하게 된다. 범주

12) https://www.synthpop.org.uk/get-started.html#firstsynthesis

13) 가능하다면 범주의 수가 많은 변수는 범주의 종류를 통합하여 범주의 개수를 줄이는 것도 바람직하다.

의 개수를 고려하는 것은 범주의 수가 많은 변수를 재현순서의 끝으로 이동하면 계산 문제를 완화하는 데 도움이 될 수 있기 때문이다(Drechsler & Reiter, 2011). 특히 범주의 개수가 너무 많으면 분포를 보존하기 어렵고, 적당한 범주 개수를 가지는 범주형 변수의 경우에는 순서에 상관없이 유용성이 좋다는 실험 결과가 제시된 바 있다.

여러 시행착오 끝에 결정한 재현자료 생성 순서는 다음과 같다. 먼저 시도코드(CP_CD), 시군구코드(CDW_CD), 기업체-산업분류-대(BR_ETPR_INDST_LCLS)에 따라 하위 그룹을 설정하고 각 그룹별로 재현자료를 생성한다. 따라서 시도코드(CP_CD), 시군구코드(CDW_CD), 기업체-산업분류-대(BR_ETPR_INDST_LCLS)의 분포는 원자료와 동일하게 생성된다. 다음으로 지역별로 산업분류를 생성하기 위하여 각 하위 그룹별로 기업체-산업분류-중(BR_ETPR_INDST_MCL), -소(BR_ETPR_INDST_SCL), 사업체-산업분류-대(BR_SANUP), -세분류(BR_SNB_CD)를 생성한 후, 등록일자(RGST_DATE), 개업일자(OPBIZ_DATE), 폐업여부(CLSB_YN), 폐업일자(CLSB_DATE), 매출액, 종사자수를 차례로 생성한다. 이러한 순서에 따른 재현자료 생성 과정의 주의 사항 및 변수별 세부 사항을 차례로 살펴보면 다음과 같다.

먼저 재현자료 생성 모형에 대하여 논의하자. R패키지에서 제공하는 비모수적 모형에서는 CART(R패키지 rpart) 및 CTREE(R패키지 party)라는 두 가지 방식을 사용할 수 있고, 두 방법 모두 유용하다는 연구 결과가 있다(Raab et al., 2017). 첫 번째 CART의 경우에는 트리를 만들 때 가능한 모든 분할을 평가하고 최종 트리의 크기를 결정한다. 이를 위하여 복잡성 매개 변수를 설정해야 한다. SBR과 같이 크기가 큰 데이터의 경우 이 매개 변수를 매우 작은 값으로 설정해야 하는데, 그러면 모형에 범주의 개수가 많은 변수를 포함시키는 경향이 생긴다. 반면 CTREE 방식은 평가해야 하는 노드의 수를 줄이기 위하여 예비 테스트를 실시하여 중요한 변수를 먼저 선택한다. SBR 재현자료를 생성할 때 먼저 CART 방식을 사용하고 CTREE 방식을 보조적으로 사용하였으나, CART 방식을 사용하면 이상치를 확대 해석하면서 재현자료에 이상치가 과하게 생성되도록 하는 경향이 있었다. 때문에 최종적으로 CTREE 방식을 사용하고, 계산 문제 등으로 재현 자료가 생성되지 않는 경우에 CART 방식을 이용하도록 모형을 구축하였다. 그럼에도 불구하고 자료의 수가 너무 적거나 분포가 균일한 경우, 두 방식 모두에 의하여 재현자료가 생성되지 않는다. 그러한 자료는 별도로 생성한 하위 그룹에 따로 저장하고, 시도코드(CP_CD)로만 분류하여 재현자료를 생성하였다.

다음으로 R패키지에서 rules 및 rvalues를 이용하여 재현 조건을 설정하는 기능을 살펴보자. rules는 변수별 조건을 생성하고, rvalues는 변수별 조건에 따른 생성값을 고정하는 기능을 한다. rules에서 조건으로 설정된 변수는 반드시 rvalues에서 값이 설정된 변수보다 먼저 재현되어야 한다. 예를 들면 변수 중에서 폐업여부(CLSB_YN)가 NA이면 폐업일자(CLSB_DATE) 또한 NA여야 한다는 조건을 설정할 필요가 있다. 폐

업여부(CLSB_YN)가 먼저 재현되어야 폐업일자(CLSB_DATE)의 값이나 범위를 결정할 수 있기 때문이다. 다음은 이러한 과정에 대한 코드를 보여 준다.

```
# RULES AND VALUES
rules <- list(CLSB_DATE=="is.na(CLSB_YN)")
rvalues <- list(CLSB_DATE=NA)
```

SBR의 경우에는 위의 예를 제외하면 일반적으로 적용할 수 있는 다른 조건을 찾기 어려웠다. 이는 자료에서 오타나 오기, 누락 등이 관찰되어 특정 규칙을 적용하기에 부적절하기 때문이었다. 따라서 SBR 재현자료 시범 생성에서는 폐업여부에 따른 폐업일자 생성 외에 다른 조건은 적용하지 않았다.

한편 매개변수 m은 재현할 데이터 세트의 개수이며, 이 프로젝트에서는 1세트만 생성한다. 또 다른 매개변수 minbucket은 재현방식별 모든 최종 노드에서 관찰되는 최소 자료의 개수를 나타내며, 3으로 설정하였다. 만약 minbucket을 크게 설정하면 최종 노드에서 선택할 수 있는 자료의 종류가 다양해지므로 재현자료가 원자료와 달라질 가능성이 높아 노출위험이 낮아진다. 반대로 자료의 유용성은 떨어지게 된다. 마지막 매개변수 seed는 의사 난수 생성기(pseudo-random number generator)를 고정하는 데에 사용되며, 동일한 자료가 다시 생성되도록 하드는 데 사용된다.

또 다른 주의사항으로 R패키지 synthpop을 사용하여 재현자료를 생성할 때는 변수의 유형(class)을 미리 정해주어야 한다는 점이 있다. syn() 함수가 실행되면 변수 범주 수에 따라서 변수 유형이 factor, binary, constant 등으로 지정되므로 주의하지 않으면 에러가 발생할 수 있다. 이번에 실시한 재현자료 시범생성 사업에서는 연속형은 numeric으로, 범주형은 character로 설정하고 작업하였으며, 시군구 코드는 60개 이상의 범주가 존재하므로 범주형 변수임에도 불구하고 연속형으로 설정하였다.

마지막으로 참고할 사항은 R패키지가 제공하는 numtocat 기능이다. 이 기능은 범주의 수가 많은 연속형 변수를 요인(factor)으로 변환해 주어, 붓스트랩으로 재현자료를 생성할 수 있도록 한다. 이번 프로젝트에서는 연령, 출생년도, 매출액, 종사자 수와 같은 정수형 변수에 이 기능을 활용하였다. 다만 각 요인에 해당하는 범위를 사용자가 결정할 수는 없기 때문에 샘플링 과정에서 원자료와 다른 결과가 관찰되기도 하였다. 이 기능을 사용하는 부분의 코드는 다음과 같다.

```
# DATA TYPE
cats <- c("RPRS_SEXDSTN","CP_CD","CDW_CD","CLSB_YN","BR_SANUP","BR_ETPR_INDST_LCLS")
conts <- c(colnames(sbr)[!colnames(sbr) %in% cats])
for (col in colnames(sbr)) {
  if (col %in% cats) {
    sbr[[col]] <- as.character(sbr[[col]])
  } else {
    sbr[[col]] <- as.numeric(sbr[[col]])
  }
}

# numtocat VARIABLES
ls.numtocat <- c("RPRS_BRYY","BR_ETPR_SLE_AMTC","BR_ETPR_WOKE_TOTL")
```

이번에는 변수별 재현자료 생성 세부 사항을 살펴보자. 하위그룹 분류에 사용된 시도코드(CP_CD), 시군구코드(CDW_CD), 기업체-산업분류-대(BR_ETPR_INDST_LCLS) 변수들의 경우 각 하위 그룹에서 1가지 값만을 가진다. 따라서 상수이므로 재현방식(method) 설정에서 재현순서(visit.sequnce)의 처음 3개의 변수는 아무 것도 하지 않도록(“”) 지정한다. 4번째 변수인 기업체-산업분류-중(BR_ETPR_INDST_MCL)이 첫 번째로 생성되는 변수가 되며, 예측변수가 없으므로 복원추출 방식을 사용하여 재현자료를 생성한다. 마지막 두 변수인 매출액 및 종사자 수를 생성할 때는 NULL값이 적은 사업체의 매출액 및 종사자 수를 기업체의 매출액 및 종사자 수보다 먼저 생성하고, 매출액 및 종사자 수의 비율을 최대한 유지하도록 설정하였다. 이상의 내용을 반영하여 재현자료를 생성하는 부분의 코드는 다음과 같다.

```
# METHODS
method.cart<-rep(c("","cart"), times=c(3,ncol(temp)-3))
method.ctree<-rep(c("","ctree"), times=c(3,ncol(temp)-3))

# SYNTHESIS CART & CTREE
tryCatch (
  {
    temp.syn_ <- syn(temp, m=1, seed=1234, method=method.ctree, ctree.minbucket=3,
                     visit.sequence=seqs, numtocat=ls.numtocat, rules=rules,
rvalues=rvalues)
    fname <- paste('./output/sbr/syn/sbr_',k,"_",l,"_",i,".csv",sep="")
    write.csv(temp.syn$syn,fname,row.names=F)
  },
  error = function(e) {
    tryCatch (
      {
        message("CTREE FAILED. Trying CART....")
        temp.syn_ <- syn(temp, m=1, seed=1234, method=method.cart, cart.minbucket=3,
                         visit.sequence=seqs, numtocat=ls.numtocat, rules=rules,
rvalues=rvalues)
        fname <- paste('./output/sbr/syn/sbr_',k,"_",l,"_",i,".csv",sep="")
        write.csv(temp.syn$syn,fname,row.names=F)
        message("CART SUCCEEDED.\n")
      },
      error = function(e) {
        message("CART ALSO FAILED.\n")
        fname <- paste('./output/sbr/resid/sbr_',k,"_",l,"_",i,".csv",sep="")
        write.csv(temp,fname,row.names=F)
      },
      finally = {
        message("")
      }
    )
  },
  finally = {
    message("")
  }
)
```

재현자료를 생성하면서 발견한 SBR 자료의 특징은 정리하면 다음과 같으므로 향후 제공받은 재현자료를 분석할 때 주의할 필요가 있다. 먼저 SBR은 조사자료가 아니며 복수의 원천 자료를 통계적 기법으로 연계한 자료이므로 실제 기업체 또는 사업체 자료와 차이가 있을 수 있다. 또한 수집 및 관리 단계에서 작업자가 수기로 자료를 관리하면서 발생했을지도 모르는 오기/오타, 변수 형식 변경 등 각종 에러가 관찰된다. 다음으로 SBR 설명자료에 나오는 기업체-사업체 정의 및 기업체-사업체 간

연계는 조사 단계에서 사용한 특정 기준이 제공되지 않았으므로 직접적으로 연계시키기 어렵다. 세 번째로 산업분류, 자료 출처 등의 분류 정보는 원천 자료의 조사 단계에서 임의로 지정된 것이므로 실제 사업체 성격과 다를 수 있다. 네 번째로 SBR 자료에서 NULL 값은 존재하지 않거나, 누락되었거나, 0이거나 혹은 수집 단계에서 발생한 오류이다. 이로 인하여 원자료 혹은 재현자료를 분석한 결과가 실제 상황을 반영하지 않을 수 있다. 마지막으로 재현자료를 이용해서는 특정 기업체나 사업체, 또는 종사자에 대하여 연도 간 연계하는 것은 불가능하다. 특정 기업체나 사업체에 대한 시계열 분석은 재현자료를 이용하여 수행할 수 없음을 주의하기 바란다.

SBR의 총 55개 변수들 가운데 재현자료 생성에 관련된 사항을 변수별로 간단히 정리하면 다음과 같다. 참고로 <표 5-2>에서는 총 55개 변수들 중에서 전처리 과정을 거쳐 재현자료 생성에 사용한 변수명들을 한눈에 볼 수 있도록 정리하였다. 최종 산출물에 포함하도록 결정한 변수는 진하게 표시하였다.

기준연도(BASE_YEAR)
생성하지 않으며, 다른 변수의 재현자료를 생성한 후 원자료 기준연도를 일괄적으로 부여한다.

사업체고유번호(ESTM_NTE_NUM)
전처리 작업 후 변수별 오름차순으로 정렬한 자료의 사업체고유번호를 따로 저장한 뒤, 재현자료를 동일한 변수 순서로 오름차순 정리한 후 열을 기준으로 결합(cbind())하여 부여한다.

대표자성별(RPRS_SEXDSTN)
범주형 변수로 재현자료를 생성한다.

대표자출생년도(RPRS_BRYY)
연속형 변수로 numtocat 기능을 사용하여 재현자료를 생성한다.

시도코드(CP_CD)
범주형 변수로 재현자료를 생성하지 않으며, 하위 그룹 분류에 사용된다. 재현자료는 원자료와 동일한 수의 시도코드(CP_CD)를 가진다.

시군구코드(CDW_CD)
범주형 변수로 재현자료를 생성하지 않으며, 하위 그룹 분류에 사용된다. 재현자료는 원자료와 동일한 시군구코드(CDW_CD) 수와 시도코드(CP_CD)에 대한 조건부 분포를 가진다.

등록일자(RGST_DATE)
YYYYMMDD 형태가 아닌 데이터는 NA로 변환하고, 입력된 자료의 최소 RGST_DATE로부터 일(day) 수를 산출하여 RGST_DAYS라는 연속형 변수를 새로 생성하고, numtocat 기능을 사용하여 재현자료를 생성한다.

개업일자(OPBIZ_DATE)
YYYYMMDD 형태가 아닌 데이터는 NA로 변환하고, RGST_DATE으로부터 일(day) 수를 산출하여 OPBIZ_DAYS라는 연속형 변수를 새로 생성하고, numtocat 기능을 사용하여 재현자료를 생성한다. 등록일자와 개업일자는 선후가 정해지지 않았으며, 몇 년씩 차이가 나기도 한다.

폐업일자(CLSB_DATE)
YYYYMMDD 형태가 아닌 데이터는 NA로 변환하고, RGST_DATE으로부터 일(day) 수를 산출하여 CLSB_DAYS라는 연속형 변수를 새로 생성하고, numtocat 기능을 사용하여 재현자료를 생성한다. 폐업한 사업체의 경우는 모두 입력된 자료의 기준연도 내에서 발생하지만, 해당 변수는 지역별, 산업별, 기간별 특징이 반영되는 계속기업(going concern) 개념으로 간주하여 등록일자(RGST_DATE)를 기준으로 하는 기간으로 설정하였다.

폐업여부(CLSB_YN)
범주형 변수로 재현자료를 생성한다.

사업체-매출액(BR_AMTC)
기업체와 사업체 간 1:N 관계를 가지고 있고, 사업체 매출액은 기업체와 동일하거나 부분이 되므로 기업체-매출액(BR_ETPR_SLE_AMTC)으로 나누어 비율을 산정한 뒤, 연속형 변수로 설정하고 재현자료를 생성한다.

사업체-종사자-총계(BR_EMP_MF_T)
기업체의 종사자 수와 같거나 적은 종사자 수를 생성하도록 기업체-종사자-총계(BR_ETPR_WOKE_TOTL)로 나누어 비율로 산정한 뒤, 연속형 변수로 설정하고 재현자료를 생성한다.

사업체-상용근로자-총계(BR_EMP_T)
사업체 종사자 수보다 많은 수를 생성할 수 없도록 사업체-종사자-총계(BR_EMP_MF_T)로 나누어 비율로 산정한 뒤, 연속형 변수로 설정하고 재현자료를 생성한다.

사업체-상용근로자-남(BR_EMP_M)
전체 사업체 상용근로자 수보다 많은 수의 재현자료를 생성하지 않도록 사업체-상용근로자-총계(BR_EMP_T)로 나누어 비율로 산정한 뒤, 연속형 변수로 설정하고 재현자료를 생성한다.

사업체-임시및일용근로자-남(BR_IMSI_M)
사업체 종사자 총계에서 상용 근로자 수를 뺀 나머지가 임시 및 일용 근로자로 산출된다. 사업체 임시 및 일용 근로자 총계보다 많은 수의 자료가 생성되지 않도록 사업체-임시 및 일용 근로자-합계(BR_IMSI_T)로 나누어 비율로 산정한 뒤, 연속형 변수로 설정하고 재현자료를 생성한다.

사업체-산업분류-대(BR_SANUP)
범주형 변수로 재현자료를 생성한다.

사업체산업분류-세분류(BR_SNB_CD)
범주형 변수이나 synthpop이 허용하는 factor level 수 60개보다 많으므로 연속형 변수로 설정하고 재현자료를 생성한다. 2018년 자료는 9차, 10차 산업분류가 섞여 있으므로 자료의 길이가 2~5자리로 일정하지 않고, 메타데이터와 다르다.

기업체-매출액(BR_ETPR_SLE_AMTC)
연속형 변수로 설정하고 numtocat 기능을 사용하여 재현자료를 생성한다.

기업체-종사자-총계(BR_ETPR_WOKE_TOTL)
연속형 변수로 설정하고 numtocat 기능을 사용하여 재현자료를 생성한다.

기업체-상용근로자-합계(BR_ETPR_CE_LARR_SUM)
기업체 종사자 총계보다 많은 자료가 생성되지 않도록 기업체-종사자-총계(BR_ETPR_WOKE_TOTL)로 나누어 비율로 산정한 뒤, 연속형 변수로 설정하고 재현자료를 생성한다.

기업체-상용근로자-남(BR_ETPR_CE_LARR_M)
기업체 상용근로자 합계보다 많은 수의 재현자료가 생성되지 않도록 기업체-상용근로자-합계(BR_ETPR_CE_LARR_SUM)로 나누어 비율로 산정한 뒤, 연속형 변수로 설정하고 재현자료를 생성한다.

기업체-임시및일용근로자-남(BR_ETPR_TEMR_DALY_LARR_M)
기업체 종사자 총계에서 기업체 상용 근로자 합계를 뺀 수가 기업체 임시 및 일용 근로자 총계가 된다. 기업체 임시 및 일용 근로자 총계보다 많은 수의 자료가 생성되지 않도록 기업체 임시 및 일용 근로자-합계(BR_ETPR_TEMP_DALY_LARR_SUM)로 나누어 비율로 산정한 뒤, 연속형 변수로 설정하고 재현자료를 생성한다.

기업체-산업분류-대(BR_ETPR_INDST_LCLS)
범주형 변수로 재현자료를 제공한다.

기업체-산업분류-중(BR_ETPR_INDST_MCL)
범주형 변수로 재현자료를 생성한다.

기업체-산업분류-소(BR_ETPR_INDST_SCL)
범주형 변수로 재현자료를 생성한다.

<표 5-2> 통계기업등록부 변수별 전처리 및 재현방법

no	항목 한글명	항목 영문명	속성	재현	재현여부 및 방법
1	**기준연도**	**BASE_YEAR**		**X**	**재현자료 생성 후 기준연도 일괄 부여**
2	**사업체 고유번호**	**ESTM_NTE_NUM**		**X**	**재현자료 생성 후 랜덤한 사업체고유번호 부여**
3	**대표자 성별**	**RPRS_SEXDSTN**	**범주형**	**O**	**재현**
4	**개인대체식별번호**	**PERS_EXC_IDNTFC_NUM**		**X**	**재현자료 생성 후 랜덤한 개인대체식별번호 부여**
5	**대표자 출생년도**	**RPRS_BRYY**	**연속형**	**O**	**재현**
6	법정동코드	LGLD_CD		X	재현하지 않음
7	행정구역코드	AD_CLS_CD		X	재현하지 않음
8	**시도코드**	**CP_CD**	**범주형**	**O**	**재현**
9	**시군구코드**	**CDW_CD**	**범주형**	**O**	**재현**
10	사업장 도로명코드	BZNT_ROD_NM_CD		X	재현하지 않음
11	**등록일자**	**RGST_DATE**	**연속형**	**O**	**자료의 최소 등록일자로부터 일(day) 수를 산출하여 재현**
12	**개업일자**	**OPBIZ_DATE**	**연속형**	**O**	**등록일자로부터 일(day) 수를 산출하여 재현**
13	**폐업일자**	**CLSB_DATE**	**연속형**	**O**	**등록일자로부토 일(day) 수를 산출하여 재현**
14	**폐업여부**	**CLSB_YN**	**범주형**	**O**	**재현**
15	자료구분	DATA_DIV_CD		X	수집 자료의 출처로, 재현하지 않음
16	조직구분	ORG_DIV_CD		X	재현하지 않음
17	**활동구분**	**BR_ACT_CD**		**X**	**재현자료 생성 후 사업체-매출액 또는 종사자 수 중 하나만이라도 0보다 클 경우 1(활동), 2(비활동)으로 부여하고, 폐업한 경우 3(폐업) 부여**
18	매칭구분	MACH_DIV_CD		X	수집 자료 간 검증코드로, 재현하지 않음
19	조직형태코드	ORG_FORM_CD		X	임의 부여된 코드로, 재현하지 않음
20	**BR사업체 고유번호**	**BR_BZNT_NTE_NUM_TMPR**		**X**	**재현자료 생성 후 랜덤한 사업체고유번호 부여**
21	최종 BR사업체 고유번호	FNL_BR_BZNT_NTE_NUM		X	재현하지 않음 (모두 null)
22	공공기관 유형코드	PBLINSTT_TYPE_CD		X	재현하지 않음

no	항목 한글명	항목 영문명	속성	재현	재현여부 및 방법
23	비영리사업자_산업중분류	NFCR_INDST_MCLS_CD		X	재현하지 않음
24	**BR사업체_매출액**	**BR_AMTC**	**연속형**	**O**	**기업체_매출액으로 나누어 비율로 재현하고, 재현자료 생성 후 재산출**
25	**BR사업체_종사자_총계**	**BR_EMP_MF_T**	**연속형**	**O**	**기업체_종사자_총계로 나누어 비율로 재현하고, 재현자료 생성 후 재산출**
26	**BR사업체_상용근로자_합계**	**BR_EMP_T**	**연속형**	**O**	**사업체_종사자_총계로 나누어 비율로 재현하고, 재현자료 생성 후 재산출**
27	**BR사업체_상용근로자_남**	**BR_EMP_M**	**연속형**	**O**	**사업체_상용근로자_합계로 나누어 비율로 재현하고, 재현자료 생성 후 재산출**
28	**BR사업체_상용근로자_여**	**BR_EMP_F**		**X**	**재현자료 생성 후 산출 (BR_EMP_T - BR_EMP_M)**
29	**BR사업체_임시및일용근로자_합계**	**BR_IMSI_T**		**X**	**재현자료 생성 후 산출 (BR_EMP_MF_T - BR_EMP_T)**
30	**BR사업체_임시및일용근로자_남**	**BR_IMSI_M**	**연속형**	**O**	**사업체_임시및일용근로자_합계로 나누어 비율로 재현하고, 재현자료 생성 후 재산출**
31	**BR사업체_임시및일용근로자_여**	**BR_IMSI_F**		**X**	**재현자료 생성 후 산출 (BR_IMSI_T - BR_IMSI_M)**
32	**BR사업체_산업분류_대**	**BR_SANUP**	**범주형**	**O**	**재현**
33	**BR사업체_산업분류코드_세분류**	**BR_SNB_CD**	**범주형**	**O**	**재현**
34	**BR기업체 구분**	**BR_ETPR_DIV_CD**		**X**	**값이 1인 자료만 필터링하여 재현하고 1을 일괄 부여**
35	**BR기업체 고유번호**	**BR_ETPR_NTE_NUM**		**X**	**재현자료 생성 후 랜덤한 기업체고유번호 부여**
36	**BR기업체_활동여부**	**BR_ETPR_ATV_YN**		**X**	**재현자료 생성 후 기업체-매출액과 또는 종사자수 중 하나만이라도 0보다 클 경우 1(활동), 2(비활동)으로 부여하고, 폐업한 경우 3(폐업) 부여**
37	기업규모	COM_SCL_NFCR_COR_CD		X	재현하지 않음
38	중소기업 구분코드	SMLPZ_DIV_CD		X	재현하지 않음

