Ⅳ 재현데이터의 효용과 노출위험 측정

1. 노출위험

데이터 노출위험 종류는 다음과 같은 세 가지로 분류할 수 있다. 첫째, 신원 노출(Identity disclosure)의 위험이다. 이 위험은 데이터 침입자가 공개된 다른 데이터 레코드와 결합하는 경우 발생한다. 예를 들어 데이터 내부 특정 관측치들을 외부 정보와 연결했을때 극단적인 값을 가진 개인이 있다면 이를 식별해 낼 수 있다. 여기서 극단치는 매우 큰 숫자일 수도 있지만 매우 드문 범주가 될 수도 있다. 식별이 성공하면 침입자는 노출된 관측치의 다른 모든 정보 접근도 가능하게 된다. 둘째, 속성 노출(Attribute disclosure)의 위험이다. 속성 노출위험은 신원 노출 없이도 발생할 수 있는데, 예를 들어 어떤 병원에 입원한 56~60세의 모든 여성 환자가 암에 걸렸다는 것을 보여주는 데이터가 공개된 경우, 침입자는 특정 개인을 식별할 필요 없이 이 병원에 입원한 적이 있는 모든 56~60세의 여성 환자의 의학적 정보를 알 수 있게 된다. 즉, 개인의 특정 속성이 노출되는 것이다. 마지막으로 추론 노출(Inferential disclosure)위험인데 이는 예측 성능이 뛰어난 모형이 주어졌을 때 데이터에 기록된 속성을 모형에 대입하면 민감한 개인정보를 신뢰할 만한 수준으로 추론할 수 있게 되는 노출위험을 말한다.[47]

공개된 데이터가 안전한지를 확인하기 위해서는 이를 측정할 수 있는 기준이 있어야 한다. 이것이 위험 측정(Risk measurement)이라는 개념이다. 적절한 위험 측정이 가능하다면 데이터 관측치별 위험성을 파악하여 특정 관측치들에만 가명처리 기법을 적용할 수도 있고, 데이터 변수(테이블 데이터의 경우 칼럼)별로 통계적 노출 제어를 적용할 수도 있다. 전통적인 위험 측정 방법은 다음 세 가지이다. 먼저 k-익명성[48]은 노출 제어 이후 데이터에 각각의 익명화된 범주에 속하는 개체의 개수로 노출의 위험을 측정하는 방법이다. 각 익명화된 범주에 속하는 개체의 개수가 임계치인 k개 이상이면 노출의 위험이 낮다고 판단한다. 두 번째로 l-다양성[49]은 k-익명성 노출위험 측정 후에도 특정 동질집합에서

47) Park, M. J., and Kim, H. J.(2016)

48) Samarati and Sweeney(1998)

서로 같은 민감한 정보를 재식별이 가능하다는 단점을 보완한 방법이다. 주어진 데이터 집합에서 가명처리되는 개체들은 동질집합에서 적어도 l개의 서로 다른 민감한 정보를 가졌는지 확인하여 노출위험을 판단한다. 마지막으로 t-근접성[50]은 민감정보 분포와 전체 데이터의 민감한 정보 분포의 거리 차이가 t 이하이면 노출의 위험이 낮다고 판단하는 방법이다. 거리를 측정하는 방법으로는 〈그림 Ⅳ-1〉과 같이 Variational distance, Kullback-Leibler divergence, Earth Mover's distance 방법 등이 있다. 하지만 k-익명성, l-다양성, t-근접성의 경우 사용할 모수인 k, l, t의 값의 결정이 자의적일 수 밖에 없다는 점에 유의해야 한다.

〈그림 Ⅳ-1〉 t-closeness 거리 측정법

거리 측정 방법	식
Variational distance	$d(p,q)=\sum_{i=1}^{m}\frac{1}{2}\lvert p_i-q_i\rvert$
Kullback–Leibler divergence	$d(p,q)=\sum_{i=1}^{m}p_i\log\frac{p_i}{q_i}=H(p)-H(p,q),$ $where\ H(p)=\sum_{i=1}^{m}p_i\log p_i,\ H(p)=\sum_{i=1}^{m}p_i\log q_i$
Earth Mover's distance	$d(p,q)=Work(p,q,f)=\sum_{i=1}^{m}\sum_{j=1}^{m}d_{ij}f_{ij}$

자료: 김승현(2020)

2. 데이터의 효용과 노출위험의 상충관계

일반적으로 데이터 노출 제어 정도 혹은 정보보호의 강도가 높을수록 원데이터가 가지는 유용성 혹은 효용(Data utility)[51]이 떨어지게 된다. 이는 원데이터 정보에 손실이 생겼다는 의미로, 사용자가 의도한 데이터 기반 모형 구현을 통한 분석·예측이 실패하거나 왜곡될 가능성이 증가하게 된다. 반대로 데이터의 정보보호 수준을 낮춘다면 데이터의 효용은 높아지겠지만 반대로 노출위험은 그만큼 커진다. 이를 데이터의 효용-정보보호의 상충관

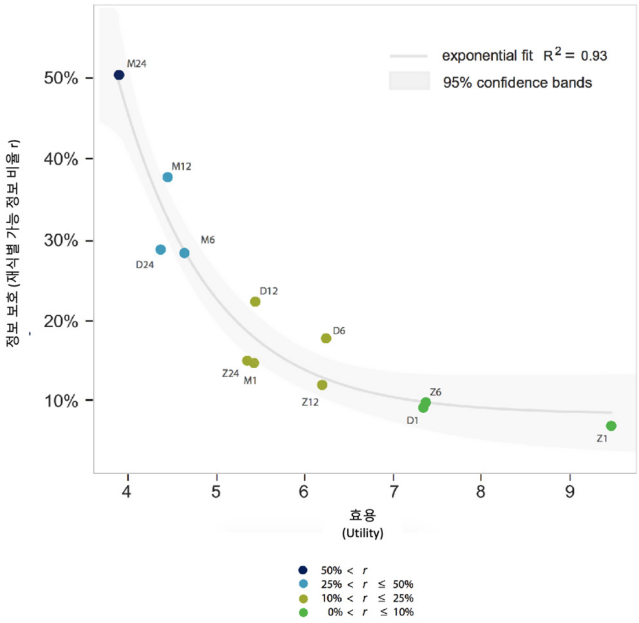
49) Machanavajjhala et al.(2007)

50) Li et al.(2007)

51) 데이터 효용의 측정에 대해서는 다음 절에서 자세히 설명함. 여기서 데이터의 효용은 원데이터가 가지고 있는 유용한 정보 정도의 원론적인 정의로 사용했음

계(Utility-Privacy trade-off)라 하는데 〈그림 Ⅳ-2〉는 이러한 관계를 보여주고 있다.

〈그림 Ⅳ-2〉 데이터 정보보호와 효용의 상충관계 예시



자료: Calacci, D. Berke, A., Larson, K., and Pentland, A.(2019)

〈그림 Ⅳ-2〉에서 가로축은 데이터 유용성을, 세로축은 데이터 정보보호의 수준 혹은 강도를 나타내며 이 둘의 상충관계는 아래로 볼록한 곡선로 표현된다. 그림 내에서 우하단의 데이터셋 Z1은 최대의 유용성을 가지지만 동시에 최하의 정보보호 수준을 보여주고 있어 노출 제어가 되지 않은 원데이터에 가깝고, 좌상단 데이터인 M24의 경우 반대로 최고의 정보보호 수준을 보여주지만 유용성이 최하임을 알 수 있다. 즉, Z1을 원데이터라고 할 때 여기 노출 제어의 수준을 높힘에 따라 커브를 따라 좌상의 방향으로 움직여 노출 제어의 수준을 최대로 끌어올리게 되면 M24의 위치로 이동하게 되는 것이다. 이처럼 유용성과 정보보호를 동시에 최대화할 수는 없다는 사실을 받아들인다면, 가장 적절한 데이터 노출 제어의 수준은 그림에서 커브의 중간지점에 존재하는 데이터에 상응한다는 것을 알 수 있다. 그러나 이를 위해서는 데이터의 유용성과 정보보호(혹은 노출위험)를 동시에 측정하는 작업이 선행되어야만 하는데 이 둘을 동시에 측정하는 연구는 최근에 와서야 등장하고 있다.