no	항목 한글명	항목 영문명	속성	재현	재현여부 및 방법
39	BR기업체_3개년 평균 매출액	BR_ETPR_3YRS_AVG_SLS		X	재현하지 않음
40	**BR기업체_매출액**	**BR_ETPR_SLE_AMTC**	**연속형**	**O**	**재현**
41	**BR기업체_종사자_총계**	**BR_ETPR_WOKE_TOTL**	**연속형**	**O**	**재현**
42	**BR기업체_상용근로자_합계**	**BR_ETPR_CE_LARR_SUM**	**연속형**	**O**	**기업체_종사자_총계로 나누어 비율로 재현하고, 재현자료 생성 후 재산출**
43	**BR기업체_상용근로자_남**	**BR_ETPR_CE_LARR_M**	**연속형**	**O**	**기업체_상용근로자_합계로 나누어 비율로 재현하고, 재현자료 생성 후 재산출**
44	**BR기업체_상용근로자_여**	**BR_ETPR_CE_LARR_F**		**X**	**재현자료 생성 후 산출 (BR_ETPR_CE_LARR_SUM - _LARR_M)**
45	**BR기업체_임시및일용근로자_합계**	**BR_ETPR_TEMR_DALY_LARR_SUM**		**X**	**재현자료 생성 후 산출 (BR_ETPR_WOKE_TOTL - _LARR_SUM)**
46	**BR기업체_임시및일용근로자_남**	**BR_ETPR_TEMR_DALY_LARR_M**	**연속형**	**O**	**기업체_임시및일용근로자_합계로 나누어 비율로 재현하고, 재현자료 생성 후 재산출**
47	**BR기업체_임시및일용근로자_여**	**BR_ETPR_TEMR_DALY_LARR_F**		**X**	**재현자료 생성 후 산출 (BR_ETPR_TEMR_DALY_LARR_SUM - _LARR_M)**
48	**BR기업체_산업분류_대**	**BR_ETPR_INDST_LCLS**	**범주형**	**O**	**재현**
49	BR기업체_산업분류_대_매출액	BR_ETPR_INDST_LCLS_SLS		X	재현하지 않음
50	**BR기업체_산업분류_중**	**BR_ETPR_INDST_MCL**	**범주형**	**O**	**재현**
51	BR기업체_산업분류_중_매출액	BR_ETPR_INDST_MCL_SLS		X	재현하지 않음
52	**BR기업체_산업분류_소**	**BR_ETPR_INDST_SCL**	**범주형**	**O**	**재현**
53	BR기업체_산업분류_소_매출액	BR_ETPR_INDST_SCL_SLS		X	재현하지 않음
54	기업집단명	COM_GROP_NM		X	재현하지 않음
55	기업집단_금융여부	COM_GROP_FNC_YN		X	재현하지 않음

3. 재현자료 평가

재현자료의 노출위험과 정보손실을 측정하여 재현자료를 평가한다. 먼저 노출위험은 SBR DB에서 사업체의 행정구역과 산업분류에 대하여 원자료에서 유일한 레코드(행)가 재현자료에서 다시 유일하게 생성된 비율을 산출하여 측정하였다. SBR 재현자료에서는 이러한 비율이 미미하였다. 재현자료로 생성하는 변수의 수가 많아서 유일한 개체가 동일하게 생성되는 확률이 낮음을 고려하면 예상되는 결과이다. 이러한 자료는 제거하고 제공할 수도 있으며, 이러한 자료만 따로 모아서 유일한 개체가 아닐 때까지 반복하여 새로 생성해서 대체할 수도 있다. R패키지가 제공하는 기본 데이터 세트인 SD2011을 이용하여 동일하게 생성된 유일한 자료를 제거하는 예는 다음과 같다.

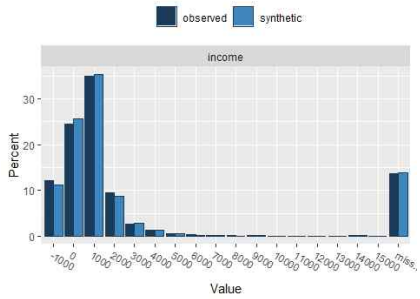
```
# 데이터 선택 및 재현
> ods <- SD2011[1:500,c("sex","age","edu","smoke")]
> s1 <- syn(ods)
# 동일하게 재현된 자료 제거(s2$syn 사용)
> s2 <- sdc(s1, ods, rm.replicated.uniques=TRUE)
no. of replicated uniques: 61

# 동일하게 재현된 자료 확인(TRUE)
> s3 <- replicated.uniques(s1, ods)
# 동일하게 재현된 자료 index 추출
> rem <- which(s3$replications)
# 동일하게 재현된 자료만 따로 추출
> dup <- s1$syn[rem,]
```

한편 재현자료의 정보손실을 측정하는 방식은 매우 다양하다. 이번 프로젝트에서는 단변량 비교, 이변량 비교, 신뢰구간 비교 등을 수행하였다. 이를 차례로 살펴보면 다음과 같다.

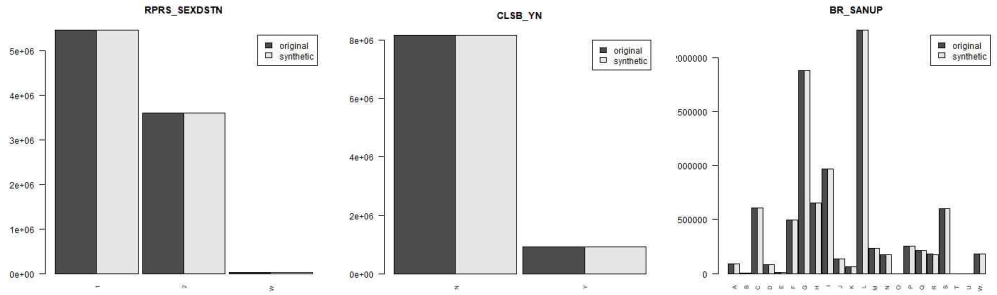
가. 단변량 비교(Comparison of univariate distributions)

R패키지 synthpop의 compare() 함수는 syn() 함수에 의하여 생성되는 synds 클래스 객체를 사용하여 원자료와 재현자료의 단변량 변수의 분포를 시각적으로 비교하는 기능을 제공하고 있다. 다만 SBR과 같은 대용량 자료의 경우 하위 그룹으로 나누어 재현자료를 생성하는데 메모리 사용이 많으므로 각 하위 그룹마다 생성되는 synds 객체를 모두 저장할 수 없다. 때문에 compare() 함수와 동일한 결과를 보여주는 함수를 따로 구현하여 사용하였다. 용량의 문제가 없다면 compare() 함수를 사용하여 동일한 작업을 수행할 수 있다. R패키지가 제공하는 기본 데이터 세트인 SD2011을 이용하여 compare() 기능을 사용하는 예는 다음과 같다.

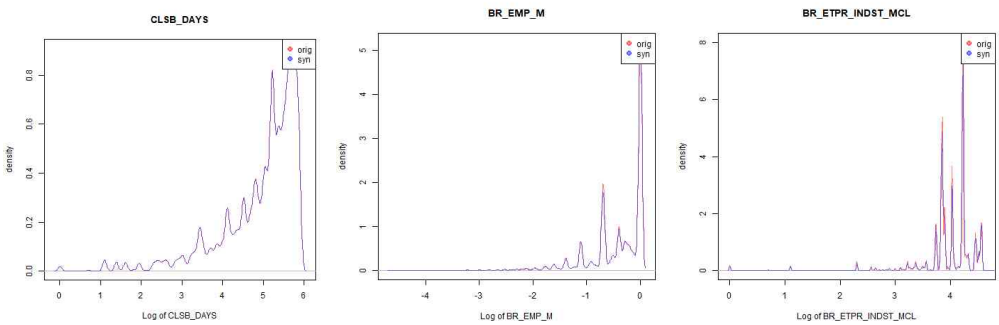
```
# 변수 선택 및 재현자료 생성
R> ods <- SD2011[,c("sex","age","edu",
                    "smoke","income")]
R> s1 <- syn(ods)

# compare()
R> compare(s1, ods, vars = "income")
```



한편 SBR 일부 범주형 변수에 대하여 도수분포표를 비교한 결과는 다음과 같다.



SBR 일부 연속형 변수에 대하여 분포를 비교한 결과는 다음과 같다. 원자료에서 각 변수의 분포가 비대칭이 심하면서 한쪽 꼬리가 길므로 로그 변환하였다.

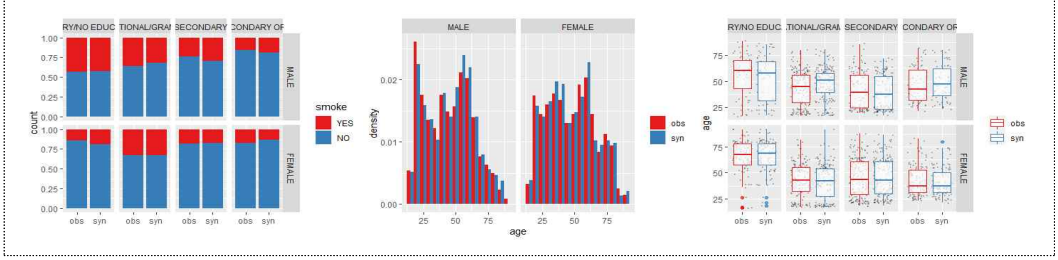


나. 이변량 비교(Comparison of bivariate distributions)

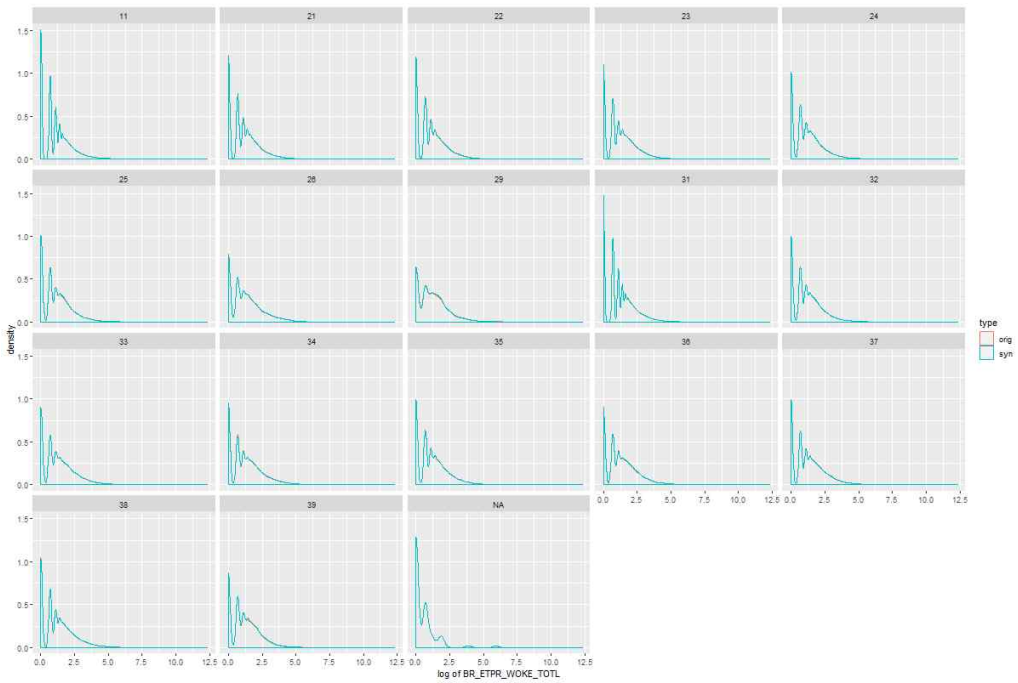
R패키지 synthpop의 multi.compare() 함수는 원자료와 재현자료에 대하여 주어진 변수의 범주별 분포를 시각적으로 비교하는 기능을 제공한다. 분포를 비교하고자 하는 변수를 var에, 그룹별로 확인하고자 하는 변수를 by에 지정하면 된다. by에 지정하는

변수는 범주가 적은 변수를 선택하는 것이 바람직하다. R패키지가 제공하는 기본 데이터 세트인 SD2011을 이용하여 multi.compare() 기능을 사용하는 예는 다음과 같다.

```r
# 변수 선택 및 데이터 재현
vars <- c("sex", "age", "edu", "smoke")
ods <- na.omit(SD2011[1:1000, vars])
s1 <- syn(ods)

# 범주형 변수
multi.compare(s1, ods, var = "smoke", by = c("sex","edu"))
multi.compare(s1, ods, var = "age", by = c("sex"), y.hist = "density", binwidth = 5)
multi.compare(s1, ods, var = "age", by = c("sex", "edu"), cont.type = "boxplot")
```



한편 SBR 일부 범주형-연속형 변수에 대한 비교는 다음과 같다. 대부분의 이변량 변수 비교에 대하여 SBR 재현자료는 적은 정보손실을 가지는 것으로 나타난다.



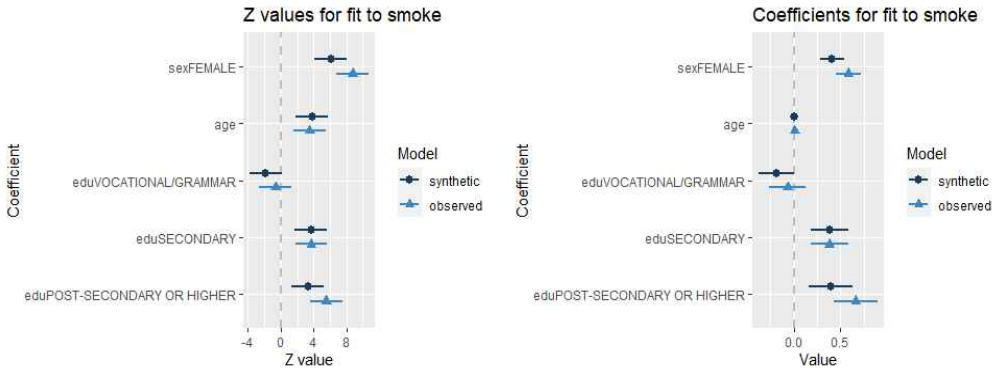
<그림 5-1> CP_CD별 BR_ETPR_WOKE_TOTL(BR기업체-종사자-총계) 분포

다. 95% 신뢰구간 비교

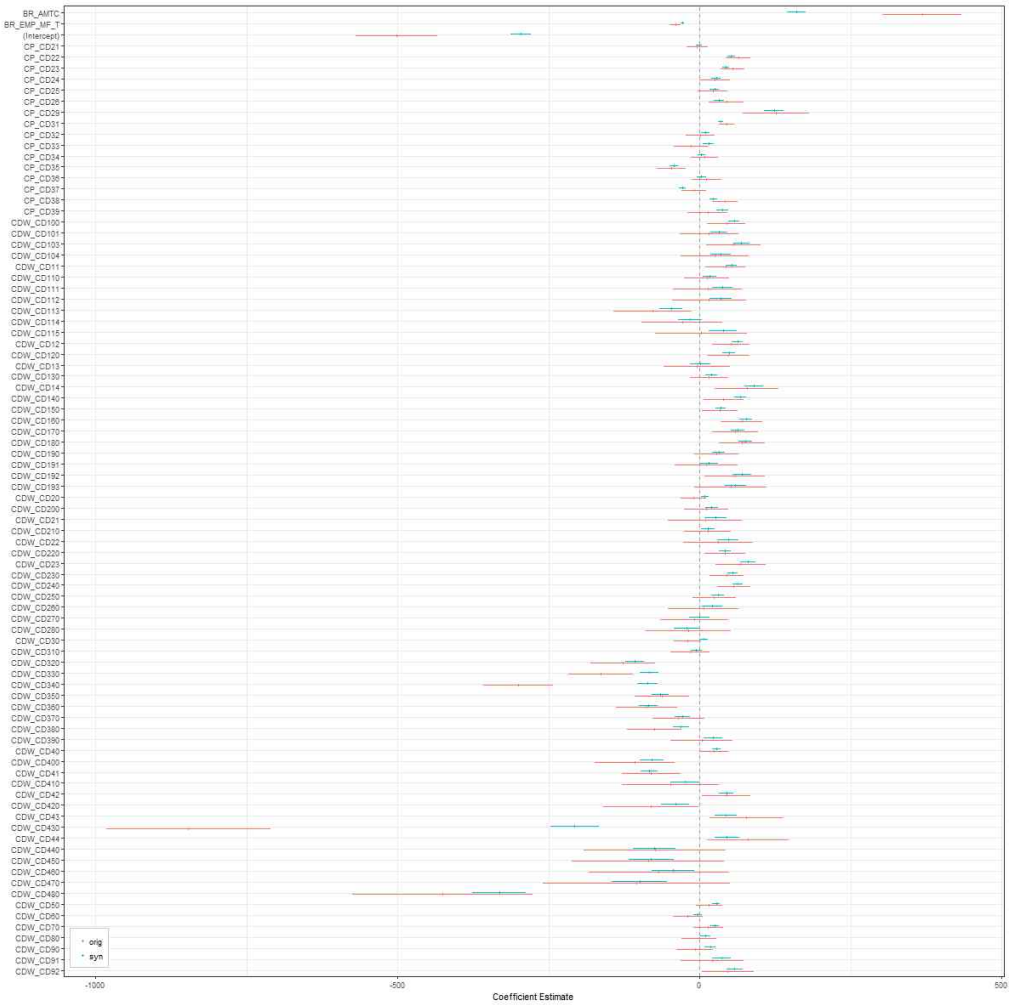
Snoke et al.(2018)은 재현자료의 정보손실을 측정하는 방법으로 원자료와 재현자료 각각에서 구한 일반 선형 회귀 모형의 계수를 비교하였다. 계수 추정값에 대한 신뢰구간의 중복(Interval Overlap, IO)을 계산하고 그 평균을 정보손실 측도로 사용한 것이다. 이 측도는 신뢰구간 중복이 없으면 음의 값을 가지며, 중복이 적게 발생할수록 값이 작아진다. 한편 R패키지 synthpop의 compare() 함수는 유사한 정보손실 측정 결과를 표와 그림으로 제공한다. 먼저 재현자료를 이용한 선형 회귀 모형 적용 결과를 glm.synds() 함수를 통해 얻고, 이를 원자료와 함께 compare() 함수의 인수로 사용하면 된다. 기본 데이터 세트인 SD2011을 이용하여 이러한 기능을 사용하는 예는 다음과 같다.

```
# 재현자료 생성 및 회귀모형 설정
> ods <- SD2011[,c("sex","age","edu","smoke")]
> s1 <- syn(ods)
> f1 <- glm.synds(smoke ~ sex + age + edu, data=s1, family="binomial")

# summary
> f1

Call:
glm.synds(formula = smoke ~ sex + age + edu, family = "binomial", data = s1)

Combined coefficient estimates:
               (Intercept)                 sexFEMALE
               0.152265849               0.388853557
                       age      eduVOCATIONAL/GRAMMAR
               0.007046393               0.148170344
             eduSECONDARY eduPOST-SECONDARY OR HIGHER
               0.647429637               0.827761430

# compare(): z-value
> compare(f1, ods)

Call used to fit models to the data:
glm.synds(formula = smoke ~ sex + age + edu, family = "binomial", data = s1)

Differences between results based on synthetic and observed data:
                           Std. coef diff p value CI overlap
(Intercept)                     1.9413183   0.052  0.5047566
sexFEMALE                      -0.8111420   0.417  0.7930722
age                            -1.4775469   0.140  0.6230678
eduVOCATIONAL/GRAMMAR          -0.2028081   0.839  0.9482623
eduSECONDARY                   -0.9865696   0.324  0.7483195
eduPOST-SECONDARY OR HIGHER    -0.5983188   0.550  0.8473648

Measures for one synthesis and 6 coefficients
Mean confidence interval overlap:  0.7441405
Mean absolute std. coef diff:  1.002951
Lack-of-fit: 9.240844; p-value 0.16 for test that synthesis model is compatible
with a chi-squared test with 6 degrees of freedom
```

```
# compare(): regression coefficients
> compare(f1, ods, print.coef = TRUE, plot = "coef")

Call used to fit models to the data:
glm.synds(formula = smoke ~ sex + age + edu, family = "binomial", data = s1)

Estimates for the observed data set:
                                Beta    se(Beta)         Z
(Intercept)                0.242725695 0.144911210  1.6749960
sexFEMALE                  0.584698020 0.066878325  8.7427133
age                        0.006803412 0.001946671  3.4948951
eduVOCATIONAL/GRAMMAR      -0.065900591 0.098898951 -0.6663427
eduSECONDARY               0.381817511 0.102912518  3.7101173
eduPOST-SECONDARY OR HIGHER  0.662876527 0.119733693  5.5362572

Combined estimates for the synthetised data set(s):
                           xpct(Beta) xpct(se.Beta)   xpct(z)
(Intercept)                0.334123190   0.144911210  2.305710
sexFEMALE                  0.408806958   0.066878325  6.112697
age                        0.007354362   0.001946671  3.777917
eduVOCATIONAL/GRAMMAR      -0.186014756   0.098898951 -1.880857
eduSECONDARY               0.378897022   0.102912518  3.681739
eduPOST-SECONDARY OR HIGHER  0.391433086   0.119733693  3.269197

Differences between results based on synthetic and observed data:
                         Std. coef diff p value CI overlap
(Intercept)                  0.63071377   0.528  0.8391007
sexFEMALE                   -2.63001595   0.009  0.3290652
age                          0.28302165   0.777  0.9277993
eduVOCATIONAL/GRAMMAR        -1.21451405   0.225  0.6901693
eduSECONDARY                 -0.02837836   0.977  0.9927605
eduPOST-SECONDARY OR HIGHER    -2.26705978   0.023  0.4216578

Measures for one synthesis and 6 coefficients
Mean confidence interval overlap:  0.7000921
Mean absolute std. coef diff:  1.175617
Lack-of-fit: 18.00781; p-value 0.006 for test that synthesis model is compatible
with a chi-squared test with 6 degrees of freedom
```



R패키지 sysnthpop이 제공하는 함수로는 SBR 재현자료에 대하여 신뢰구간 중복 관련 결과를 산출할 수 없어서 R패키지 dotwhisker를 사용하여 유사한 함수를 구현하여 사용하였다. 그 결과는 다음과 같다.



<그림 5-2> SBR 재현자료에 대한 dot-whisker plot 결과

대부분의 경우 신뢰구간 중복이 유의미하게 이루어지므로 선형 회귀 분석 결과가 통계적으로 현저히 차이가 난다고 보기 어렵다. 또한 재현자료 이용 목적이 원자료를 대체하기 위함이 아니라, 데이터센터를 방문하여 원자료를 분석하기 이전의 사전 분석용이라면 어느 정도 수준의 정보손실은 용인될 수 있다고 판단된다.

제2절 종사자-기업체 연계 DB(SBRC)의 재현자료 생성

1. 종사자-기업체 연계 DB(SBRC) 개요

종사자-기업체 연계 DB(이하 SBRC)는 종사자의 기본 정보, 근무지에 대한 정보, 자료의 출처 및 추가 정보로 구성되어 있다. 종사자 기본 정보로는 개인대체식별번호, 연령, 성별, 주거지 행정 구역 코드 등과 더불어 각 근무지별 시작일자, 종료일자, 종료일자를 기준으로 산출된 연말근무여부, 근무일수, 근무월수 등이 있다. 종사자 근무지에 대한 정보로는 기업체식별번호, 근무지행정구역코드, 조직분류, 대산업분류, 중산업분류 및 해당 기업체의 종사자 수 등이 있다. 한편 사회보험가입여부 변수는 자료출처 변수를 기준으로 생성된다. 이 변수는 분석 목적에 따라 종사자가 가입한 사회보험 종류의 정보로 활용할 수 있다.

2018년 기준으로 SBRC 총 49,317,030건 중에서 조직구분이 개인(1) 혹은 법인(2)로 분류된 자료 49,218,231건이 재현자료 생성 대상 원자료가 된다. 이 중 개인대체식별번호가 존재하는 건수는 48,362,034이며, 종사자수로는 25,727,852명이 된다. 개인대체식별번호를 가지는 25,727,852명의 개별 종사자 중에서 7,728,738명이 2018년에 2개 이상의 근무지 기록을 가지고 있으며, 건수를 기준으로는 약 60%인 30,381,487건이 된다. 개인대체식별번호가 없거나 2018년에 1개의 근무지 기록만을 가지는 종사자는 총 18,836,744건으로 구성되어 있다. 개인대체식별번호가 없는 종사자는 1개의 근무지 기록만을 가지는 것으로 가정하였다.

<table>
<tr><th>원자료 건수</th><th>개인대체식별번호</th><th>근무지</th></tr>
<tr><td rowspan="3">총 : 49,218,231 건</td><td rowspan="2">있음 : 48,362,034 건</td><td>2개 이상 : 30,381,487 건</td></tr>
<tr><td rowspan="2">1개 기록 : 18,836,744 건</td></tr>
<tr><td>없음 : 856,197 건</td></tr>
</table>

다음 <표 5-3>에서는 재현자료 생성에서 고려한 변수를 정리하였다.

<표 5-3> 종사자-기업체 연계 DB (2018)

no	항목 한글명	항목 영문명	속성	자료 길이	Null 건수	unique (NA포함)
1	기준연도	BASE_YEAR	문자	4	-	1
2	기업체고유번호	BR_GI_KEY	문자	10	1,857,275	1,859,295
3	개인대체식별번호	SRTY_DSMT_NUM	문자	1	857,275	2,572,785
4	연령(만 나이)	FULL_APS	숫자	2	244,644	108
5	성별	SD_CD	문자	1	185,053	3
6	거주지(행정구역코드)	HM_AD_CLS_CD	문자	7	2,221,566	3,740
7	근무지(행정구역코드)	WR_AD_CLS_CD	문자	7	217,195	3,509
8	자료출처	GB	문자	1~5	-	58
9	사회보험가입여부	WRK_JOIN_YN	문자	1	-	2
10	연말근무여부	YEND_WRK_YN	문자	1	30,692,193	2
11	시작일자	ACQ_DATE	문자	8	22,745,644	14,646
12	종료일자	LOSS_DATE	문자	8	22,745,644	369
13	당해연도근무월수	FNL_WRK_MM_CNT	숫자	1~2	-	12
14	당해연도근무일수	WRK_DD_CNT_TOT	숫자	1~3	-	365
15	조직구분	BR_JOJIK	문자	1	98,798	3
16	기업체_산업분류(대)	BR_GI_LCLS_CD	문자	1	98,798	23
17	기업체_산업분류(중)	BR_GI_MCLS_CD	문자	2	183,727	77
18	기업체_종사자수	BR_GI_EMP_T	숫자	1~6	98,798	1,719

2. 재현자료 생성

R패키지 synthpop을 이용해 SBRC의 재현자료를 생성하기 위하여 하위그룹을 분류하는 방식은 다음과 같다. 먼저 개인대체식별번호(SRTY_DSMT_NUM)가 없는 종사자는 개인대체식별번호(SRTY_DSMT_NUM)가 존재하는 다른 기록과 연결할 만한 근거가 없으므로, 각기 다른 개인으로 가정하고 해당 연도 내에 1개의 근무지를 가지는 그룹으로 간주한다. 개인대체식별번호(SRTY_DSMT_NUM)가 존재하는 레코드의 경우에는 개인대체식별번호(SRTY_DSMT_NUM)로 묶어 관찰빈도(n)를 구한다. 근무지가 1개로 분류된 그룹(ids.sing)은 2018년 기준 약 1,800만 명으로 파악된다. ids.sing 그룹에 대하여 근무지행정구역코드1(WR_AD_CLS_CD_1) 및 2(WR_AD_CLS_CD_2)로 하위 그룹을 생성한다. 개인대체식별번호(SRTY_DSMT_NUM)가 존재하는 종사자 중 2개 이상의 근무지를 가지는 나머지 종사자 그룹은 2018년 기준 약 3,000만 명으로, 근무지 중복을 허용하면 최소 2개에서 최대 143개 근무지를 가지는 것으로 나타났다. 근무지가 2개 이상으로 분류된 그룹(ids.multi)은 총 근무지 수(n)와 근무순서(order) 변수를 사용하여 하위 그룹을 생성한다. SBRC에서 종사자의 정보를 최대한 보존하여 재현자료를 생성하기 위하여 이와 같이 하위 그룹을 형성하였다. 다음으로 재현자료 생성 순서는 기업체행정지역(WR_AD_CLS_CD_1, WR_AD_CLS_CD_2), 산업분류(BR_GI_LCLS_CD, BR_GI_MCLS_CD), 조직분류(BR_JOJIK) 및 종사자 정보 변수들(GB, HM_AD_CLS_CD_1, HM_AD_CLS_CD_2, FULL_APS, SD_CD, ACQ_DATE, LOSS_TYPE, LOSS_DATE)이다. 최종 노드 수(minbucket)를 최대한 충분히 확보할 수 있도록 자료의 범위가 가장 넓은 종사자 수(BR_GI_EMP_T)를 마지막에 배치하였다.

재현자료 생성 조건을 설정하는 rules 및 rvalues는 다음과 같이 결정하였다. 일용직(GB가 4, 7 또는 47)인 경우, 시작일자(ACQ_DATE), 종료타입(LOSS_TYPE), 종료일자(LOSS_DAYS) 관련 정보가 전무하여 NA로 설정하였다. 또한, 시도코드가 없는 경우에는 시군구코드도 생성될 수 없기 때문에 이를 NA로 처리하도록 설정하였으며, 시작일자(ACQ_DATE)가 NULL인 경우, 종료일자도 NA 처리하도록 하였다. 관련 코드는 다음과 같다.

```
# RULES AND VALUES
rules <- list(LOSS_DAYS="GB=='4' | GB=='47' | BG=='7' | is.na(ACQ_DAYS)",
              HM_AD_CLS_CD_2="is.na(HM_AD_CLS_CD_1)",
              WR_AD_CLS_CD_2="is.na(WR_AD_CLS_CD_1)")
rvalues <- list(LOSS_DAYS=NA, HM_AD_CLS_CD_2=NA, WR_AD_CLS_CD_2=NA)
```

한편 SBRC 재현자료 생성 과정에서 numtocat 기능은 정수형 자료에만 적용하였으며, 관련 코드는 다음과 같다.

```
# DATA TYPE
cats <- c("SD_CD","HM_AD_CLS_CD_1","GB","BR_JOJIK","BR_GI_LCLS_CD","WR_AD_CLS_CD_1")
conts <- c(colnames(ids.sing)[!colnames(ids.sing) %in% cats])
for (col in colnames(ids.sing)) {
  if (col %in% cats) {
    ids.sing[[col]] <- as.character(ids.sing[[col]])
  } else {
    ids.sing[[col]] <- as.numeric(ids.sing[[col]])
  }
}

# numtocat VARIABLES
ls.numtocat <- c("FULL_APS","ACQ_DAYS","LOSS_DAYS","BR_GI_EMP_T")
```

SBRC의 재현자료를 생성하는 과정을 정리하면 다음과 같다. 재현 모형을 고려할 때, 종사자의 기본 정보를 우선적으로 생성하고, 생성된 종사자의 정보를 설명변수로 사용하여 근무지 관련 변수를 생성하였다. 만약 종사자별 근무지 수를 고려하지 않고 자료를 생성하는 경우라면, 각각의 기록(건)이 동일인의 기록인지 아닌지를 구분할 수 없으므로 종사자의 이직 및 취직 활동 전반에 관한 정보가 보존되지 못한다. 뿐만 아니라 총 4,900만여 명의 종사자 수가 산출되지 않으며, 종사자별로 동일한 기업체에서 퇴직한 이후에 재근무한 정보를 보존하지 못하므로 기업체 수 역시 종사자 수만큼 커지는 현상도 발생했다. 이를 보완하기 위하여 기준연도별 총 근무지수(n), 근무순서 및 각 종사자별 근무지(RB_ID)를 구분하여 기업체고유번호(BR_GI_KEY)를 설정하였다.

다음으로 시작일자(ACQ_DATE)는 이전 근무지의 종료일자(LOSS_DATE)로부터 기간(일)을 계산하였으며, 근무순서가 1인 자료에 한하여 2018년 기준 가장 빠른 시작일자로부터 기간(일)을 산출[14]하여 ACQ_DAYS라는 변수를 새로 생성하였다. 종료일자(LOSS_DATE)는 해당 기업체에 근무한 시작일자(ACQ_DATE)로부터 근무한 기간(day)으로써, LOSS_DAYS라는 새로운 변수를 만들고 이를 이용하여 재현자료를 생성하였다. 또한 종료일자가 누락(NA)이 아닌 다른 이유로 알 수 없는 경우나 계속 근무하고 있는 경우는 '88881231', '99991231', '99999999'로 표시되었는데, 이를 나타내는 LOSS_TYPE이라는 새로운 변수를 만들어 재현조건으로 사용하였다.

재현자료 기준연도 내 근무지 수가 1개인 그룹(ids.sing)과 근무지 수가 2개 이상인 그룹(ids.multi) 각각에서 재현자료 생성이 이루어진다. 먼저 ids.sing은 근무지행정구역코드1(WR_AD_CLS_CD_1)과 근무지행정구역코드2(WR_AD_CLS_CD_2)에 의하여 결정된 하위 그룹에서 먼저 CTREE 방식으로 재현자료를 생성하고, 에러가 발생하면 CART 방식으로 전환하여 다시 재현자료 생성을 시도한다. 만약 CART 방식까지 실패한 경우 이러한 레코드들만 따로 모아서 근무지행정구역코드1(WR_AD_CLS_CD_1)

14) 향후 재현자료 생성 모형을 개선할 때는 출생년도보다 빠른 시작일자를 가질 수 없도록 전처리 후 rules/rvalues를 설정하는 것을 추천한다.

로만 하위 그룹을 정하고 다시 재현자료를 생성한다.

기준연도 내 2개 이상의 근무지를 가지는 그룹(ids.multi)에서 재현자료 생성 방식은 조금 다르다. 하위 그룹으로 나누는 기준이 근무지 수(n)와 근무순서(order)인 만큼 재현자료가 생성되지 않는 레코드들에 대해서 다시 재현자료를 생성할 수 없다. 동일 근무지 수(n) 그룹 내 근무순서(order)별 종사자 수가 같아야 하기 때문이다. 각 하위 그룹에서 재현자료를 생성하는 코드는 다음과 같다. 먼저 근무지 수가 1개인 그룹(ids.sing), 두 번째로 이 그룹에서 재현자료가 생성되지 않은 레코드 그룹, 마지막으로 2개 이상의 근무지를 가지는 그룹(ids.multi) 각각에 대한 코드를 차례로 나열하였다.

```
# ids.sing METHODS
method.cart<-rep(c("","cart"), times=c(2,ncol(temp)-2))
method.ctree<-rep(c("","ctree"), times=c(2,ncol(temp)-2))

# SYNTHESIS CART & CTREE
tryCatch (
  {
    temp.syn  <- syn(temp, m=1, seed=1234, method=method.ctree, ctree.minbucket=3,
                     visit.sequence=seqs, numtocat=ls.numtocat, rules=rules,
rvalues=rvalues)
  },
  error = function(e) {
    tryCatch (
      {
        message("CTREE FAILED. Trying CART....")
        temp.syn_ <- syn(temp, m=1, seed=1234, method=method.cart, cart.minbucket=3,
                         visit.sequence=seqs, numtocat=ls.numtocat, rules=rules,
rvalues=rvalues)
        message("CART SUCCEEDED.\n")
      },
      error = function(e) {
        message("CART ALSO FAILED.\n")
        # CART와 CTREE 모두 재현에 실패한 데이터를 따로 모아둠
        fname <- paste('./output/sbrc/resid/sbrc_WR_',i,"_WR2_",l,".csv",sep="")
        write.csv(temp,fname,row.names=F)
      },
      finally = {
        message("")
      }
    )
  },
  finally = {
    message("")
  }
)
```

```
## CART와 CTREE 모두 재현에 실패한 데이터(resid) 불러오기
fnames <- list.files("./output/sbrc/resid", pattern="*.csv",full.names=T)
ldf <- lapply(fnames, fread)
resid <- do.call(rbind, ldf)

# DATA TYPE
cats <- c("SD_CD","HM_AD_CLS_CD_1","GB","BR_JOJIK","BR_GI_LCLS_CD","WR_AD_CLS_CD_1")
conts <- c(colnames(resid)[!colnames(resid) %in% cats])
for (col in colnames(resid)) {
  if (col %in% cats) {
    resid[[col]] <- as.character(resid[[col]])
  } else {
    resid[[col]] <- as.numeric(resid[[col]])
  }
}

# numtocat VARIABLES
ls.numtocat <- c("FULL_APS","ACQ_DAYS","LOSS_DAYS","BR_GI_EMP_T")

# RULES AND VALUES
rules <- list(LOSS_DAYS="GB=='4' | GB=='47' | BG=='7' | is.na(ACQ_DAYS)",
              HM_AD_CLS_CD_2="is.na(HM_AD_CLS_CD_1)",
              WR_AD_CLS_CD_2="is.na(WR_AD_CLS_CD_1)")
rvalues <- list(LOSS_DAYS=NA, HM_AD_CLS_CD_2=NA, WR_AD_CLS_CD_2=NA)

# VISIT.SEQUENCE
seqs <- c(1:ncol(resid))

# CART와 CTREE로 댜시 재현
JS <- unique(resid$WR_AD_CLS_CD_1)
for (i in JS) {
  if (is.na(i)) {
    temp <- subset(resid, is.na(WR_AD_CLS_CD_1))
  } else {
    temp <- subset(resid, WR_AD_CLS_CD_1==i)
  }
    # subgroup temp 재현코드
  }
}
```

```
# ids.multi METHODS
method.cart<-rep(c("","cart"), times=c(2,ncol(temp)-2))
method.ctree<-rep(c("","ctree"), times=c(2,ncol(temp)-2))

# SYNTHESIS CART & CTREE
tryCatch (
  {
    temp.syn_ <- syn(temp, m=1, seed=1234, method=method.ctree, ctree.minbucket=3,
                     visit.sequence=seqs, numtocat=ls.numtocat, rules=rules,
rvalues=rvalues)
  },
  error = function(e) {
    tryCatch (
      {
        message("CTREE FAILED. Trying CART....")
        temp.syn_ <- syn(temp, m=1, seed=1234, method=method.cart, cart.minbucket=3,
                         visit.sequence=seqs, numtocat=ls.numtocat, rules=rules,
rvalues=rvalues)
        message("CART SUCCEEDED.\n")
      },
      error = function(e) {
        message("CART ALSO FAILED.\n")
      },
      finally = {
        message("")
      }
    )
  },
  finally = {
    message("")
  }
)
```

참고로 SBRC의 재현자료를 분석할 때 다음의 사항을 주의할 필요가 있다. SBRC의 기업체고유번호(BR_GI_KEY)는 종사자별로 다르게 지정되었으므로 종사자 간 기업체고유번호(BR_GI_KEY)가 같다고 하더라도 동일 근무지로 간주하는 것은 바람직하지 않다. 또한 자료구분 및 매칭구분은 자료의 출처 및 연계과정에서 자료 매칭 여부를 임의로 표기한 것이거나 자료의 출처를 통한 사회보험 가입여부를 추가로 정제한 결과를 임의로 재현한 것이다. 따라서 실제와 차이가 발생할 수 있으므로 분석 시 주의가 필요하다.

재현자료 생성에 관련된 사항을 변수별로 간단히 정리하여 설명하면 다음과 같다. 참고로 <표 5-4>에서는 전처리 과정을 거쳐 재현자료 생성에 사용한 변수명을 한눈에 볼 수 있도록 정리하였다. 최종 산출물에 포함하도록 결정한 변수는 진하게 표시하였다.

기준연도(BASE_YEAR)
재현자료 생성 후 입력된 자료의 기준연도를 일괄적으로 부여한다.

기업체고유번호(BR_GI_KEY)
종사자별 각기 다른 기업체 고유번호가 부여되며, 종사자 간 동일한 기업체 고유번호를 가진다고 하더라도 동일한 기업체가 아닐 수 있다. 아래에서 보듯이, 종사자 1의 근무지 1은 종사자 2의 근무지 1과 동일하지 않을 수 있으며, 이는 종사자별 근무한 기업체 수와 특징, 그리고 이직 패턴 정보를 보존하기 위함이다.

SRTY_DSMT_NUM (개인대체식별번호)	BR_GI_KEY (기업체식별번호)	종사자 정보	기업체 정보
1	1	...	...
1	1	...	...
1	2	...	...
2	1	...	...
2	2	...	...

SBRC는 특정 기업체의 종사자 정보를 분석하기 위한 목적으로 사용되어야만 하며, 특정 기업체를 식별하거나 또는 해당 기업체에 근무하는 특정 개인을 식별할 수 없다.

개인대체식별번호(SRTY_DSMT_NUM)
종사자의 근무지 수와 근무 순서에 따라 그룹별 랜덤으로 생성된다.

만 나이(FULL_APS)
연속형 변수로 설정하고, R패키지의 **numtocat**[15] 기능을 사용해 재현자료를 생성한다.

성별(SD_CD)
범주형 변수로 설정하여 재현자료를 생성한다.

---

15) numtocat 기능을 SBRC에 사용한 결과, 사용자가 분류되는 요인을 설정할 수 없기 때문에 생성된 재현자료가 원래 데이터 분포와 다르거나 잘못된 값인 경우들이 관찰되었다. 만 나이(FULL_APS)가 원자료가 가지는 최대값을 넘어서게 생성된 경우 모두 최대값으로 변경하였다.

주거지행정구역코드(HM_AD_CLS_CD)
R패키지가 범주형 변수로 허용하는 60개 요인 수준 수보다 많기 때문에 시/도를 나타내는 최초 2자리 숫자인 HM_AD_CLS_CD_1과 시/군/구를 나타내는 다음 3자리인 HM_AD_CLS_CD_2로 나누어 작업한다. 시/군/구 이하 행정구역코드는 재현자료를 생성하지 않는다.

근무지행정구역코드(WR_AD_CLS_CD)
위의 주거지행정구역코드와 마찬가지로 WR_AD_CLS_CD_1(시/도) 및 WR_AD_CLS_CD_2(시/군/구)로 나누어 작업하며, 시/군/구 이하 행정구역코드는 재현자료를 생성하지 않는다.

자료출처(GB)
조사자료의 출처를 나타내며, 주로 일용직을 구분하기 위해 사용되었다. 자료의 출처는 오류가 있을 수 있지만, 사회보험가입여부 및 종류를 개략적으로 설명할 수 있는 변수이므로 재현하는 변수로 포함하였다. 범주형 변수로 정의한 뒤 재현자료를 생성한다.

시작일자(ACQ_DATE)
종사자별 이전 근무지의 종료일자 이후 다음 근무지로 이직까지 걸린 기간(일)을 나타내는 구직기간(ACQ_DAYS) 변수를 새로 만들어 재현자료 생성에 이용하였다. 이 변수는 구직기간 또는 구직까지 걸린 시간으로 이해할 수 있다. 단, 근무순서가 1인 자료는 자료의 최소 시작일자를 기준으로 시간(일)으로 산출하여 최초 근무시작일은 따로 구분하였다. 연속형 변수로 설정하고 **numtocat** 기능을 사용하여 재현자료를 생성한다.

종료일자(LOSS_DATE)
종사자별 근무지의 시작일로부터 종료일까지 걸린 기간(일)을 계산한 것으로, 해당 기업체에서의 근무기간을 나타내는 LOSS_DAYS를 새로 만들어 재현자료 생성에 이용하였다. 기준연도의 12월 31일 외에도 '88881231', '99991231', '99999999' 등 종료일이 불확실한 자료들을 구분하기 위해 각각에 0~3까지 코드를 부여하여 종료타입(LOSS_TYPE)이라는 새로운 변수를 생성하였다. LOSS_TYPE이 1~3의 값을 가지는 자료는 LOSS_DAYS를 해당연도의 12월 31일을 종료일자로 임의 가정하여 산출하나, 원자료 형태로 복원할 시 '88881231', '99991231', '99999999'로 복원된다. 연속형 변수로 설정되어 **numtocat** 기능을 사용하여 재현자료를 생성한다.

조직구분(BR_JOJIK)
SBRC는 조직구분(BR_JOJIK)이 개인(1) 또는 법인(2)인 경우만 필터링하여 재현자료를 생성한다. 지역별, 산업별, 또는 종사자 그룹별 근무지 정보를 보존하기 위해 범주형 변수로 정의하여 재현자료를 생성한다.

기업체 산업분류-대(BR_GI_LCLS_CD)
A~W까지 값을 가지며, 범주형 변수로 정의하여 재현자료를 생성한다.

기업체 산업분류-중(BR_GI_MCLS_CD)
범주형 변수이나, R 패키지 synthpop이 허용하는 요인 수준의 수를 넘기 때문에 연속형 변수로 설정하고 재현자료를 생성한다.

기업체 종사자 수(BR_GI_EMP_T)
연속형 변수로 설정하고 **numtocat** 기능을 사용하여 재현자료를 생성한다.