3. 효용과 노출위험의 측정 기법들

통계적인 관점에서 볼 때 잘 생성된 재현데이터는 원데이터가 속한 동일한 모집단으로부터 독립적으로 랜덤 추출된 또 다른 표본으로 이해할 수 있다. 따라서 좋은 재현데이터는 원데이터와 충분히 유사하여 그 효용을 보존하는 동시에 개인정보 노출위험을 최소화하기 위해 충분히 다른 것이어야 한다. 재현데이터의 유용성과 정보보호는 서로 상충관계에 있어 이 둘 사이의 균형을 달성하는 것은 데이터 재현(합성)의 중요한 목표이자 고려해야 하는 핵심사항이다.

사실 위의 상충관계는 재현데이터뿐 아니라 다른 비식별화 처리된 데이터에 대해서도 마찬가지이지만 여기서는 논의의 편의를 위해 재현데이터로 지칭하겠다. 재현데이터의 효용측정은 문헌에서 두 개의 다른 수준에서 논의되어왔다. 첫째, 개별 분석의 수준에서 재현데이터와 원데이터 간의 데이터 요약 또는 추정된 통계량을 비교하는 방식이 있다. 즉, 두 데이터에서 각각 계산된 분석 통계량의 값이 충분히 가까우면 재현데이터는 높은 효용을 갖는 것으로 간주되는 것이다. 그러나 이러한 분석 특정(Analysis-specific) 측정 방법은 사용자가 수행할 분석의 종류에 따라 측정된 효용의 정도가 다르게 된다. 이 방식은 사용자가 수행하는 분석의 종류나 범위가 알려져 있지 않는 것이 일반적이기 때문에 적용 가능한 범위가 제한적이다. 이러한 단점을 극복하기 위해, 특정 분석이 아니라 표본의 수준에서 데이터 효용을 측정하는 척도 역시 연구되어왔다. 이는 재현 전후의 데이터 분포의 차이를 측정함으로써 모든 가능한 분석에 대한 데이터의 효용을 평가하는 것으로 글로벌(Global)척도라 부른다. 최근 소개된 경향점수평균제곱오차(Propensity Score Mean Squared Error; PMSE) 기법[52]이 대표적인 예이다.

PMSE는 분류기로부터 얻어지는 경향점수에 기반한다. 설명을 위해 테이블 형태의 정형 원데이터와 이로부터 생성된 재현데이터를 얻었다고 가정하자. 이 두 데이터를 상하로 이어붙인 후 새로운 변수(열)를 추가하고 각 행이 원데이터(0)인지 재현데이터(1)인지의 정보를 입력한다. 다음 단계로 통합된 데이터에 분류기를 적용해 각 행이 실제 속하는 데이터로 잘 분류가 되는지(0/1)를 확인하게 되는데, 여기서 분류기는 로지스틱회귀, CART, 혹은 다른 이진 분류기(Binary classifier)를 적절히 사용할 수 있다. 이제 n_0와 n_1을 각각 원데이터와 재현데이터의 크기라고 하고, 분류기가 각 행에 대해 재현데이터(1)에 속할

52) Snoke, J. et al.(2018)

확률을 추정한 값을 $\hat{p_i}$라고 하자. 즉 $\hat{p_i}$ 값이 1에 가까울수록 재현데이터로 분류될 가능성이 높아진다. 만약 재현이 성공적이라면 분류기는 재현 전후의 데이터를 구별하기 어렵게 되어 $\hat{p_i}$ 값이 $c = n_1/(n_0 + n_1)$ 에 가깝게 될 것이 자명하다. 이러한 성질을 이용하기 위해 PMSE는

$$pMSE = \frac{1}{n_0 + n_1} \sum_{i=1}^{n_0+n_1} \left(\hat{p_i} - c\right)^2.$$

으로 정의된다. 따라서 성공적인 재현데이터에 대해서 PMSE는 0에 가깝게 되고, 반대로 재현데이터와 원데이터의 차이가 많은 경우 PMSE의 값은 커지게 됨을 알 수 있다. PMSE가 0에 충분히 가까운지 아닌지를 구분하는 경계에 대한 논의는 해당 논문에서 확인할 수 있다.