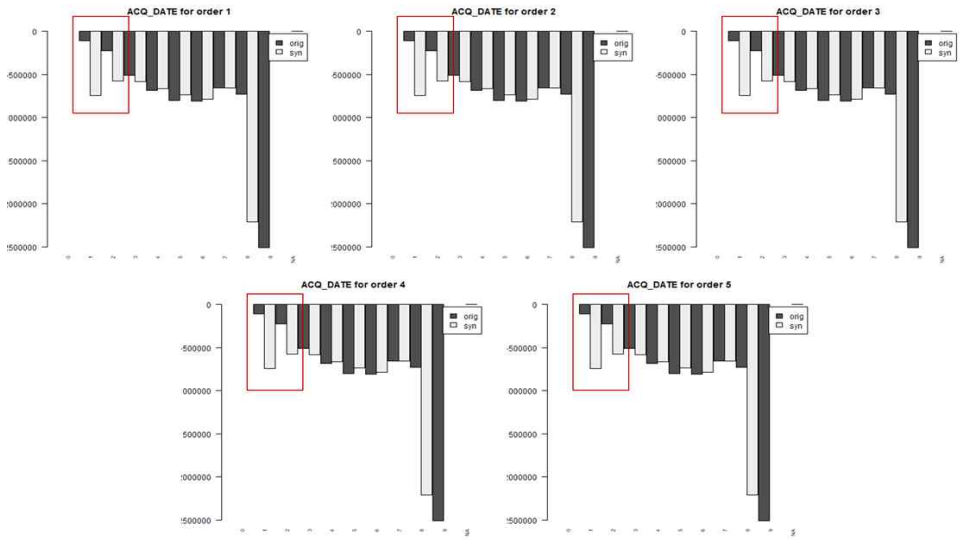
<표 5-4> 종사자-기업체 연계DB 변수별 전처리 및 재현 방법

no	항목 한글명	항목 영문명	속성	재현	재현여부 및 방법
1	**기준연도**	**BASE_YEAR**		**X**	**재현자료 생성 후 해당연도 일괄 부여**
2	**기업체고유번호**	**BR_GI_KEY**		**X**	**재현자료 생성 후 종사자별 BR_ID로 대체**
3	**개인대체식별번호**	**SRTY_DSMT_NUM**		**X**	**재현자료 생성 후 종사자별 식별번호 부여**
4	**연령(만 나이)**	**FULL_APS**	**연속형**	**O**	**재현**
5	**성별**	**SD_CD**	**범주형**	**O**	**재현**
6	**근무지 수(근무지 중복 허용)***	**n**	**범주형**	**△**	**subgroup 구분 조건으로 사용**
7	근무지*	BR_ID	범주형	O	재현
8	(근무지별) 근무 순서*	order	범주형	△	subgroup 구분 조건으로 사용
9	**거주지(행정구역코드)**	**HM_AD_CLS_CD**	**범주형**	**△**	**재현**
10	**근무지(행정구역코드)**	**WR_AD_CLS_CD**	**범주형**	**O**	**재현**
11	**자료출처**	**GB**	**범주형**	**O**	**재현**
12	**사회보험가입여부**	**WRK_JOIN_YN**		**X**	**재현자료 생성 후 자료구분이 1~3을 포함하는 경우 Y, 포함하지 않는 경우 N 부여**
13	**연말근무여부**	**YEND_WRK_YN**		**X**	**재현자료 생성 후 LOSS_DATE가 '1231'을 포함하는 경우 Y, 포함하지 않는 경우 N 부여**
14	**시작일자**	**ACQ_DATE**	**연속형**	**O**	**종료일부터 다음 근무지 시작일까지 걸린 기간(일)을 산출하여 재현**
15	종료타입*	LOSS_TYPE	범주형	O	0은 2018년 이내 종료일자가 존재하는 경우, 1~3은 누락 외 종료일자 종류

no	항목 한글명	항목 영문명	속성	재현	재현여부 및 방법
16	종료일자	LOSS_DATE	연속형	O	근무 후 퇴직까지 걸린 기간(일)을 산출하여 재현
17	당해년도근무월수	FNL_WRK_MM_CNT		X	재현자료 생성 후 개월 수 계산
18	당해년도근무일수	WRK_DD_CNT_TOT		X	재현자료 생성 후 일 수 계산
19	조직구분	BR_JOJIK	범주형	O	재현
20	기업체_산업분류(대)	BR_GI_LCLS_CD	범주형	O	재현
21	기업체_산업분류(중)	BR_GI_MCLS_CD	연속형	O	재현
22	기업체_종사자수	BR_GI_EMP_T	연속형	O	재현

마지막으로 재현자료 생성 과정에서 CTREE 방식을 우선적으로 사용하면서 CART 및 CTREE 방식을 모두 도입한 이유를 좀 더 상세히 설명하면 다음과 같다. 이러한 분류 알고리즘은 분할의 적합성을 측정하는 기준이 각각 다르다. 공통적으로는 주어진 부분집합에서 가능한 데이터 분할의 경우마다 목표 변수의 동질성(homogeneity)을 측정하여 어느 분할이 적합한지 판단한다. 이는 공변량(covariate) 집합의 값을 기반으로 일변량 종속변수 분할을 재귀적으로 수행하여 이루어진다. 다만 분할 변수의 선택 및 분할 프로세스에 대해 적용하는 규칙이 알고리즘마다 다르다.

분류회귀트리 방식인 CART는 공변량 집합에 이질적인 요소가 얼마나 포함되었는지를 측정하는 지표로 지니 불순도(Gini impurity)를 사용한다. 이는 어떤 집합에서 한 항목을 뽑아 해당 라벨을 무작위로 추정할 때 틀릴 확률을 말한다. 집합에 존재하는 항목이 모두 같다면 지니 불순도는 최소값(0)을 가지며, 이 집합은 완전히 순수하다고 말한다. 일반적으로 트리 기반 알고리즘은 각 노드에서 부분 최적값을 찾아낼 때 휴리스틱 기법을 이용하므로 최적 트리 선택을 보장하기 어렵다. 특히 CART는 분할이 많거나 결측값이 많은 변수를 선택하는 등, 분할변수 선택에 있어서 편향성을 보인다고 알려져 있다(Strobl, 2014). 예를 들면 다음 <그림 5-3>에서 SBRC의 연령대별 근무시작일 생성 비교 결과와 같이, CART 방식을 이용하면 0~20세 사이의 이상치를 과대 생성하는 결과를 얻을 수도 있다.



<그림 5-3> 5개 근무기록을 가지는 그룹의 연령대별 근무시작일 생성 비교

이러한 CART의 문제점을 조건부 추론으로 해결하고자 제안된 알고리즘이 CTREE이다(Hothorn et al., 2006). 지니 불순도, 정보 획득량(information gain) 등, 정보 측정

측도를 최대화하는 변수를 찾는 rpart류의 알고리즘과는 다르게, CTREE는 과적합을 피하는 변수 선택을 위하여 유의성 검증을 실시한다.[16] 이는 알고리즘이 주어진 공변량에 대하여 분할의 과정을 재귀적으로 수행할 때 모든 시작 부분에서 순열을 검정한다는 의미이다. 즉, 분류된 결과인 관찰값의 라벨을 재배열하여 산출한 검정 통계량을 모두 계산함으로써 귀무가설하에서 검정 통계량의 분포를 산출한다. 만약 공변량과 종속변수가 모두 수치형인 경우, 가능한 모든 종속변수의 순열과 공변량 사이의 상관관계를 계산한다. 다음으로 각각의 순열 검정에 유의확률을 산출하고, 다른 공변량에 대한 유의확률과 비교한다. 공변량과 종속변수가 모두 범주형인 경우에는 검정 통계량이 분할표에서 계산된다(Christiano & Gina, 2020). 이제 주어진 공변량 집합에서 각 공변량에 대한 유의확률을 계산하고 가장 작은 유의확률을 가진 공변량을 선택한다. 공변량을 선택하고 나면, 순열 검정을 다시 수행하여 미리 설정된 minbucket 기준을 만족하는 가능한 모든 분할을 탐색한다.

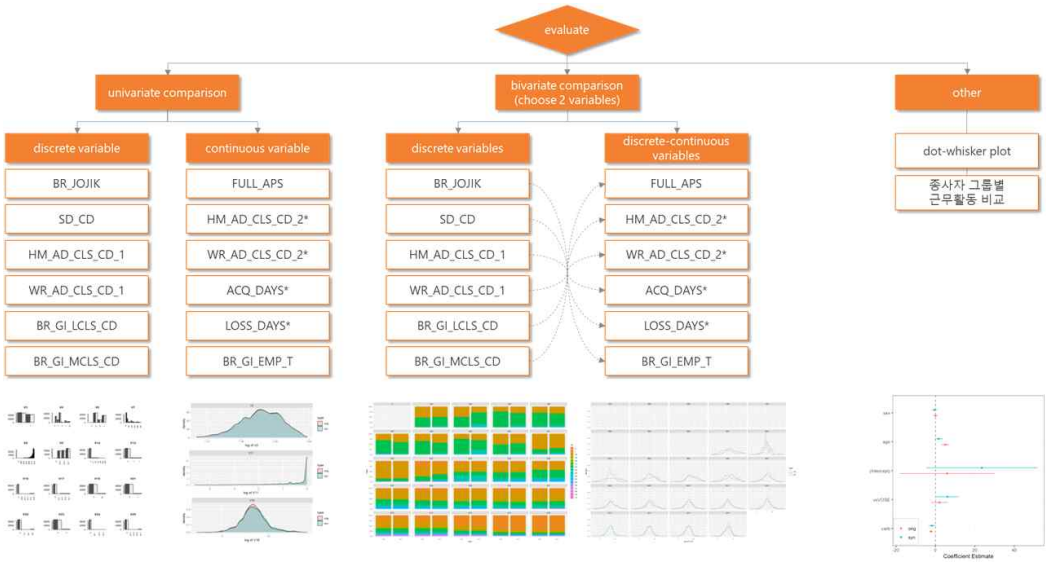
정리하면 두 알고리즘의 차이점은 CTREE가 유의성 검정 통계 이론을 이용한 공변량 선택 체계를 사용하여 CART의 잠재적 편향을 피한다는 것이다. 이에 기입 오류 등 노이즈가 많은 SBR 혹은 SBRC에 대하여 재현자료를 생성할 때 CTREE를 우선적으로 적용하고, CTREE가 적용되지 않는 부분 집합에 대하여 CART를 이용하였다. 이는 원본과 동일한 자료를 최소한으로 포함되도록 하여 노출위험을 낮추기 위함이다. R패키지 synthpop은 R패키지 rpart의 CART(Classification and Regression Tree) 및 R패키지 party의 CTREE(Conditional Inference Trees) 기능을 차용하고 있다.

3. 재현자료 평가

재현자료의 노출위험과 정보손실을 측정하여 재현자료를 평가한다. 먼저 노출위험의 경우, SBRC는 종사자의 근무순서에 따른 근무기간 등이 달라지므로 레코드(행) 단위로 노출위험을 비교하여 산출하기 어렵다. 특정 종사자를 식별하기 위해서는 종사자 정보뿐만 아니라 근무지수, 근무순서 및 순서별 근무기간과 근무지 정보가 모두 일치해야 하므로 노출위험이 낮다고 판단된다. 그러나 노출위험 측정에 관한 명확한 판단을 위해서는 학계의 연구가 더 필요한 실정이다.

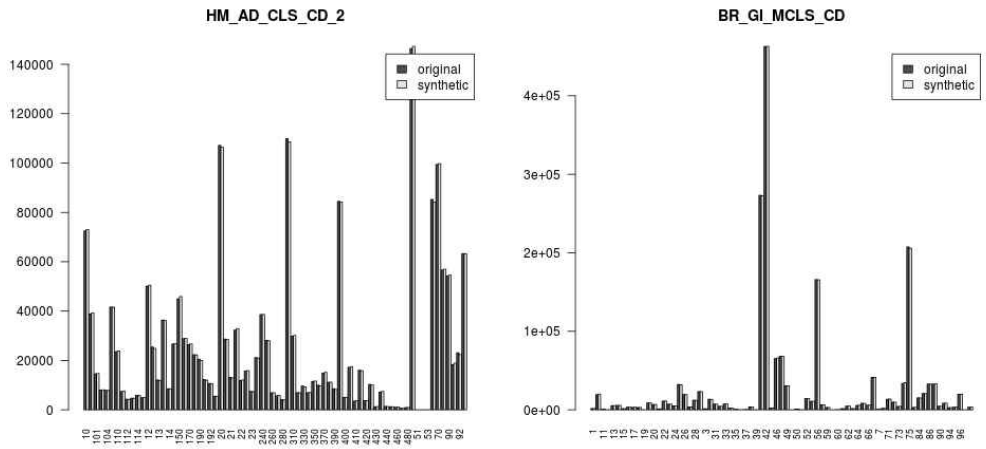
한편 재현자료의 정보손실을 측정하는 방식은 매우 다양하다. 이번 프로젝트에서는 단변량 비교, 이변량 비교, 신뢰구간 비교 등을 수행하였다. 이를 차례로 살펴보면 다음과 같다. 참고로 SBRC의 경우 아래 그림에 정의된 변수 유형에 따라 하단에 나타나는 그림들을 활용한 시각적인 비교를 수행할 수 있다.

16) R package ‘party’ (https://cran.r-project.org/web/packages/party/vignettes/party.pdf)

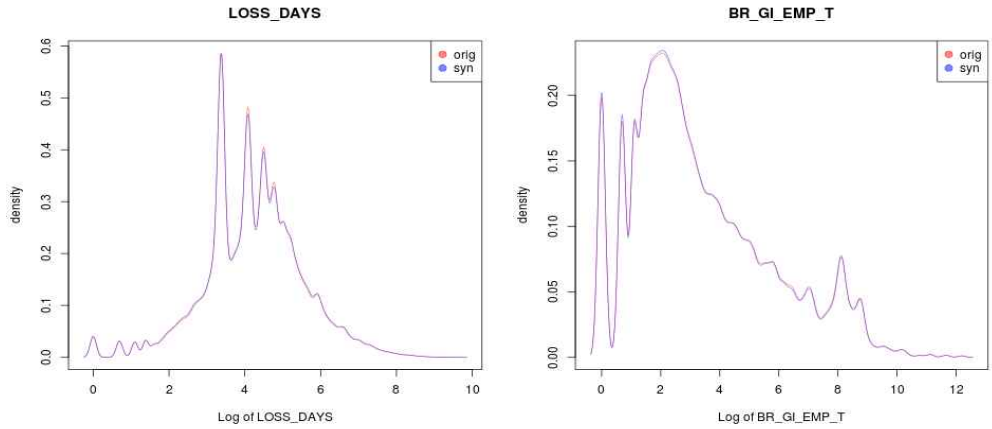


가. 단변량 비교(Comparison of univariate distributions)

R패키지 synthpop의 compare() 함수 기능을 사용할 수 있으나 메모리의 문제로 유사한 함수를 구현하여 사용하였다. SBR 일부 범주형 변수에 대하여 도수분포표를 비교한 결과는 다음과 같다.

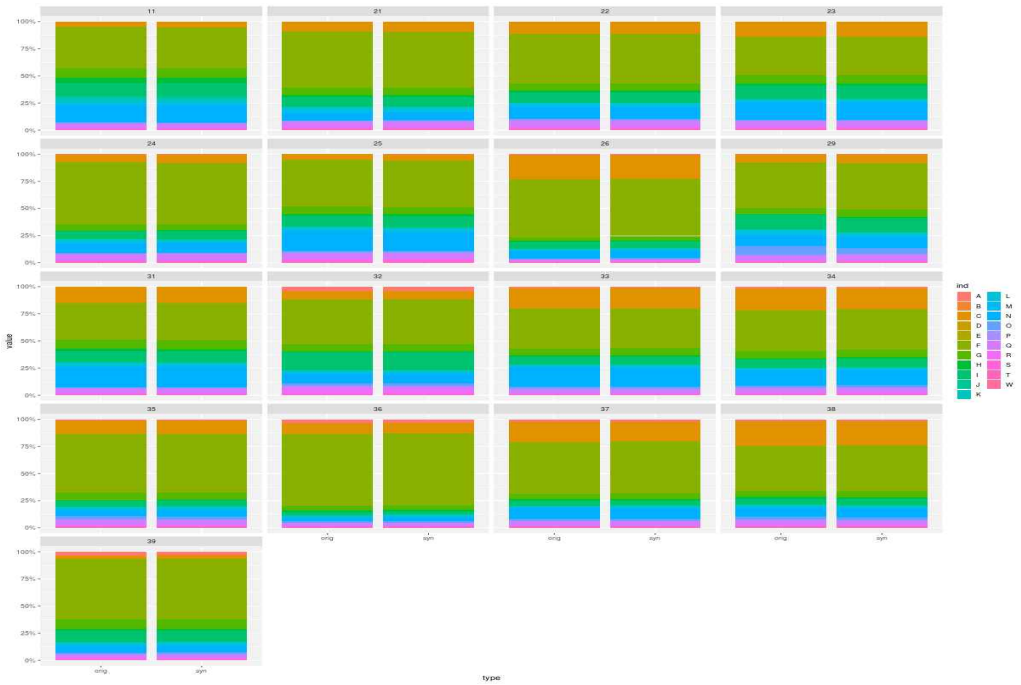


한편 SBRC 일부 연속형 변수에 대하여 분포를 비교한 결과는 다음과 같다. 원자료에서 각 변수의 분포가 비대칭이 심하면서 한쪽 꼬리가 길므로 로그 변환하였다.

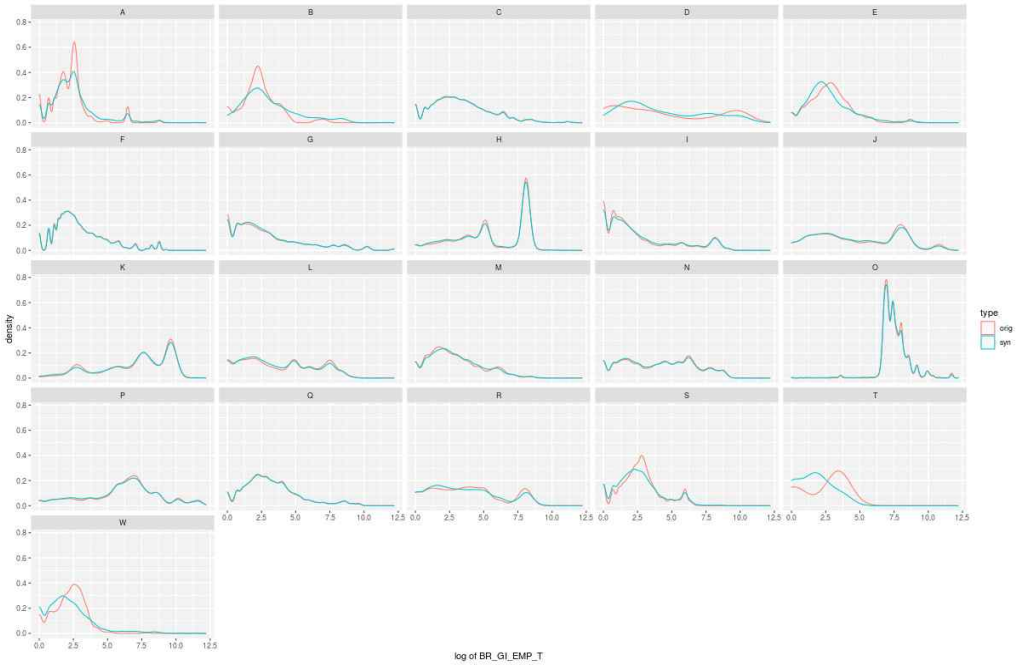


나. 이변량 비교(Comparison of bivariate distributions)

SBRC 일부 범주형-범주형 변수 및 범주형-연속형 변수에 대한 비교는 다음과 같다. 대부분의 이변량 변수 비교에 대하여 SBR 재현자료는 적은 정보손실을 가지는 것으로 나타난다.



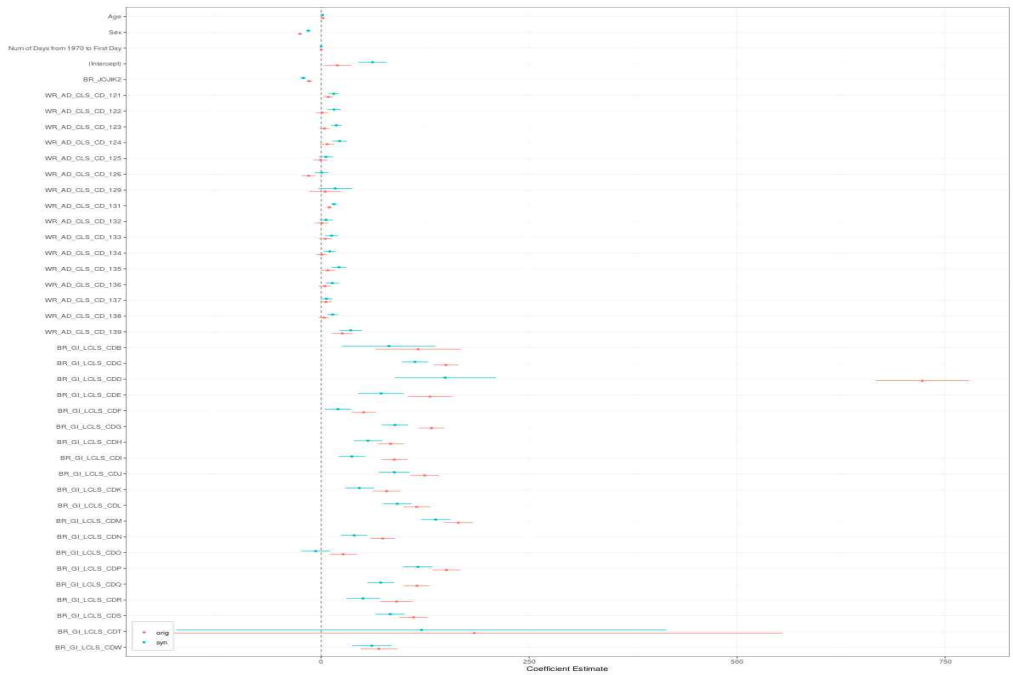
[그림 5-4] BR_GI_LCLS_CD 및 WR_AD_CLS_CD_1의 분포



<그림 5-5> BR_GI_LCLS_CD별 BR_GI_EMP_T의 분포

다. 95% 신뢰구간 비교

R패키지 dotwhisker로 함수를 구현하여 사용하였다. 그 결과는 다음과 같다.



<그림 5-6> SBRC 재현자료에 대한 dot-whisker plot 결과

SBR 결과와 다르게 SBRC에서는 신뢰구간 중복이 유의미하게 이루어지지 않는 경우가 많았다. 이는 자료의 구조가 복잡하므로 재현자료가 원자료의 분포를 세세하게 다 보존하기 어렵기 때문이라고 추측된다. 다만 재현자료 이용 목적이 원자료를 대체하기 위함이 아니라, 데이터센터를 방문하여 원자료를 분석하기 이전의 사전분석용이라면 어느 정도 수준의 정보손실은 용인될 수 있다고 판단할 수 있다.

참고로 SBRC의 경우 원자료의 오기, 오타 등을 수정하지 않고 이상치를 그대로 사용하였으나, CART는 모든 범주를 골고루 생성하는 특징이 있으므로 정보손실이 커지는 효과가 나타났을 수 있다. 따라서 이상치를 제거하거나, 재현자료를 생성할 변수나 변수별 범주의 수를 줄이면 재현자료의 품질이 향상(정보손실이 축소)될 수 있으리라 예측된다. 한편 향후에는 시군구 변수나 산업분류 변수의 조건을 따로 설정하지 않아도 되는 syn.nested() 함수를 활용하거나, 변수별로 예측 변수를 직접 지정하여 모형의 정확성을 높이도록 predictor.matrix 매개 변수를 설정하는 방법을 생각해 볼 수 있다. 한편 SBR 및 SBRC 모두 일부 변수는 그대로 사용한 부분 재현자료이며, 이러한 재현자료의 노출위험 측정 연구는 앞으로 지속될 필요가 있다.

제 6 장

딥러닝 모형을 이용한 재현자료 생성

재현자료는 원자료와 유사한 분포를 가지도록 생성한 자료를 의미한다. 유사성을 높이기(정보손실을 줄이기) 위하여 재현자료를 생성하는 과정에서 비모수적 모형을 사용하거나 일부 자료만을 대체하는 방식을 이용해왔다. 그러나 원자료의 값이 그대로 생성되는 경우에는 노출위험이 없다는 가정을 담보하기 어려우므로, 재현자료 역시 매스킹 처리와 유사한 노출위험-정보손실 상충관계를 가지게 된다. 이를 해결하기 위하여 차등정보보호를 만족하는 재현자료 생성 방식도 연구되고 있으나, 이는 노출위험 제어 수준은 보장되지만 정보손실 발생이 과다한 측면이 있다. 따라서 노출위험-정보손실 모든 측면에서 가장 효율적인 방법에 관한 연구는 여전히 진행 중인 상황이라고 할 수 있다. 현재까지 각국의 통계기관에서는 베이지안 다중대체에 근거한 통계적 모형을 이용하여 재현자료를 생성하는 방식을 활용하여 왔으나 최근 관련 연구 영역이 확대되고 있다.

이 장에서는 정보손실을 줄여 재현자료를 생성하기 위하여 딥러닝 기법을 이용하는 연구 결과를 소개한다. 이를 위하여 우선 적대적 생성망(GAN)을 이용하여 재현자료를 생성하는 방식을 소개하고, 다음으로 후속 연구로 제안된 조건부 적대적 생성망을 이용하는 방법을 소개한다. 한편 적대적 생성망에 대한 기본적인 설명은 부록에 정리하였다. 참고로 개체별로 적재된 데이터를 통계기관에서 마이크로데이터라고 부르지만, IT분야의 문헌에서는 테이블(table) 데이터라고 언급하는 경우가 많다.

종래의 익명화 기술은 원본자료의 레코드와 익명화된 데이터의 레코드 사이에 1:1 대응 관계가 여전히 존재하고 그로 인하여 재식별화가 가능하다. 이에 대응하고자 재현자료 생성 기술들이 연구되었으며, 특히 적대적 생성망은 몇 가지 그 적합한 특성으로 인하여 재현자료 생성 기술로 각광을 받고 있다. 실제로도 다양한 유형의 데이터에 대해서 적대적 생성망 활용 기술들이 연구되고 있으며, 이미지나 텍스트에 대해서는 특히 많은 연구가 수행되었다. 하지만 테이블 데이터 또는 마이크로데이터에 대한 활용은 오랫동안 무시되어 왔다. 하지만 2018년도 VLDB 학회에 Table GAN (Park et al., 2018)이라는 기술이 공개되면서 많은 후속 연구들이 여러 연구 그룹에 의해서 이루어져 왔다. Table GAN(이하 TGAN)이 제시하는 적대적 생성망 기반의 데

이터 생성은 아래와 같은 장점이 있다.

- 원본 레코드와 재현 레코드 사이에 1:1 대응 관계가 존재하지 않는다. 따라서 재식별화가 원칙적으로 불가능하다.
- 레코드의 모든 값들이 재현된 값으로 공개되어도 안전하다.
- 잘 설계된 기술로 생성된 재현자료를 이용하여 설계/학습된 기계 학습 모델은 원본자료를 위해서 활용될 수 있다.
- 재현자료는 프라이버시 걱정 없이 외부에 공개될 수 있다.

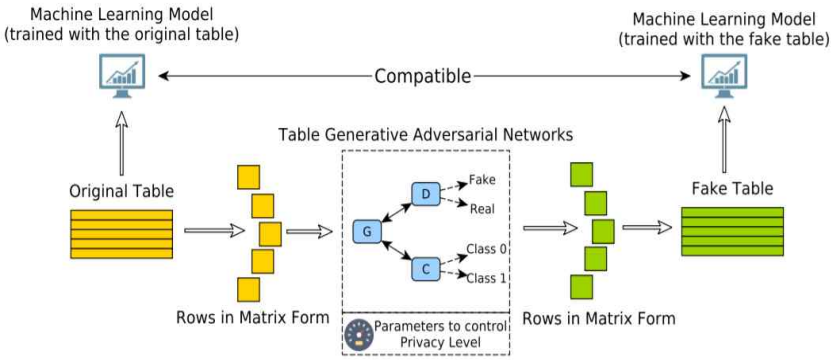
하지만 이러한 적대적 생성망 기술은 기계 학습 분류기 알고리즘을 목표로 하는 멤버쉽 공격(Shokri et al., 2017)에 취약한 측면이 있다. 종래의 익명화는 기계 학습 알고리즘이 아니라 멤버쉽 공격에는 영향을 받지 않는다. 따라서 거시적인 측면에서 보면 익명화 기술과 적대적 생성망 기반의 재현자료 생성 기술은 완전히 반대되는 성격을 기술로 이해될 수 있다.

제1절 적대적 생성망을 활용한 재현자료 생성

TGAN(Park et al., 2018)은 초창기의 재현자료 생성 모델이다. TGAN은 재현자료로 학습된 기계학습 모델과 원래 데이터로 학습된 기계학습 모델이 호환될 수 있도록 높은 성능으로 데이터를 재현하는 것을 목표로 하고 있다. 그뿐만 아니라 필요에 따라 모델의 재현성능을 인위적으로 조절할 수 있는 기능도 갖고 있다. 즉 신뢰도가 높은 상대에게는 성능을 높여 원본과 유사한 재현자료를 제공하고, 그렇지 않은 상대에게는 성능을 낮춰 원본과 차이가 큰 재현자료를 제공할 수 있다. 이러한 TGAN의 구조, 손실함수, 학습 방법, 성능 평가 등을 차례로 살펴보자.

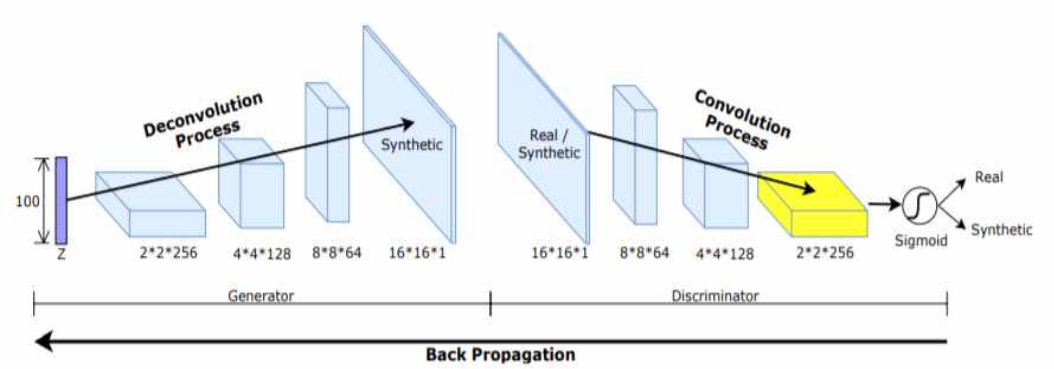
1. TGAN의 구조

<그림 6-1>은 TGAN의 구조를 한눈에 보여준다. 일반적인 적대적 생성망과 비슷한 구조를 가지지만, 몇 가지 특징적인 부분이 있다. 테이블의 레코드들을 행렬 형태로 가공하여 활용하고, 생성자(Generator, G), 식별자(Discriminator, D) 이외에도 구분자(Classifier, C)라는 신경망이 추가로 존재한다.



<그림 6-1> TGAN의 구조

TGAN은 1차원의 레코드를 2차원의 행렬 형태로 바꿔서 활용한다. 이러한 변환을 통해 초창기 적대적 생성망들 중 가장 각광받는 DCGAN(<그림 6-2>)을 활용할 수 있다. DCGAN은 이미지 데이터 재현에서 우수한 성능을 보이는 적대적 생성망 모델인데, 테이블의 각 레코드를 2차원 행렬 형태로 만듦으로써 DCGAN을 활용한다. 1차원 레코드를 2차원으로 변환해 모델을 학습하고, 생성자가 생성한 2차원 데이터를 1차원으로 환원하면서 재현 레코드를 얻는다. 2차원으로 변환할 때, 필요한 경우 0을 채워서 사용한다. 예를 들어 23개의 속성(Attribute)들이 있는 레코드를 5x5 행렬로 변환한다면 마지막에 0을 두 개 추가해서 사용한다. 생성자가 생성한 2차원 데이터를 1차원으로 바꿀 때는 0이 채워진 부분을 버리고 원래 1차원 형태로 복원해 활용한다.



<그림 6-2> 식별자와 생성자 구조

구분자는 식별자와 같은 신경망 구조를 갖지만 다른 역할을 한다. 식별자는 입력 데이터의 진위를 가리지만, 구분자는 입력 데이터의 클래스(Class)를 판단해 준다. 예를 들어 콜레스테롤 수치, 나이, 당뇨 여부 변수가 있을 때, 식별자는 3개 속성 모두

입력받아 데이터의 진위를 가린다. 반면, 구분자는 콜레스테롤 수치, 나이 변수를 받아서 당뇨 여부를 판단한다. 구분자는 의미적 무결성(Semantic integrity)이 높은 재현 자료를 생산하도록 돕는다. 클래스마다 속성들 값의 분포는 다른데, 구분자는 생성자가 클래스에 맞게 속성값들을 생성하도록 돕는다. 예를 들어, 당뇨 환자라면 콜레스테롤의 수치는 값이 큰 곳에 집중된 분포를 가질 것이다. 구분자는 재현 레코드가 당뇨 환자라면 콜레스테롤 수치가 높아지도록 만들어준다.

2. 손실 함수

TGAN에서는 일반적인 적대적 신경망의 손실 함수뿐만 아니라 목적에 맞게 설계된 추가적인 손실 함수를 사용함으로써 재현성능을 높이는 한편, 재현성능을 제어할 수 있도록 한다. 구분 손실 함수를 최소화함으로써 데이터의 의미적 무결성을 높이고, 정보 손실 함수를 최소화하여 높은 재현 성능 및 성능에 대한 제어를 얻을 수 있다.

TGAN은 아래와 같은 일반적인 손실함수를 사용해 식별자와 생성자를 훈련시킨다. 식별자는 원본과 재현 레코드들을 구분하려고 노력하고, 생성자는 원본 수준의 재현 레코드를 만들려고 노력하는 과정에서 수준 높은 재현자료가 생성된다. 아래의 L_{orig}는 제로섬 최소-최대 게임 학습 방법의 손실 함수이다.

$$L_{orig} = \min_G \max_D V(G,D) = E[\log D(x)]_{x \sim p_{data}(x)} + E[\log(1 - D(G(z)))]_{z \sim p(z)}$$

구분 손실 함수 L_{class}^{C}와 L_{class}^{G}는 구분자와 생성자를 학습하는 데 활용된다. 구분자를 학습할 때는 자료에서 클래스 속성을 제외한 나머지(아래 수식에서 $remove(x)$)를 구분자의 입력으로 넣고 도출된 클래스와 실제 클래스와의 차이를 측정한다. 그리고 이러한 차이를 줄이는 방향으로 구분자를 훈련한다. 즉, 구분자는 원본 레코드들을 활용하여 원본 레코드의 클래스 속성을 다른 속성들로 잘 추론하게 하는 훈련을 진행한다. 생성자를 학습할 때도 이와 비슷하게 손실함수를 계산한다. 다만 실제 데이터 대신 재현자료인 $G(z)$를 입력으로 넣는다는 점과 학습의 대상이 생성자라는 점만 다르다.

$$L_{class}^{C} = E[|l(x) - C(remove(x))|]_{x \sim p_{data}(x)}$$

$$L_{class}^{G} = E[|l(G(z)) - C(remove(G(z)))|]_{z \sim p(z)}$$

정보 손실 함수 $L_{\infty o}^{G}$는 식별자에서 [0,1] 범위의 스칼라 값으로 출력이 요약되기 직전의 마지막 은닉 계층의 벡터들의 통계량을 이용해서 정의되는 손실함수이다. <그림 6-2>에서 노란색 은닉 계층의 벡터들의 통계량을 이용한다. 이 값은 식별자가 최종적으로 진위를 판단하기 바로 전에 사용하는 값으로써 진위에 대한 정보가 들어있다고 볼 수 있다. 그러므로 실제 데이터와 재현자료를 식별자에 넣었을 때 이 위치의 값들이 비슷한 분포를 보인다면, 재현자료가 실제 데이터에 가깝다고 말할 수 있다. 원본 샘플과 재현 샘플에 해당하는 벡터들(아래 수식에서 f_x 또는 $f_{G(z)}$)의 평균 벡터와 표준 편자 벡터들의 차이를 계산한다. 이들의 거리(L2 norm)를 줄이는 방향으로 학습한다. 이때, 임계점($\delta_{mean}, \delta_{sd}$)을 두어서 임계점 이하로 내려가면 학습이 일어나지 않도록 제어한다. 이것을 통해 재현성능을 조절할 수 있다. 질 좋은(정보손실이 적은) 데이터를 생성하고 싶다면 임계점을 낮추고, 그 반대의 경우라면 임계점을 높이면 된다.

$$L_{mean} = \| E[f_x]_{x \sim p_{data}(x)} - E[f_{G(z)}]_{z \sim p(z)} \|_2$$

$$L_{sd} = \| SD[f_x]_{x \sim p_{data}(x)} - SD[f_{G(z)}]_{z \sim p(z)} \|_2$$

$$L_{\infty o}^{G} = \max(0, L_{mean} - \sigma_{mean}) + \max(0, L_{sd} - \sigma_{sd})$$

3. 학습 방법

기본적으로는 일반적인 적대적 생성망 학습 방식과 유사하다. 식별자와 생성자가 번갈아 가면서 학습된다. 다만, 기존 적대적 생성망과 달리 TGAN에 추가된 것이 있기 때문에 다른 양상을 보인다. L_{orig}를 이용해 식별자를 학습한 뒤, L_{class}^{C}를 이용해 구분자를 학습한다. 이후 정보 손실 함수의 계산을 위한 통계량을 계산한다. 이러한 통계량은 모든 데이터를 이용해 계산해야 하지만 메모리 문제로, 전체 데이터를 이용할 수 없다. 그래서 기존 통계량의(f_{mean}^{X}, f_{sd}^{X}, f_{mean}^{Z}, f_{sd}^{Z}) 정보를 새로 계산된 정보를 이용해 조금씩 업데이트하는 이동 평균(Moving average) 방식으로 계산한다. 이때 업데이트의 양은 w 변수로 조절할 수 있으며, 해당 연구에서는 0.99로 설정하고 학습했다. 최종적으로 L_{orig}, L_{class}^{G}, $L_{\infty o}^{G}$를 모두 이용해 생성자를 학습한다.

Input: real records: $\{x_1, x_2, \cdots\} \sim p(x)$

Ouput: a generative model G

G ← a generative neural network

D ← a discriminator neural network

C ← a classifier neural network

/* Initializing to zero vectors */

$f_{mean}^{X} \leftarrow 0; f_{mean}^{X} \leftarrow 0; f_{mean}^{Z} \leftarrow 0; f_{sd}^{Z} \leftarrow 0$

while until convergence of loss values **do**

  Create a mini-batch of real records $X_{\text{mini}} = \{x_1, \cdots, x_n\}$

  Create a mini-batch of latent vector inputs for G $Z_{\text{mini}} = \{z_1, \cdots, z_n\}$

  Perform the SGD update of the discriminator D with ι_{orig}^{D}

  Perform the SGD update of the classifier C with ι_{class}^{C}

  /* Moving average update of the mean and standard deviation of features */

  $f_{mean}^{X} = w \times f_{mean}^{X} + (1-w) \times E\left[f_x\right]_{x \in X_{\text{mini}}}$

  $f_{sd}^{X} = w \times f_{sd}^{X} + (1-w) \times SD\left[f_x\right]_{x \in X_{\text{mini}}}$

  $f_{mean}^{Z} = w \times f_{mean}^{Z} + (1-w) \times E\left[f_{G(z)}\right]_{z \in z_{\text{mini}}}$

  $f_{sd}^{Z} = w \times f_{sd}^{Z} + (1-w) \times SD\left[f_{G(z)}\right]_{z \in z_{\text{mini}}}$

  Perform the SGD update of the generator G with $\iota_{\text{orig}}^{C} + \iota_{\infty o}^{G} + \iota_{class}^{G}$

end

return G

<그림 6-3> Table GAN 학습방법

4. TGAN 성능 비교 평가 방법

TGAN의 성능 평가는 두 가지 방면으로 이루어진다. 첫 번째는 데이터의 재현 성능에 관한 것이고 다른 것은 프라이버시 보호와 관련된 것이다. 아래의 결과들은 2가지 방면의 성능을 고려했을 때, TGAN이 기존의 익명화나 잡음 추가 기술, 다른 생성 모델보다 우수함을 보여준다.

모델의 범용성을 보여주기 위해, 다양한 도메인의 데이터를 활용했다. 급여, 개인

기록, 건강, 미국 국내선 항공 승객운항 데이터를 활용했다. 도메인별 데이터 이름은 LACity, Adult, Health, Airline이다. LACity, Adult, Airline 데이터는 급여, 노동시간, 티켓 가격을, 즉 회귀 분석을 위한 속성들을 가지고 있다. 반면, Health는 당뇨병 여부, 즉 구분(Classification)을 위한 속성을 가지고 있다. 회귀 데이터들은 해당 속성 값의 중앙값(Median)에 대한 초과 여부 속성을 추가해 구분 작업도 수행할 수 있도록 했다.

위에서 언급했듯이 재현성능과 프라이버시 보호의 두 가지 방면에 대한 평가가 있다. 재현 성능에 대한 평가 방법도 세부적으로 두 가지 방법이 존재한다. 첫 번째는 누적확률분포를 이용하는 것이다. 속성별로 원본자료와 여러 기술을 통해 재현된 데이터의 누적확률분포를 구해 그 유사성을 보는 것이다. 다른 하나는 기계학습 모델을 사용하는 것이다. 동일한 테스트 데이터에 대해서 원본자료로 학습된 기계학습 모델의 점수와 재현자료로 학습된 기계학습 모델의 점수가 비슷한지 보는 것이다. 즉, 두 가지 방식으로 학습된 모델이 서로 호환 가능한지(Model compatibility) 살펴본다. 분류 작업에서는 F-1을, 회귀 작업에서는 MRE(Mean relative error)를 평가 지표로 사용한다. 기계 학습 모델 점수의 유사성이 중요한 것이지 높은 점수가 중요한 것이 아니기 때문에, 기계학습 모델에 대한 최적 하이퍼 파라미터 선정은 하지 않으며, 다양한 하이퍼 파라미터를 이용하여 해당 비교 실험을 여러 번 수행한다.

프라이버시 보호에 대한 성능 평가를 위해서는 원본 레코드와의 거리를 계산한다. 원본 레코드와 각 재현 레코드 간 거리는 가장 가까운 원본 레코드를 하나 선정하여 계산한다. 예를 들어, 1, 2, 3의 원본 레코드가 존재하고 A, B, C 재현 레코드를 생성했다고 가정해보자. A와 1, 2, 3 사이의 거리를 측정했을 때, 2와의 거리가 0.3으로 가장 짧다면 A의 원본자료와의 거리는 0.3이다. 해당 연구에서는 각 속성을 [0, 1] 범위로 정규화 후 유클리디안 거리를 이용한다. B, C와 원본자료와의 거리 역시 이러한 방식으로 측정된다. 그리고 이 거리의 평균과 표준편차를 이용해 프라이버시 보호 수준을 측정한다. 이러한 방법을 사용하는 이유는 기존의 방법과 다른 적대적 생성망의 특성 때문이다. 적대적 생성망 기반의 재현 방법은 원본자료와 재현자료가 1:1 대응되지 않는다. 따라서 재현 레코드와 가장 유사한 원본 레코드를 찾아서 거리를 측정한다.

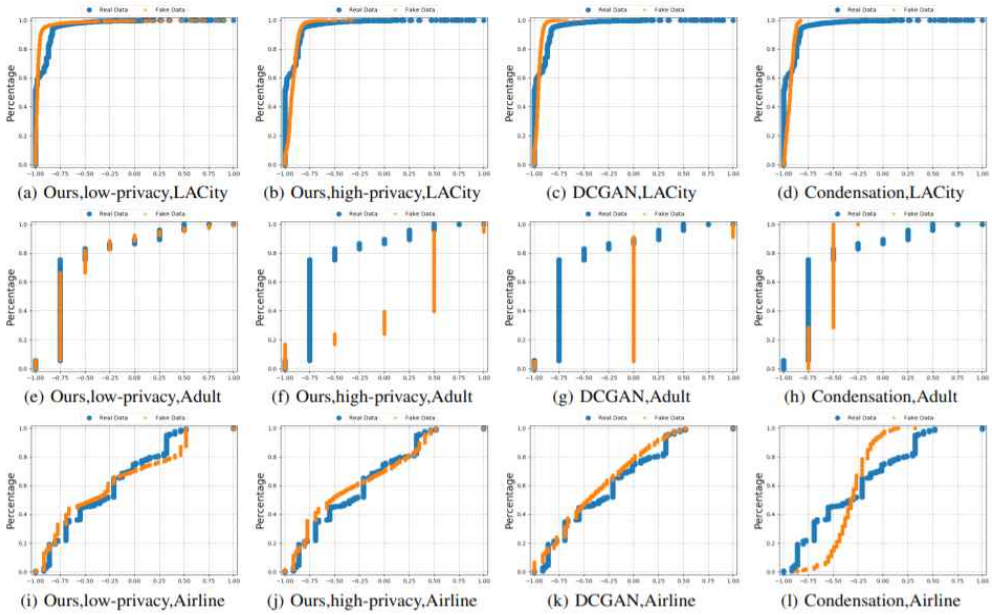
5. TGAN 성능 평가를 위한 비교 모델

두 가지 세팅의 TGAN 모델을 사용했다. 하나는 저수준 프라이버시(Low-privacy) 모델($\delta_{mean} = 0$, $\delta_{sd} = 0$)이고 다른 하나는 고수준 프라이버시(High-privacy) 모델($\delta_{mean} = 0.2$, $\delta_{sd} = 0.2$)이다. 두 모델 모두 δ_{mean}와 δ_{sd}를 조절하여 획득될 수 있다. 이와 비교되는 모델로 전통적인 익명화 및 민감정보 변형(Perturbation) 방법, 다른 생성 모델을 사용했다. ARX에 의해서는 k-익명성과 t-근접성 조합, 차등정보보호와

δ-disclosure 조합 등의 익명화 방법을 사용했다. 이러한 방법들은 준식별자만 변경하고 민감 정보는 바꾸지 않는다. sdcMicro로는 준식별자에 대해서는 마이크로 집계(Microaggregation)를 적용하고 민감정보에 대해서는 랜덤화(Post randomization)를 적용했다. 익명화 및 변형 방법에 대해서는 다양한 하이퍼 파라미터를 적용한 후, 프라이버시 보호, 재현 성능 측면에서 가장 좋은 균형을 보이는 것을 비교 대상으로 선정하였다. 비교를 위하여 사용된 다른 생성모델로는 응결 방법과 DCGAN이 있다.

6. TGAN 성능 평가 결과 – 누적 확률 분포

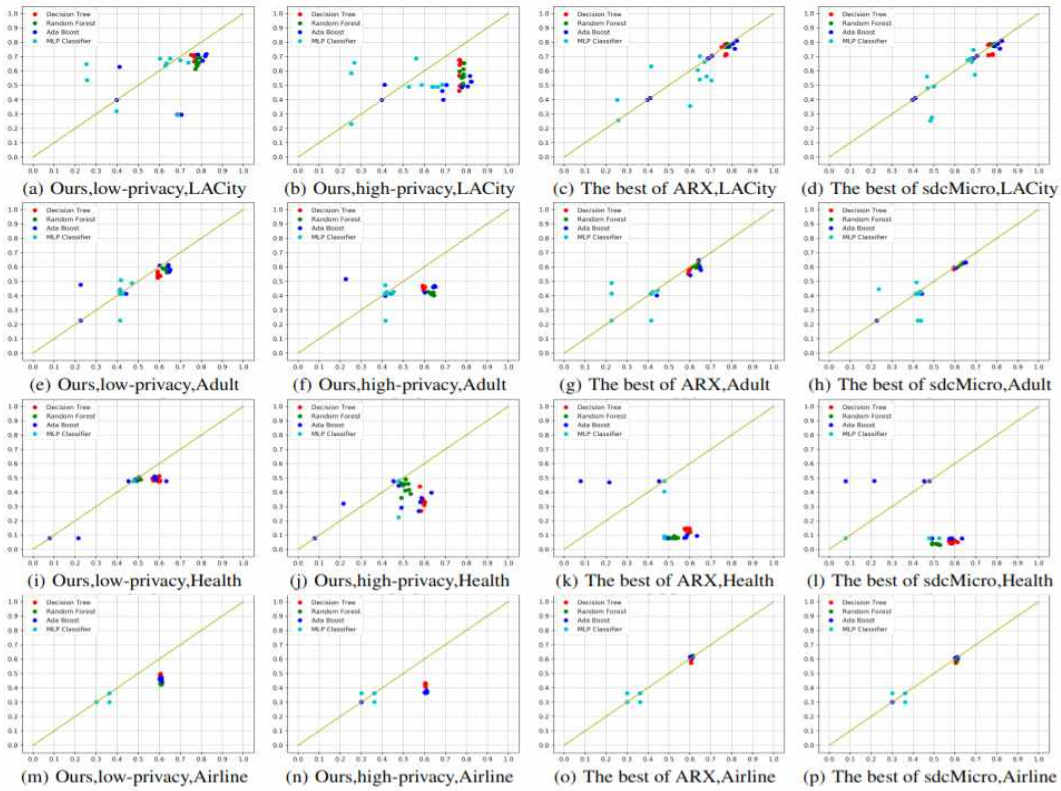
누적확률분포 측면에서 TGAN은 우수한 재현성능을 보인다. 아래 <그림 6-4>는 LACity의 base salary, Adult의 work class, Airline의 destination airport ID 속성의 누적확률분포를 나타낸 것이다. ARX, sdcMicro는 민감정보를 완전히, 거의 바꾸지 않기 때문에 나타내지 않았다. (a~d), (e~h)에서는 저수준 프라이버시 TGAN이 다른 방법보다 높은 재현 성능을 갖는다는 것을 보여준다. (i~l)에서는 DCGAN과 TGAN 모두 모든 값의 범위를 성공적으로 재현한 것을 보여준다. 다른 속성에 대해서도 비슷한 양상을 보인다. 이 결과를 통해서 저수준 프라이버시 TGAN이 가장 우수한 성능을 가지고 고수준 프라이버시 TGAN이 그 다음으로 좋은 성능을 가진다는 것을 알 수 있다. DCGAN은 테이블 데이터를 위해 만들어진 모델이 아니기 때문에 3번째 정도의 성능을 보여주고, 응결 방법이 가장 안 좋은 성능을 보여준다.



<그림 6-4> 원본자료와 재현자료의 특정 속성에서의 누적확률분포

7. TGAN 성능 평가 결과 – Model Compatibility

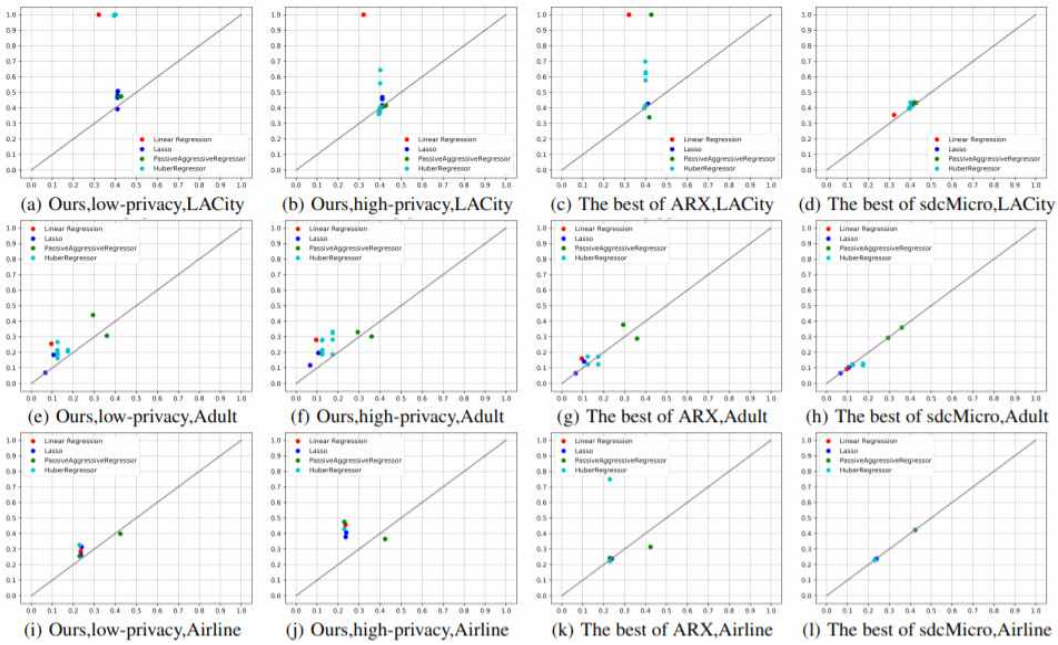
재현자료로 학습한 모델이 원래 데이터로 학습한 모델과 점수가 비슷한지(호환이 가능한지) 확인한다. <그림 6-5>와 <그림 6-6>은 각각 분류, 회귀 모델에 대한 점수 비교를 나타내는 그림이다. X축은 원래 데이터의 F-1(또는 MRE) 점수를 Y축은 재현 자료에 대한 F-1(MRE) 점수를 나타낸다. 대각선에 가까울수록 두 개의 기계학습 모델이 비슷하게 학습되었다는 것이므로 재현 성능이 우수함을 나타낸다. 분류 실험의 경우 10개 종류의 하이퍼 파라미터를 갖는 Decision Tree, Random Forest, AdaBoost, Multi-layer Perception의 네 가지 종류의 총 40개의 모델을 사용했다. 회귀 실험의 경우 10개 종류의 하이퍼 파라미터를 갖는 Linear Regression, Lasso Regression, Passive Aggressive Regression, Hurber Regression 등 총 40개의 모델을 활용했다. DCGAN과 응결 방법은 성능이 좋지 않아서 표시하지 않았다.



<그림 6-5> 분류 model compatibility

<그림 6-5> (a~d)에서 ARX, sdMicro가 가장 좋은 성능을 보이고 저수준 프라이버시 TGAN은 뒤따르는 성능을 보인다. (e~h), (m~p)에서 역시 ARX, sdMicro이 가장 우

수한 성능을, TGAN 모델이 그보다 조금 나쁜 성능을 보여준다. (i~l) Health 데이터에 대해서는 TGAN 모델이 가장 좋은 성능을 갖는다. <그림 6-6>에서는 sdcMicro가 가장 좋은 성능을 가지고 TGAN, ARX가 뒤따른다. 기존의 방법들이 좋은 성능을 보이는 이유는 민감정보를 변형하지 않거나 거의 손실 없이 변형한다는 의미이다. 하지만 이는 정보유출로 이어질 수 있는 위험성이 있다.



<그림 6-6> 회귀 model compatibility

8. TGAN 성능 평가 결과 – 프라이버시 보호

<그림 6-7>은 원본자료와 재현자료 레코드 간 거리의 평균과 표준편차를 이용해서 프라이버시 보호 수준을 나타낸 것이다. 표에서 상위 네 개의 열은 준식별자와 민감정보를 고려한 것이고, 아래 네 개의 열은 민감정보만 고려한 것이다. 그 수치가 높을수록 프라이버시 보호 정도가 높다고 할 수 있다. ARX는 민감 정보에 대해서는 아무런 변화를 주지 않기 때문에 민감 정보만 고려했을 때, 거리가 0이 나온다. 저수준 TGAN이 ARX, sdcMicro보다 더 큰 거리를 갖는 것을 볼 수 있다. 큰 거리를 갖는 점, 재현자료와 원본자료 간의 1:1 대응이 되지 않는 점에서 TGAN이 프라이버시 보호에 있어서 우수한 성능을 보여주는 것을 알 수 있다. 또한 고수준 프라이버시 TGAN이 저수준 프라이버시 TGAN보다 큰 거리를 갖는 것을 통해, 재현 성능과 프라이버스 수준에 대한 제어가 가능한 것을 확인할 수 있다.

<표 6-1> 재현 데이터와 원본 데이터 사이의 거리

	table-Gan (low-privacy)	table-GAN (high-privacy)	The best of ARX	The best of sdcMicro	DCGAN
QIDs + Sensitive attributes					
LACity	0.96 ± 0.22	1.48 ± 0.3	0.68 ± 0.52	0.07 ± 0.17	0.83 ± 0.31
Adult	0.75 ± 0.19	1.84 ± 0.23	0.59 ± 0.17	0.54 ± 0.12	0.88 ± 0.24
Health	2.53 ± 0.43	2.75 ± 0.41	0.61 ± 0.25	1.23 ± 0.34	2.85 ± 0.42
Airline	1.21 ± 0.21	1.23 ± 0.27	1.46 ± 0.32	0.98 ± 0.41	0.86 ± 0.15
Only Sensitive attributes					
LACity	0.68 ± 0.18	1.24 ± 0.17	0 ± 0	0.05 ± 0.13	0.54 ± 0.18
Adult	0.45 ± 0.14	1.25 ± 0.17	0 ± 0	0.2 ± 0.1	0.82 ± 0.24
Health	2.4 ± 0.38	2.56 ± 0.39	0 ± 0	0.22 ± 0.2	2.68 ± 0.41
Airline	0.96 ± 0.19	1.08 ± 0.26	0 ± 0	0.69 ± 0.36	0.76 ± 0.16

프라이버시 보호와 모델 호환성은 종합적으로 고려되어야 한다. TGAN이 연구되었던 시점을 기준으로 두 가지 모두를 만족하는 기술은 존재하지 않는다. 하지만 둘 사이에서의 협의점(Trade-off)을 찾을 수 하나의 아이디어가 제공되면서 후속 연구들이 현재 활발히 진행 중에 있다.

제2절 조건부 적대적 생성망을 활용한 재현자료 생성

TGAN은 초창기 기술로 실제적인 응용 측면에서 몇 가지 문제점을 가지고 있다. 그 후로도 재현자료 생성에 대한 연구는 많았으나, 범용적으로 데이터 유용성과 프라이버시 보호를 모두 제공하는 기술은 아직 좀 더 연구가 필요하다. 재현자료 생성과 관련된 기술적 난제를 정리하면 아래와 같다.

- 많은 원본자료에서 다양한 형태의 속성들이 섞여 있다. 수치 속성도 연속이나 불연속 형태가 있으며, 불연속 형태도 순서가 정의되는 것과 그렇지 않은 것으로 나뉜다. 추가로 문자열까지 섞여 있으면 재현자료 생성의 난이도는 폭발적으로 높아질 수 있다.
- 많은 경우에 속성값은 정형화할 수 없는 분포를 따르는 경우가 많다. 이미지의 경우 픽셀값의 분포가 가우시안 분포를 따르는 경우가 많으므로, 이는 딥러닝이

이미지를 잘 처리할 수 있는 이유 중의 하나이다. 그러나 그 외의 많은 경우에 테이블 데이터는 정형화되지 않는 분포를 따른다.

- 한 속성에서도 다중 모드(multi-modal)를 발견할 수 있는 경우가 많다. 다중 모드 데이터는 딥러닝이 학습하기 어려운 대표적인 문제이다. 예를 들어 이미지를 생성하는 적대적 신경망을 학습할 때 다양한 모드의 이미지들이 섞여 있으면 그 학습의 난이도가 단일 모드 이미지들로만 학습할 때보다 훨씬 높다.
- 어떠한 경우에서는 특정 값이 희소한 특성을 보인다. 각 속성에 다양한 값들이 존재할 수 있지만, 재현자료를 생성할 때는 많은 경우 희소한 값들은 무시되고 전혀 재현되지 않는 문제점이 발생할 수 있다.

위의 문제점들은 딥러닝 기반의 재현자료 생성 기술에서 더 심각하게 발현되는 특징이 있다. Conditional Tabular GAN(Xu et al., 2019)은 이러한 문제점들을 해결하려는 거의 최초의 적대적 신경망 기반의 재현자료 생성 기술이다.

조건부 적대적 테이블 생성망(Conditional Tabular GAN, 이하 CTGAN)은 기온과 같은 연속적 혹은 나이와 같은 순서가 있는 불연속적인 수치형 데이터와 성별과 같은 두 개 이상의 종류를 가지고 있는 범주형 데이터로 구성된 테이블 데이터를 재현할 수 있는 딥러닝 모델이다. CTGAN은 테이블 데이터 재현에서의 기술적 난제를 해결하기 위해 몇 가지 묘안을 사용한다. 정형화되지 않은 다중 모드(multi-modal) 분포를 학습하기 위해 모드별 정규화(Mode-specific normalization) 전처리 방법을 활용하고, 범주형 데이터의 특정 값이 희소한 문제(Imbalanced class problem)를 해결하기 위해 조건부 적대적 생성망(Conditional GAN)을 활용한다. 또한 신경망을 구성할 때도, 테이블 데이터가 잘 학습될 수 있는 구조를 사용한다.

1. 모드별 정규화 (Mode-specific Normalization)

수치형 데이터의 정형화되지 않은 다중 모드 문제를 해결하기 위해 모드별 정규화를 사용해 전처리를 진행한다. 이전 모델에서는 최댓값을 1, 최솟값을 −1로 변환하는 최소-최대 정규화(Min-max normalization)를 사용했다. 하지만 이렇게 변환하게 되면 수치형 데이터가 가우시안 분포가 아닌 다른 분포 형태를 가질 때, 역전파 과정에서 그래디언트가 사라져서 정상적으로 모델이 훈련되지 않는 문제(Gradient vanishing problem)가 발생할 수 있다. 또한, 적대적 생성망이 데이터의 다중 모드를 잘 발견하지 못하는 문제가 발생할 수도 있다. 모드별 정규화는 이러한 약점을 해결하기 위해, 전처리 과정에서 일반적인 형태의 분포를 혼합 가우시안 형태로 나타내고 특정 레코드의 모드를 명시적으로 나타내도록 전처리한다. 변분 가우시안 혼합 모델(Variational

gaussian mixture)을 사용해서 정규화를 진행하는데, 구체적인 방법은 아래와 같다.

1) <그림 6-8>의 가장 왼쪽 그림과 같이 수치형 데이터의 확률 분포를 혼합 가우시안 분포의 형태로 추정한다. 아래 식 형태로 가우시안 혼합 분포를 추정한다. 정해진 k 개의 혼합분포를 찾은 다음 가중치(μ_k)가 특정 임계치 ϵ이상인 경우의 분포만을 남긴다.

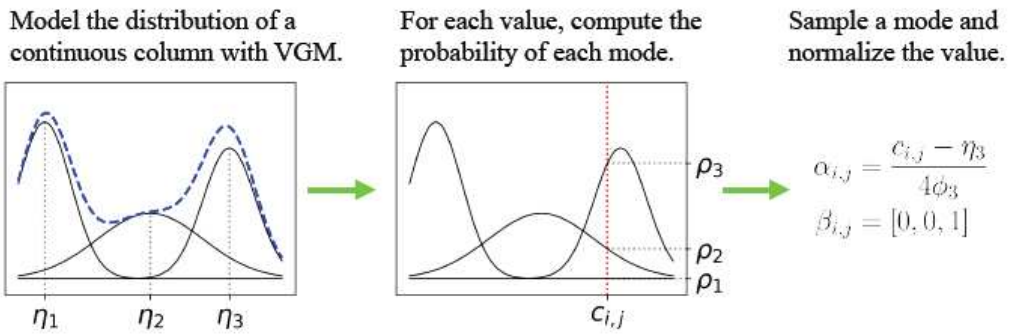
$$P(c_{i,j}) = \sum_{all\,k} \mu_k N(c_{i,j}; \eta_k, \phi_k)$$

($c_{i,j}$는 i번째 수치형 열의 j번째 레코드, μ_k, η_k, ϕ_k는 각 k번째 가우시안 분포의 가중치, 평균, 표준편차)

2) 특정 값 $c_{i,j}$에 대해서 각 가우시안 분포에서의 확률$\mu_k N(c; \eta_k, \phi_k)$을 구한다. 그리고, $P(k) = \dfrac{\mu_k N(c_{i,j}; \eta_k, \phi_k)}{\sum_{all\,p} \mu_p N(c_{i,j}; \eta_p, \phi_p)}$ 확률을 이용해 $c_{i,j}$가 어떤 가우시안 분포에서 나왔는지 선택한다. <그림 6-8>에서 ρ_1, ρ_2, ρ_3 3개의 확률을 이용해 하나의 가우시안 분포를 선택한다. ρ_1, ρ_2, ρ_3 = [0.1, 0.3, 0.1]이라고 하면, 1번 가우시안 분포가 선택된 확률은 20%(0.1/0.5)이고, 2번은 60%(0.3/0.5), 3번은 20%(0.1/0.5)이다.

3) 위의 단계에서 선택한 가우시안 분포의 평균과 표준편차를 이용해 값 $c_{i,j}$를 $\alpha_{i,j}$로 정규화하고 어떤 정규분포가 선택되었는지 원핫 벡터(One-hot vector) $\beta_{i,j}$로 표시한다. 예를 들어, 위의 단계에서 3번 가우시안 분포가 선택되었다면 $\alpha_{i,j}$, $\beta_{i,j}$는 아래와 같이 표현된다.

$$\alpha_{i,j} = \frac{c_{i,j} - \eta_3}{4\phi_3}, \quad \beta_{i,j} = [0, 0, 1]$$



<그림 6-8> 모드별 정규화의 예시

테이블에서 j번째 레코드(r_j)는 수치형 값과(c_{ij})와 범주형 값인 원핫 벡터(d_{ij})로 구성된 형태에서 모드별 정규화를 이용해 변환된 수치형 데이터(α_{ij}, β_{ij})와 범주형 데이터의 원핫 벡터($d_{i,j}$)로 구성된 형태가 된다. 이러한 방법으로 전처리된 테이블은 각 레코드에 대해서 상세한 모드 정보를 가지게 된다.

$$r_i = \alpha_{1,j} \oplus \beta_{1,j} \oplus \ldots \oplus \alpha_{N_c,j} \oplus \beta_{N_c,j} \oplus d_{1,j} \oplus d_{1,j} \oplus \ldots \oplus d_{N_d j}$$

($\oplus$는 결합(concatenation) 기호)

2. 조건부 적대적 생성망(Conditional GAN)

조건부 적대적 생성망 설명에 앞서 일반적인 적대적 생성망에 대해서 알아야 한다. 일반적인 적대적 생성망은 서로 적대적 관계에 있는 생성자(Generator)와 식별자(Discriminator)가 존재한다. 생성자는 표준 다변수 가우시안 분포에서 추출한 잡음 벡터를 입력 받아 데이터 샘플을 재현하게 되고, 식별자는 입력으로 받은 데이터 샘플이 원본 샘플인지 재현 샘플인지 구분한다. 조건부 적대적 생성망에서는 이러한 기본적인 적대적 생성망 구조의 생성자 입력에 특정 조건이 추가된다. 따라서 생성자는 특정 조건에 맞는 데이터를 생산해내는 작업을 학습하게 된다. 즉, 주어진 조건을 만족하는 재현자료를 얻을 수 있다. 예를 들면 조건으로 “남성”이라는 성별을 주고 생성자가 남성에 적합한 키와 몸무게를 재현하도록 하는 방식이다. 또 다른 예로써 동물 이미지 재현자료를 만드는 적대적 생성망을 생각해보자. 일반적인 적대적 생성망의 경우 어떤 잡음 벡터를 입력했을 때 어떤 동물이 생성될지 사용자가 조절할 수 없다. 하지만 조건부 적대적 생성망은 “강아지”이라는 조건을 명시하여 강아지 이미지가 재현되도록 할 수 있다.

조건부 적대적 생성망은 범주형 자료에서 특정 클래스가 희소한 문제(Imbalanced data problem)를 해결하기 위해 사용할 수 있다. 이러한 문제가 있는 경우, 일반적인 적대적 생성망은 그 희소한 값을 잘 학습하지 못해, 이후 데이터 재현 시 다수의 클래스만 재현하게 된다. 이러한 문제는 적대적 생성망이 다수의 클래스 데이터 위주로 학습되기 때문에 발생하게 된다. 조건부 적대적 생성망에서는 생성자가 학습 과정에서도 희소한 범주형 속성값에 대해서 노출되는 것을 조건으로 조절할 수 있어서 이러한 문제에 훨씬 더 잘 대응하는 측면이 있다. 희소한 클래스가 생성되도록 범주형 열에 대한 조건을 넣어주면 되기 때문이다. 조건부 적대적 생성망으로 실제 데이터를 재현하기 위해서 다음의 3가지 문제를 해결해야 한다.

1) 조건을 어떠한 형태로 바꿔 입력으로 넣을 것인가?

2) 어떻게 입력한 조건에 맞게 생성자가 데이터를 생성하게 할 것인가?

3) 조건부 적대적 생성망이 어떻게 실제 조건부 분포를 학습하게 할 것인가?

위 문제를 해결하고 나면 생성한 재현자료의 조건부 분포($P_g(row|D_{i^*}=k)$)는 실제 데이터의 조건부 분포($P(row|D_{i^*}=k)$)와 같아지고 다음의 식을 통해 실제 데이터 분포를 얻을 수 있다.

$$P(row)=\sum_{k\in D_{i^*}} P_g(row|D_{i^*}=k)P(D_{i^*}=k)$$

CTGAN에서는 조건 벡터(Conditional vector), 생성자의 추가적인 손실함수, 샘플링 훈련 전략(Training-by-sampling method)을 통해서 각각의 문제를 해결하고 있다.

가. 조건 벡터(Conditional Vector)

CTGAN은 범주형 속성값을 조건으로 사용한다. 예를 들어 테이블에 “성별”과 “직업”이라는 두 개의 범주형 데이터 열이 있을 경우, 특정 성별이나 특정 직업 하나를 조건으로 주게 된다. 예를 들어 조건을 성별 데이터 열의 남성이라고 하면 생성자는 그에 맞는 재현 샘플을 만들어야 하며, 조건은 직업 데이터 열의 교수라고 하면 생성자는 또 그에 맞는 샘플을 재현하도록 훈련된다. 범주형 데이터 열에 대한 조건을 만들기 위해 원핫 벡터(One-hot vector) 형태의 마스킹 벡터(m_i)를 이용한다. 범주형 열 중 하나의 열만 선택하고 그 열의 여러 클래스 중 하나의 클래스만 선택해서 조건으로 나타낸다. 즉, 아래의 수식과 같이 마스킹 벡터의 원소가 선택된 열의 선택된 클래스일 때 1로 표현되고 나머지는 0으로 표현된다.

$$m_i=[m_i^{(k)}],\ m_i^{(k)}=\begin{cases}1 \text{ if } i=i^* \text{and } k=k^*\\ 0\ otherwise\end{cases}(i^*, k^*\text{는 선택된 열과 클래스})$$

조건 벡터는 마스킹 벡터를 모두 결합한 형태인 $m_1\oplus m_2\oplus ...\oplus m_{N_d}$($N_d$는 범주형 데이터 열의 개수)로 표현된다. 예를 들어, 두 개의 범주형 열이 있고 첫 번째 열에는 {1, 2, 3} 세 개의 클래스가, 두 번째 열에는 {1, 2} 두 개의 클래스가 존재하는 경우를 생각해보자. 만약, 2번째 열에서 1이 선택되었다면, $m_1=[0,0,0],\ m_2=[1,0]$으로 표현되고 조건 벡터는 $[0,0,0,1,0]$으로 표현된다.

나. 생성자 손실 함수(Generator Loss Function)

조건부 적대적 생성망에서는 들어온 조건에 맞는 재현자료가 생성되어야 한다. 위

의 예시처럼 $[0, 0, 0, 1, 0]$ 조건이 들어왔다면, 첫 번째 범주형 열은 아무 클래스나 생성해도 되지만 두 번째 범주형 열은 1이라는 값이 생성되어야 한다. 즉, 생성된 두 번째 범주형 열 데이터($\hat{d}_2$)가 조건(m_2)과 같아야 한다. 생성자가 조건과 같은 클래스를 만들도록 학습시키기 위해, 생성자 손실 함수에 $\hat{d}_2$와 m_2의 크로스 엔트로피 손실을 추가한다.

$$Cross\,Entropy = \sum_{k=1}^{|D_i|} m_i^{(k)} \log \hat{d}_i^{(k)} \quad (|D_i|\text{는 i번째 범주형 열의 클래스 개수})$$

다. 샘플링 훈련 전략(Training-by-sampling)

생성자가 만든 재현자료의 조건부 분포와 실제 데이터 조건부 분포를 같게 만들기 위해서는 식별자에서 두 분포 간의 거리(차이)를 정확하게 추정해야 한다. 조건 벡터와 훈련 데이터를 샘플링하는 것이 거리 추정에 도움을 줄 수 있다. 식별자가 훈련될 때, 모든 가능한 경우의 범주형 열의 값들에 노출되어야 제대로 차이를 추정할 수 있다. 즉, 모든 사용 가능한 조건 벡터와 훈련 데이터를 사용해야만 제대로 된 학습이 일어날 수 있다. 균일하게 조건 벡터와 훈련 데이터를 표본을 뽑는 전략을 통해 위 문제를 해결할 수 있다. 구체적인 방법은 아래와 같다.

1) 범주형 열 중 하나를 균등한 확률로 선택한다. 예를 들어, <그림 6-9>에서 두개의 열(D_1, D_2)을 각 1/2 확률로 뽑았고, D_2가 선택되었다.

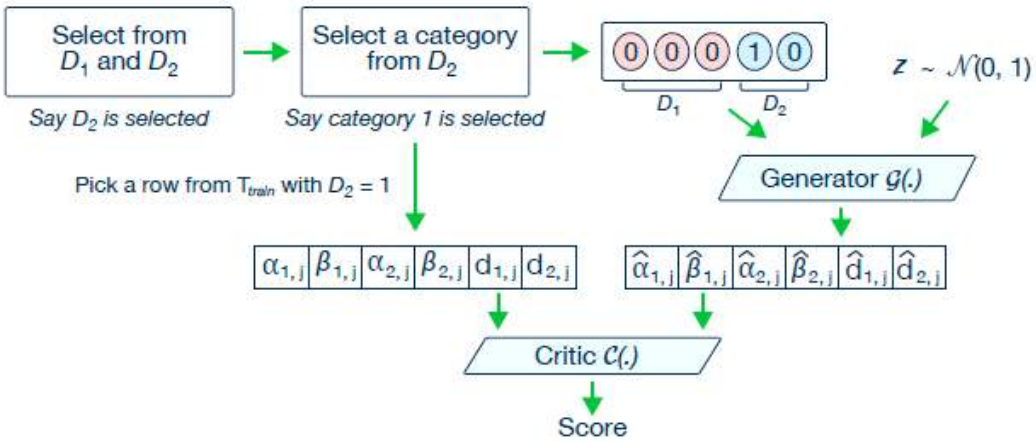
2) 선택된 범주형 열에서 각 값의 발생 빈도를 통해 확률분포를 만든다. 이때 각 빈도에 로그 함수를 취한다. 예를 들어, D_2에서 첫 번째 클래스가 100번 발생하고, 두 번째 클래스가 50번 발생했다면, 확률분포는 $(\frac{\log^{100}}{\log^{100}+\log^{50}},\ \frac{\log^{50}}{\log^{100}+\log^{50}}) \approx$ (0.54, 0.46)이다.

3) 2번에서 구한 확률분포에 따라 1개의 값을 선택한다. <그림 6-9>에서는 첫 번째 클래스가 선택되었다.

4) 선택된 열과 클래스에 따라 조건 벡터를 만들고, 훈련 데이터를 랜덤 추출한다. <그림 6-9>에서 범주형 열 D_1(클래스 3개), D_2(클래스 2개) 중에서 D_2의 첫 번째 클래스가 선택되었으므로 $m_1 = [0, 0, 0]$, $m_2 = [1, 0]$, 조건 벡터는 [0, 0, 0, 1, 0]이다. 생성된 조건 벡터와 추출한 데이터로 훈련을 진행한다.

5) 생성자에 조건 벡터와 다변수 가우시안 분포에서 추출한 잠재변수를 입력변수로 넣어서 재현자료를 생성한다.

6) 식별자(<그림 6-9> Critic)에 실제 데이터와 조건 벡터, 재현자료와 조건 벡터를 각각 넣어서 나온 결과로 두 조건부 분포의 거리(Score)를 계산해, 식별자와 생성자를 업데이트한다. 식별자는 거리를 멀어지게 하는 방향으로, 생성자는 거리를 가까워지게 하는 방향으로 학습된다.



<그림 6-9> CTGAN, 샘플링 훈련 전략의 예시(Lei Xu et al., 2019)

라. 신경망 구조(Neural Network Structure)

CTGAN은 테이블 데이터가 잘 학습될 수 있는 신경망 구조로 되어 있다. 생성자, 식별자 모두에 완전 연결 계층을 사용해, 입력값들의 모든 상관관계가 고려되도록 한다. <그림 6-10>은 생성자의 신경망 구조를 나타내고 있다. 먼저 다변수 가우시안 분포에서 만든 잠재 변수와 조건 벡터를 결합해 첫 번째 은닉(hidden) 벡터(h_0)를 만든다. 완전 연결 계층(FC), 배치 정규화 층(BN)과, ReLU함수에 h_0를 통과시킨 다음 h_0와 결합해 h_1을 만든다. 비슷한 방법으로 h_2을 만든다. 최종적으로 만들어진 은닉 벡터에 변수별로 다른 완전 연결 계층과 활성화 함수를 연결해, 여러 유형이 혼합된 테이블 데이터를 생산할 수 있게 한다. $\hat{\alpha}_i$를 만들 때는 tanh 함수를 사용했고, $\hat{\beta}_i, \hat{d}_i$와 같은 원핫 벡터로 이루어진 데이터를 만들 때는 gumbel-softmax 함수를 사용했다.

$$
\begin{cases}
h_0 = z \oplus cond \\
h_1 = h_0 \oplus ReLU\big(BN\big(FC_{|cond|+|z|\to 256}(h_0)\big)\big) \\
h_2 = h_1 \oplus ReLU\big(BN\big(FC_{|cond|+|z|+256\to 256}(h_1)\big)\big) \\
\widehat{\alpha}_i = tanh\ \big(FC_{|cond|+|z|+512\to 1}(h_2)\big) & 1 \le i \le N_c \\
\widehat{\beta}_i = gumbel_{0.2}\big(FC_{|cond|+|z|+512\to m_i}(h_2)\big) & 1 \le i \le N_c \\
\widehat{d}_i = gumbel_{0.2}\big(FC_{|cond|+|z|+512\to |D_i|}
\end{cases}
$$

<그림 6-10> CTGAN 생성자 신경망 구성(Lei Xu et al., 2019)

식별자 신경망 구성은 PacGAN으로 구성된다. 한 개의 행(조건 벡터 + 데이터)만 input으로 들어가는 것이 아니라, 10개의 행이 한꺼번에 신경망으로 들어간다. PacGAN을 통해 소수의 모드만 학습되는 문제(mode collapse)를 방지할 수 있다. 10개의 행이 합쳐져 완전 연결 계층, leaky ReLU 층을 두 번 통과하고 마지막으로 완전 연결 계층을 통과해 하나의 값을 배출한다. 과적합을 방지하기 위해 중간에 drop 층도 들어간다.

$$
\begin{cases}
h_0 = r_1 \oplus ... \oplus r_{10} \oplus cond_1 \oplus ... \oplus cond_{10} \\
h_1 = drop\big(leaky\ _{0.2}\big(FC_{10|r|+10|cond|\to 256}(h_0)\big)\big) \\
h_2 = drop\big(leaky\ _{0.2}\big(FC_{256\to 256}(h_1)\big)\big) \\
C(\bullet) = FC_{256\to 1}(h_2)
\end{cases}
$$

<그림 6-11> CTGAN 식별자 신경망 구성(Lei Xu et al., 2019)

모델을 학습하는 손실 함수로는 WGAN-GP 손실 함수를 활용했으며, Adam 최적기(learning rate = 2×10^{-4})를 사용했다. 구체적인 모델 코드는 아래 주소[17]에 있다.

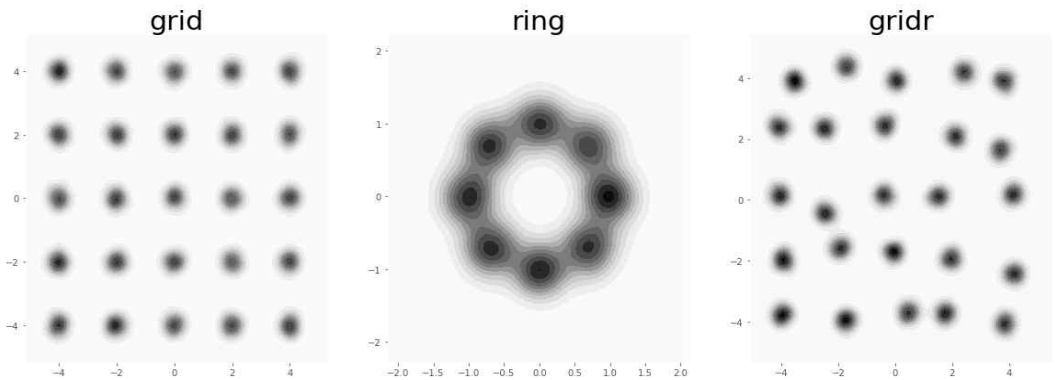
17) CTGAN 코드: https://github.com/DAI-Lab/CTGAN

3. CTGAN 성능 비교 평가 방법

CTGAN 이전에 테이블 데이터를 재현하는 많은 방법이 개발되어왔다. 베이지안 네트워크를 활용한 방법(CLBN(C Chow et al. 1968), PrivBN(Jun Zhang et al. 2017)), 적대적 생성망을 활용한 방법(MedGAN(Edward Choi et al. 2017), VeeGAN(Akash Srivastava et al. 2017) 및 TableGAN (Park et al. 2018)) 등이 있다. 공통된 평가 환경[18]을 이용하여 기존 방법과 CTGAN의 재현 데이터 생성 능력을 비교 분석한다. 또한, CTGAN에서 사용했던 여러 방법(Mode-specific normalization, Conditional GAN, 신경망 구조)들을 하나씩 제거하면서 각 방법이 얼마나 효과 있었는지 살펴본다.

다른 재현 방법과 CTGAN을 비교하기 위해서 7개의 인위적으로 만들어진 인공 데이터와 8개의 실제 존재하는 데이터를 이용한다. 모델의 재현성능은 데이터별로 다른 방법으로 측정된다. 인공 데이터의 경우 재현자료의 발생가능도(likelihood)와, 재현한 데이터로 추정한 모델에서의 테스트 데이터 발생가능도를 계산한다. 실제 데이터는 재현한 데이터로 학습시킨 기계학습 모델에 테스트 데이터를 넣어 r2, f1 점수를 확인한다.

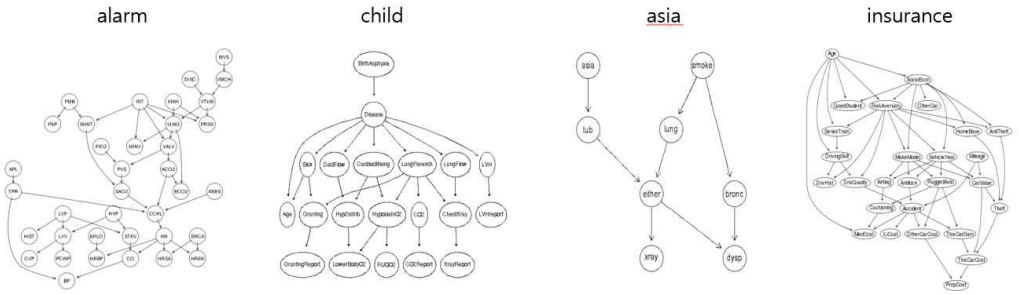
평가 자료에 관한 설명은 다음과 같다. 인공 데이터는 가우시안 혼합 모델, 베이지안 네트워크에서 생성된 데이터이다. 가우시안 혼합 모델로부터, Ring, Grid, GridR 데이터를 만들었다. 데이터는 2D 가우시안 모델이며 <그림 6-12>와 같은 형태로 구성된다. grid는 격자 형태로, ring은 고리 형태로 다중 모드가 형성되어있다. gridr은 grid에서 각 모드에 무작위로 위치변화를 준 데이터이다. 베이지안 네트워크 인공데이터는 잘 알려진 데이터 alarm, chlid, asia, insurance[19] 4개를 사용했다. <그림 6-13>은 4개의 데이터를 생성한 베이지안 네트워크의 구조이다.



<그림 6-12> 가우시안 혼합모델 인공 데이터

18) 공통된 평가 기준 코드: https://github.com/DAI-Lab/SDGym.

19) 데이터 출처: http://www.bnlearn.com/bnrepository/



<그림 6-13> 베이지안 네트워크 데이터 구조

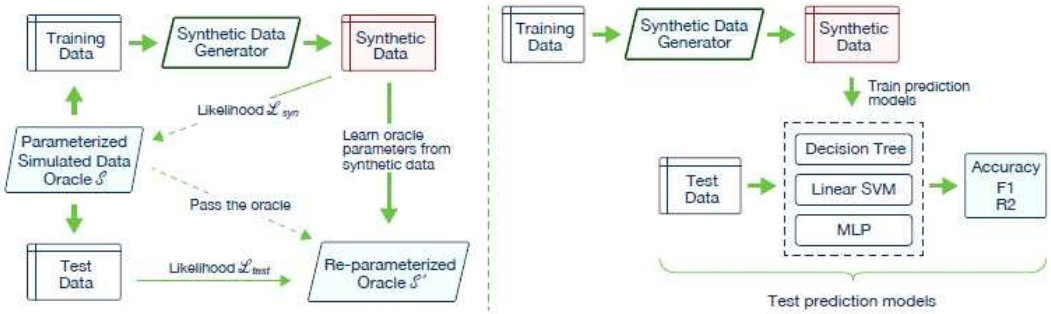
실제 데이터는 UCI machine learning repository에서 adult, census, covertype, intrusion, news를 가져왔고 Kaggle에서 credit 데이터를 가져왔다. 숫자 손글씨 이미지 데이터(MNIST)를 0, 1로 이진(binary)화해서 데이터를 만들었고, MNIST를 12픽셀로 줄여서 이진화한 데이터를 만들었다. 총 8개 데이터이며, news 데이터는 회귀(regression) 문제에 사용되고 나머지 7개는 분류(classification) 문제에 사용된다.

아래 <표 6-2>는 전체 데이터에 대한 정보를 나타낸다. #train/test는 훈련과 테스트 데이터 개수이다. #C, #B, #M은 각각 연속형 열, 클래스가 2개 존재하는 범주형 열, 클래스가 3개 이상 존재하는 범주형 열의 개수이다. C, R은 분류(classification), 회귀(regression) 문제에 사용되는 데이터임을 나타내는 것이다.

<표 6-2> 사용되는 데이터 정보(Lei Xu et al., 2019)

Simulated Data					Real Data					
name	#train/test	#C	#B	#M	name	#train/test	#C	#B	#M	task
grid	10k/10k	2	0	0	adult	23k/10k	6	2	7	C
gridr	10k/10k	2	0	0	census	200k/100k	7	3	31	C
ring	10k/10k	2	0	0	covertype	481k/100k	10	44	1	C
asia	10k/10k	0	8	0	credit	264k/20k	29	1	0	C
alarm	10k/10k	0	13	24	intrusion	394k/100k	26	5	10	C
child	10k/10k	0	8	12	mnist12	60k/10k	0	144	1	C
insurance	10k/10k	0	8	19	mnist28	60k/10k	0	784	1	C
					news	31k/8k	45	14	0	R

평가 방법은 다음과 같다. 인공 데이터의 경우에는 데이터를 만든 원본 모델(S)이 존재하기 때문에 발생가능도(likelihood)를 계산할 수 있다. 2가지 발생가능도 L_{syn}, L_{test}를 계산한다. L_{syn}는 인공 데이터 모델(S)에서 재현자료의 발생가능도이다. 발생가능도가 높을수록 원본 데이터 모델(S)에서도 생성될 가능성이 높은 데이터들이 재현된 것이므로 높은 성능을 갖는 재현 모델이라고 말할 수 있다. 하지만 이 점수는 과적합에 취약하다는 문제가 있다. 즉, 가능도가 높은 소수의 샘플들만을 재현하게 되면 재현 데이터의 전반적인 가능도가 최대화된다는 문제점이 있다. 그래서 추가적인 가능도 정의인 L_{test}를 사용한다. 모델의 큰 구조는 유지한 채 재현자료를 이용해 원래 모델과 같은 방식의 모델을 다시 처음부터 학습하여 재현 모델(S')을 생성한다. 가우시안 혼합 모델의 경우 가우시안 분포의 혼합 개수는 유지한 채, 평균과 공분산을 추정한다. 베이지안 네트워크에서는 전체적인 그래프 구조는 유지한 채 각 방향성 연결선의 조건부 분포를 추정한다. L_{test}는 테스트 샘플의 재현 모델(S')에 대한 가능도를 의미한다. 재현 모델(S')를 다시 학습하기 위해 재현자료에는 없는 사전 정보(가우시안 분포 혼합 개수, 그래프 구조)가 필요하다는 단점이 있지만, 과적합에 취약한 L_{syn}문제를 극복할 수 있고 다중모드가 하나로 통합되는 문제(mode collapse)를 발견할 수 있다. 과적합 등의 문제가 발생하면 L_{syn}는 우수한 값을 가지나 L_{test}는 아주 낮은 값을 가지게 된다. 따라서 두 가지 가능도 모두에서 우수한 값을 보이는 재현 데이터 생성 방법이 우수한 방법이다.



<그림 6-14> 인공 데이터와 실제 데이터의 모델 평가 방법(Lei Xu et al., 2019)

실제 데이터의 평가 방법은 발생가능도를 계산할 수 없어서 기계학습 모델을 이용한다. 재현자료로 기계학습 모델을 학습시키고 테스트 데이터의 예측 점수를 계산한다. 회귀 문제의 경우 r2 점수를 계산하고, 분류 문제의 경우에는 accuracy, f1 점수를 계산한다. 각각의 데이터에 대해 적절한 점수를 갖는 다른 하이퍼 파라미터를 가진 머신러닝 모델 여러 개를 사용했다. 모델 여러 개의 점수의 평균을 계산해 생성자의 점수로 계산했다. 아래 <표 6-3>은 각 실제 데이터에 사용한 모델과 재현자료·

가 아닌 실제 훈련 데이터를 이용했을 때, 테스트 점수이다.

<표 6-3> 각 실제 데이터별 기계학습 모델과 테스트 점수(Lei Xu et al., 2019)

dataset	name	accuracy	f1	macro_f1	micro_f2	r2
adult	Adaboost (estimator=50)	86.07%	68.03%			
	Decision Tree (depth=20)	79.84%	65.77%			
	Logistic Regression	79.53%	66.06%			
	MLP(50)	85.06%	67.57%			
census	Adaboost (estimator=50)	95.22%	50.75%			
	Decision Tree (depth=30)	90.57%	44.97%			
	MLP(100)	94.30%	52.43%			
covtype	Decision Tree (depth=30)	85.25%		73.62%	82.25%	
	MLP(100)	70.06%		56.78%	70.06%	
credit	Adaboost (estimator=50)	99.93%	76.00%			
	Decision Tree (depth=30)	99.89%	66.67%			
	MLP(100)	99.92%	73.31%			
intrusion	Decision Tree (depth=30)	99.91%		85.82%	99.91%	
	MLP(100)	99.93%		86.65%	99.93%	
mnist12	Decision Tree (depth=30)	84.10%		83.88%	84.10%	
	Logistic Regression	87.29%		87.11%	87.29%	
	MLP(100)	94.40%		94.34%	94.40%	
mnist28	Decision Tree (depth=30)	86.08%		85.89%	86.08%	
	Logistic Regression	91.42%		91.29%	91.42%	
	MLP(100)	97.28%		97.26%	97.28%	
news	Linear Regression					0.1390
	MLP(100)					0.1492

4. CTGAN 성능 평가

성능 평가를 위해 배치사이즈 500으로 모델을 300에폭 훈련하였다. 하나의 에폭마다 데이터의 개수÷배치사이즈만큼 샘플링을 통한 훈련(Training-by-sampling)을 진행했다. Identity 모델은 단순히 훈련 데이터에서 표본추출 한 것으로 점수를 계산했다. Identity 점수가 인공데이터의 몇몇 경우를 제외하고 얻을 수 있는 최고점수이다. 굵은 글씨로 표현된 것은 가장 좋은 성능을 낸 점수이다. 각 모델 옆에 있는 숫자는 모델의 순위이며, 1등일수록 좋은 성능을 갖는다고 할 수 있다. 모델의 순위를 계산하는 방법은 데이터별 순위를 산출한 다음 순위를 모두 합해, 합한 숫자가 작은 것부터 다시 순위를 산출했다.

<표 6-4> 각 모델 성능(Lei Xu et al., 2019)

Gaussian Mix	grid		gridr		ring	
method	L_{syn}	L_{test}	L_{syn}	L_{test}	L_{syn}	L_{test}
Identity	-3.06	-3.06	-3.06	-3.07	-1.70	-1.70
CLBN(1)	**-3.68**	-8.62	**-3.76**	-11.60	**-1.75**	**-1.70**
PrivBN(3)	-4.33	-21.67	-3.98	-13.88	-1.82	-1.71
MedGAN(6)	-10.04	-62.93	-9.45	-72.00	-2.32	-45.16
VEEGAN(5)	-9.81	-4.79	-12.51	-4.94	-7.85	-2.92
TableGAN(4)	-8.70	-4.99	-9.64	-4.70	-6.38	-2.66
CTGAN(2)	-5.63	**-3.69**	-8.11	**-4.31**	-3.43	-2.19

Bayesian Net	asia		alarm		child		insurance	
method	L_{syn}	L_{test}	L_{syn}	L_{test}	L_{syn}	L_{test}	L_{syn}	L_{test}
Identity	-2.23	-2.24	-10.3	-10.3	-12.0	-12.0	-12.8	-12.9
CLBN(2)	-2.44	-2.27	-12.4	-11.2	-12.6	-12.3	-15.2	-13.9
PrivBN(1)	**-2.28**	**-2.24**	-11.9	**-10.9**	**-12.3**	**-12.2**	**-14.7**	**-13.6**
MedGAN(4)	-2.81	-2.59	**-10.9**	-14.2	-14.2	-15.4	-16.4	-16.4
VEEGAN(6)	-8.11	-4.63	-17.7	-14.9	-17.6	-17.8	-18.2	-18.1
TableGAN(4)	-3.64	-2.77	-12.7	-11.5	-15.0	-13.3	-16.0	-14.3
CTGAN(3)	-2.56	-2.31	-14.2	-12.6	-13.4	-12.7	-16.5	-14.8

Real	adult	census	credit	cover.	intru.	mnist12	mnist28	news
method	F1	F1	F1	Macro-F1	Macro-F1	Acc	Acc	R^2
Identity	0.669	0.494	0.720	0.652	0.862	0.886	0.916	0.14
CLBN(2)	0.334	0.310	0.409	0.319	0.384	**0.741**	0.176	-6.28
PrivBN(3)	0.414	0.121	0.185	0.270	0.384	0.117	0.081	-4.49
MedGAN(5)	0.375	0.000	0.000	0.093	0.299	0.091	0.104	-8.80
VEEGAN(5)	0.235	0.094	0.000	0.082	0.261	0.194	0.136	-6.5e6
TableGAN(4)	0.492	0.358	0.182	0.000	0.000	0.100	0.000	-3.09
CTGAN(1)	**0.601**	**0.391**	**0.672**	**0.324**	**0.528**	0.394	**0.371**	**-0.43**

가우시안 혼합 데이터에서 CLBN과 PrivBN의 L_{test}점수가 좋지 않은 것을 볼 수 있다. 이는 두 모델이 연속형 변수를 이산형 변수로 쪼개 표현하기 때문에 발생하는 문제이다. MedGAN, VEEGAN, TableGAN은 다중모드가 하나로 통합되는 문제(mode collapse)를 피하지 못했다. CTGAN은 전체적인 점수 분포를 봤을 때, mode-specific normalization을 통해 다중 모드 수치형 데이터를 잘 재현하였다. 베이지안 네트워크 데이터에서 CLBN과 PrivBN이 베이지안 네트워크에 기반을 둔 모델이기 때문에 좋은 성능을 내고 있다. CTGAN은 MedGAN, TableGAN보다 조금 나은 성능을 보여주고 있다. 실제 데이터에서 CTGAN의 성능이 증명된다. 대부분의 실제 데이터에서 CTGAN은 우월한 성능을 보여주고 있다.

5. Mode-specific normalization, Conditional GAN, 신경망구조 성능 비교 평가

CTGAN에서 Mode-specific normalization, Conditional GAN, 신경망구조 3가지 트릭이 주요하게 사용되었다. <표 6-4>는 이들을 제거한 뒤 계산된 점수와 원래 CTGAN 점수의 비교를 통해 3가지 트릭의 효과를 보여준다.

Mode-specific normalization VGM을 이용해 모드의 개수를 유동적으로 추정했다. 모드의 개수를 유동적으로 추정하지 않고 고정된 모드 개수를 사용했을 때, 성능의 하락을 보였다. 5개(GMM5), 10개(GMM10)의 고정된 모드를 사용했을 때, 각각 4.1%, 8.6% 정도 성능이 떨어졌다. 또 Mode-specific normalization을 이용하지 않고 최댓값을 1, 최소값을 −1로 변환하는 Min-Max normalization을 사용했을 때, 25.7% 정도의 성능 하락을 보였다. 유동적으로 모드 개수를 추정해 전처리하는 것이 효과가 있음을 보여준다.

CTGAN을 훈련하기 위해서 조건 벡터를 input으로 넣어주고, 조건 벡터와 훈련 데이터를 표본 추출하면서 훈련한다. 훈련데이터를 표본 추출하지 않고 전체 데이터를 처음부터 끝까지 순서대로 input으로 넣어 훈련을 진행한 경우가 <표 6-5>에서 “w/o S.”에 해당한다. 이러한 세팅은 CTGAN이 다양한 조합의 훈련데이터에 노출되어 좀 더 안정적으로 학습되는 것을 방해하는 세팅이다. “w/o C.”로 표기된 항목은 “w/o S.”에서 조건부 벡터까지 사용하지 않은 경우이다. 위에서 설명한 것처럼 조건부 벡터는 임의로 선택된 범주형 속성값을 의미한다. 즉, 생성해야 할 범주형 속성값을 명시적으로 생성자에게 입력해주는 것이다. 이러한 조건을 명시하지 않으면 생성자가 알아서 가짜 샘플을 만들게 되며 사용자는 생성될 가짜 샘플의 범주형 속성값을 제어할 수 없게 된다. 이럴 경우 생성자는 빈도수가 낮은 범주형 속성값은 쉽게 무시하고 빈도수가 큰 값들에만 집중하는 mode-collapse 문제가 발생하기 쉽다. 각각

의 경우에서 17.8%, 36.5%의 성능 하락을 보였다. 특히 원본 데이터의 클래스별 샘플 수가 차이가 심한 불균형 데이터에 치명적이었는데, 대표적으로 credit 데이터의 f1 점수가 0이 나왔다. 조건부 적대적 생성망(Conditional GAN)이 불균형 데이터에 얼마나 효과적인지 보여준다.

신경망 구조 CTGAN에서는 WGANGP loss와 PacGAN(Pac = 10)을 사용했다. 일반적인 GAN 손실함수만 사용했을 때는 6.5% 정도 성능이 떨어졌고 WGANGP 손실함수만 사용했을 때는 오히려 성능이 조금 올라갔다. 일반 GAN 손실함수와 PacGAN을 사용했을 때는 GAN 손실함수만을 사용했을 때보다는 성능이 조금 올라간 −5.2% 성능을 보였다. WGANGP 손실함수가 데이터 재현에 있어서 더 적합하며, PacGAN은 WGANGP 손실함수보다 일반 GAN 손실함수에 효과가 있음을 보여준다.

<표 6-5> Mode-specific normalization, Conditional GAN, 신경망구조에 따른 성능 변화(Lei Xu et al., 2019)

Model Performance	Mode-specific Normalization			Generater		Network Architechture		
	GMM5	GMM10	MinMax	w/o S.	w/o C.	GAN	WGANGP	GAN + PacGAN
	-4.1%	-8.6%	-25.7%	-17.8%	-36.5%	-6.5%	+1.75%	-5.2%

제 7 장

결론

재현자료는 1993년 다중대체라는 통계적 기술을 이용하여 생성하도록 제안되었다(Rubin, 1993). 이후 통계적 모형을 구축하는 방법, 노출위험과 정보손실을 측정하는 방법 등이 꾸준히 연구되어 왔다. 국내에서는 박민정과 김항준(2016) 및 박민정과 김정연(2017)에서 재현자료를 소개하였으며, 이를 기반으로 국내에서 재현자료를 시범 구축한 사례가 두 차례 있었다. 통계청 통계데이터센터의 실제 DB에 대하여 재현자료를 시범 생성 사업을 진행하는 것을 계기로, 본 보고서에서는 재현자료 전반에 관한 이론적인 내용과 해외 사례를 소개하고 시범 생성 내용을 논의하였다. 나아가 향후 재현자료 생성을 위한 새로운 접근으로 연구되고 있는 GAN을 이용한 재현자료 생성 방식을 소개하였다.

재현자료 생성 관련 최근 연구 동향을 요약하면 각국 통계기관을 중심으로는 통계적 모형을 이용한 재현자료 생성 연구가 이루지면서, 동시에 딥러닝 기법을 이용한 재현자료 생성 방식도 제안되고 있다. 시범 생성 결과 통계적 모형을 이용하여 재현자료를 생성할 때는 실무적으로 자료의 특성을 반영하여 다양한 조건을 구현하며 작업하는 것이 중요하였다. 한편 GAN을 이용한 재현자료 생성 역시도 조건부 GAN을 이용하는 등 자료의 특징을 더 잘 반영할 수 있는 방향으로 발전하고 있다고 파악된다. 양쪽 접근 방식 모두에서 적절한 재현자료 생성을 위한 기술적인 노력을 경주하고 있으나, 분야가 달라 비교 연구가 부족한 실정이다. 한편 딥러닝 모델로 생성한 데이터의 정보보호 수준을 이론적으로 보장하기 위한 연구도 진행되고 있으나, 아직 실제 자료에 적용할 정도로 그 수준이 높지는 않다(Jordon et al., 2019). 따라서 재현자료 생성을 위하여 해결해야 할 기술적인 문제들이 산적해 있을 뿐만 아니라 재현자료의 노출위험 측정에 관한 이론적 발전도 더 필요하다고 볼 수 있다.

데이터 공유는 빅 데이터와 인공지능으로 대표되는 제4차 산업혁명의 핵심 요소이다. 하지만 프라이버시를 보호하면서 효과적으로 데이터를 공유하는 데는 현실적으로 큰 어려움이 따른다. 마땅한 대안이 없는 상황에서 데이터를 공유하기 위하여 재현자료를 활용하는 방안도 충분히 연구할 가치가 있고, 나아가 적절한 활용 방안을 모색할 필요도 있다. 연구자들의 끊임없는 노력으로 가까운 미래에 충분한 수준으로 노출위험을 제어한 유용한 재현자료 생성 모형이 만들어질 것을 기대한다.

참고문헌

박민정, 김정연(2017). 재현자료 작성 방법론 검토, **2017년 연구보고서**, 통계개발원.

박민정, 김항준(2016). 마이크로데이터 공표를 위한 통계적 노출제어 방법론 고찰, **응용통계연구, 39(6)**, pp. 1041-1059.

유성준, 박나리(2020). CART 기법을 이용한 개인신용정보 재현자료 생성, **통계연구, 25(1)**, 1-30.

Benedetto, G., Stanley, J. C., & Totty, E. (2018). The Creation and Use of the SIPP Synthetic Beta v7.0.

Che, Z., Cheng, Y., Zhai, S., Sun, Z., & Liu, Y.(2017). Boosting deep learning risk prediction with generative adversarial networks for electronic health records, In ICDM.

Choi, E., Biswal, S., Malin, B. Duke, J, Stewart, W. F., & Sun, J.(2017). Generating multi-label discrete patient records using generative adversarial networks. In MLHC.

Christiano, G. & Gina, L. (2020). Comparing the predictive power of the CART and CTREE algorithms, *Revista Avaliação Psicológica, 19(1)*, 87-96.

Drechsler, J.(2009). Synthetic Datasets for the German IAB Establishment Panel.

Drechsler, J. & J. P. Reiter(2011). An empirical evaluation of easily implemented, nonparametric methods for generating synthetic datasets, *Computational Statistics and Data Analysis 55*, 3232−3243.

Drechsler, J. & Reiter, J.(2010). Sampling with synthesis: a new approach for releasing public use census microdata, *Journal of the American Statistical Association, 105(492)*, 1347-1357.

Fan, J., Liu, T., Li, G., Chen, J., Shen, Y. & Du, X.(2020). Relational Data Synthesis using Generative Adversarial Networks: A Design Space Exploration, In VLDB.

Hothorn, T., Hornik, K., & Zeileis, A.(2006). Unbiased recursive partitioning: A conditional inference framework, *Journal of Computational and Graphical Statistics, 15(3)*, 651-674.

Hu, J., Reiter, J., & Wang, Q.(2014). Disclosure risk evaluation for fully synthetic categorical data, *Privacy in Statistical Databases 2014*, 185-199.

Jordon, J., Yoon, J., & van der Schaar M.(2019). Pate-gan: Generating synthetic data with differential privacy guarantees. In ICLR.

Karr, A. F., Kohnen, C. N., Oganian, A. Reiter, J. P., & Sanil, A. P.(2006). A framework for evaluating the utility of data altered to protect confidentiality, *The American Statistician, 60(3)*, 1-9.

Kinney, S. K., Reiter, J. P., & Miranda, J.(2014). Improving the Synthetic Longitudinal

Business Database.

Kinney, S. K., Reiter, J. P., Arnold P. Reznek, A. P., Javier Miranda, J., Ron S. Jarmin, R. S., John M., & Abowd, J. M.(2011). Towards Unrestricted Public Use Business Microdata: The Synthetic Longitudinal Business Database.

Little, R. J. A.(1993). Statistical analysis of masked data, *Journal of Official Statistics, 9(2),* 407-426.

Loong, B., Zaslavsky, A. M., He, Y., & Harrington, D. P.(2013). Disclosure control using partially synthetic data for large-scale health surveys, with applications to CanCORS, *Statistics in Medicine, 32(34),* 4139-4161.

McClure, D. & Reiter, J.(2016). Assessing disclosure risks for synthetic data with arbitrary intruder knowledge, *Statistical Journal of the IAOS, 32*, 109-126.

Nowok, B., Raab, G. M. & Dibben, C.(2016). synthpop: Bespoke creation of synthetic data in R, *Journal of Statistical Software, 74(11).*

Nowok, B., Raab, G. M., & Dibben, C.(2017). Providing bespoke synthetic data for the UK Longitudinal Studies and other sensitive data with the synthpop package for R, *Statistical Journal of the IAOS 33,* 785−796.

Ohm, P.(2010). Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization, *UCLA Law Review, Vol. 57*, p.1701.

Park, N., Mohammadi, M., Gorde, K., Jajodia, S., Park, H., & Kim, Y.(2018). Data Synthesis based on Generative Adversarial Networks, *VLDB, 11(10)*, 1071-1083.

Raab, G. M., Nowok, B. & Dibben, C.(2017). Guidelines for Producing Useful Synthetic Data. arVix:1712.04078vl [stat.AP].

Raghunathan, T.E., Lepkowski, J.M., Hoewyk, J.V., & Solenberger, P.(2001). A multivariate technique for multiply imputing missing values using a sequence of regression models, *Survey Methodology, 27(1)*, 85-95.

Reiter, J.(2005a). Estimating risks of identification disclosure in microdata, *Journal of the American Statistical Association, 100(472)*, 1103-1112.

Reiter, J.(2005b). Using CART to generate partially synthetic, public use microdata, *Journal of Official Statistics, 21*, 441-462.

Reiter, J. & Mitra, R.(2009). Estimating risks of identification disclosure in partially synthetic data, *Journal of Privacy and Confidentiality, 1(1)*, 99-110.

Rubin, D. B.(1993). Discussion : Statistical disclosure limitation, *Journal of Official Statistics, 9,* 461-468.

Snoke, J. Raab, G. M., Nowok, B., Dibben, C., & Slavkovic, A.(2018). General and Specific Utility Measures for Synthetic Data. *Journal of the Royal Statistical Society A, 181, Part3*, pp.663-688.

Strobl, C.(2014). Comments on fifty years of classification and regression trees. *International Statistical Review, 82(3)*, 349-352.

van Buuren S. & Groothuis-Oudshoorn, K.(2011). mice: Multivariate imputation by chained equations in R, *Journal of Statistical Software, 45(3)*.

Woo, M.-J., Reiter, J. P., Oganian, A., & Karr, A. F.(2009). Global measures of data

utility for microdata masked for disclosure limitation, *The Journal of Privacy and Confidentiality, 1(1)*, 111-124.

Xu, L., Skoularidou, M., Cuesta-Infante, A., & Veeramachaneni, K.(2019). Modeling Tabular Data using Conditional GAN, In NeurIPS.

Zhang, J. Cormode, G., Magdalena, C., Srivastava, D., & Xiao, X.(2014). PrivBayes: Private Data Release via Bayesian Networks, In SIGMOD 2014.

부 록

1. 적대적 생성망에 대한 기본적인 설명

적대적 생성망은 종래의 변분오코인코더 (Kingma et al., ICLR 2014)와 같은 생성 모델과는 완전히 다른 성격을 가지고 있다. 적대적 생성만은 생성자(Generator)와 식별자(Discriminator)의 두 가지 독립적인 신경망으로 구성된다. 두 신경망들이 아래의 수식으로 표현되는 제로섬 최소-최대 게임(Zero-sum min-max Game) 학습을 따른다.

$$\min_G \max_D V(G,D) = E[\log D(x)]_{x \sim p_{data}(x)} + E[\log(1 - D(G(z)))]_{z \sim p(z)}$$

위의 수식에서 G가 생성자, D가 식별자를 의미한다. x는 원본자료 분포 p_{data}를 따르는 특정 원본자료 샘플을 의미하며, z는 저차원 잡음 벡터를 의미한다. 보통 128 차원 정도의 잡음 벡터를 이용한다. $G(z)$는 생성자가 z를 입력으로 받아 생성한 하나의 재현자료 샘플을 의미한다. 즉, 생성자는 잡음 벡터들과 데이터 샘플들 간의 매핑 함수로 이해될 수 있다. 생성자가 잡음 벡터를 입력으로 받는 이유는 생성자가 생성하는 샘플의 종류를 외부에서 이해하고 조절할 수 있기 위해서 이다.

식별자는 위의 게임 모델 수식을 최대화하는 반면 생성자는 최소화한다. 식별자가 최대화하는 방법은 $D(x) = 1$, $D(G(z)) = 0$을 출력하는 것이다. 즉, 원본 샘플과 재현 샘플 간의 이진 식별(Binary classification) 작업을 수행하라는 의미이다. 생성자가 최소화하는 방법은 $D(G(z)) = 1$을 만들 수 있도록 원본과 비슷한 수준의 샘플을 만드는 것이다. 이러한 과정 중에서 적대적 생성망이 다른 생성 모델들과 차별화되는 특징 중의 하나가 생성자가 원본 샘플을 직접 접근하지 않는다는 것이다. 생성자의 입력은 잡음 벡터 z뿐이며, 생성한 샘플에 대해서 식별자로부터 피드백을 받아서 간접적으로 원본 샘플의 특징을 파악해야 한다는 것이다. 이러한 생성자와 식별자의 관계를 아래와 같이 몇 가지 방식으로 해석될 수 있다.

생성자와 식별자는 서로 적대적인 관계에 있으며, 생성자는 고수준의 재현 샘플을 생성해서 식별자를 속이고자 한다. 이러한 해석은 적대적 생성망에 대한 표준 해석이다. 예를 들어 생각하면 식별자는 범인을 목격한 목격자이고, 생성자는 몽타쥬를 그리는 경찰이다. 경찰은 목격자의 피드백을 받으면서 지속적으로 몽타쥬의 수준을 높여가는 작업을 수행하는데, 이러한 과정이 적대적 생성망과 비슷하다.

Abstract

Study on Synthetic Data Generation Methods with Applications to Statistics Korea RDC Data

Min-Jeong Park, Jungim Han, Noseong Park

Data serves as the fuel driving this era of the fourth industrial revolution. Data is also needed in order for a machine learning model to be learned. High quality data can enrich our lives by the advantageous application of high-quality analysis and a well-trained sophisticated AI model. Data sharing is needed in this sense, but often data is not actively shared due to privacy issues. As nations around the world reinforce their applicable Acts and regulations regarding privacy, there is a concern over further reductions in data sharing.

For the protection of privacy, data anonymity has traditionally been employed to keep the disclosure risks of personal information under control; furthermore, noise addition has been suggested to satisfy differential privacy. However, if such a technique is applied to actual data, it will be accompanied with problems including a decline in data utility. This is the reason that the generation of synthetic data is being studied as an alternative to control disclosure risks and secure the utility of data. The synthetic data suggested to overcome the limitations of the traditional anonymity methods refer to data generated from a distribution similar to that of the original data. Such synthetic data can be created through a statistical model, but recently eyes are also on a method that uses state-of-the-art deep learning technology.

This report explains statistical methods to generate synthetic data; methods that measures disclosure risks and the degree of information loss with synthetic data; case analysis of synthetic data created at domestic and overseas statistical organizations; and results of synthetic data created on a trial basis at the actual database of the Research (Statistics) Data Center at Statistics Korea. It further introduces research trends regarding synthetic data using deep-learning technologies.

Keywords : Privacy, anonymization, synthetic data, multiple imputation, GAN

집필진

- 박민정(통계청 사무관)
- 한정임(통계청 통계개발원 통계방법연구실 주무관)
- 박노성(연세대 인공지능학과 조교수)

연구보고서 2020-13

통계데이터센터 DB를 활용한 재현자료 생성 방법 연구

인 쇄	2021년 4월 12일
발 행	2021년 4월 13일
발 행 인	통계개발원장 전영일
발 행 처	통계청 통계개발원 35220 대전광역시 서구 한밭대로 713 TEL.(042)366-7100 FAX.(042)366-7123
홈페이지	http://sri.kostat.go.kr
ISSN(Online)	2733-4120

통계청
통계개발원