그러나 이 기법에는 몇 가지 제한이 있다. 첫째, 재현데이터와 원데이터가 동일한 모집단 분포를 따르더라도 점근적 PMSE의 값이 매우 다를 수 있는데 이는 제안된 방법이 재현데이터 생성 방법에 크게 의존함을 뜻한다. 둘째, PMSE 기법은 분류기에 민감하다고 알려져 있다. 즉 동일한 재현데이터에 대해 어떤 분류기를 사용하느냐에 따라 PMSE 값이 상당히 다를 수 있다는 것이다. 마지막으로, PMSE 기법은 데이터의 효용만을 평가하며 개인정보 노출위험 대한 정보를 제공하지 않는다.

전통적인 노출위험 척도인 k-익명성, l-다양성, t-근접성도 모수의 결정이 임의적이라는 점과 데이터 효용을 계량적으로 측정할 수 없다는 약점이 있다. 좀 더 최근의 척도로는 앞서 기술한 차등적 정보보호(DP)를 들 수 있다. 데이터 정보보호 연구에서 유용한 도구로 인정되고 있지만 DP 역시 한계점이 존재한다. 첫째, 반복되는 쿼리에 대해 점차적으로 효율이 떨어진다. 둘째, 쿼리의 종류에 따라 적절한 매개변수 ε의 값을 결정하는 것은 현실적으로 매우 어려운 문제이다. 실제로 특정 매개변수 값이 경우에 따라 매우 다른 정보보호 수준을 의미할 수 있어, 쿼리(분석)의 종류나 범위가 알려져 있지 않은 것이 일반적인 상황에서 DP를 사용한 마이크로데이터의 노출 제어는 현실적으로 매우 어렵다. 이 두 단점은 앞에서도 기술하였지만, 여기서는 추가로 DP 기법 역시 노출위험을 제어한 후 감소한 데이터의 효용을 측정할 수 없다는 점을 지적한다.

4. DUPI: 분포 기반 재현데이터 품질평가 척도

위와 같이 현재 사용되는 척도들은 대부분 데이터의 효용만을 혹은 노출위험만을 측정하고 있는데, 서로 상충관계에 있는 두 측면을 동시에 측정하는 척도가 필요함은 자명하다. 이 둘을 같이 고려한 재현데이터의 성능을 품질이라고 할 수 있는데, 사실 재현데이터 연구에서 재현된 데이터의 품질 측정은 재현데이터의 생성기법만큼이나 중요한 주제이다. 새로운 노출 제어(가명·익명처리 혹은 재현데이터) 기법은 결국 노출 제어를 통해 산출된 데이터의 품질을 통해 평가받아야만 하기 때문이다.

이와 관련해 기억해야 할 중요한 내용은 재현 전후 데이터를 비교할 때 분석특정(Analysis-specific) 척도의 한계점이다. 앞서 보험데이터를 이용한 예시에서 우리는 보험가입 여부를 반응변수로 둔 로지스틱회귀모형을 포함해 다양한 비교를 하였다. 만약 사용자가 특정한 분석만을 진행할 것이라는 것을 미리 안다면 재현 전후 데이터의 효용과 노출위험을 쉽게 통제할 수 있을 것이지만, 일반적으로 우리는 재현된 데이터가 어떤 분석이나 쿼리에 이용될지 알 수 없다. 예를 들어, 동일한 보험데이터에서 다른 변수를 반응변수로 두고 일반화선형모형으로 이용할 수도 있고, 전혀 새로운 모형으로 실험을 할 수도 있을 것이다. 또한 앞서 살펴본 두 변수 사이의 공분산이나 상관계수 역시 저차원의 적률만을 비교하는 것이기 때문에 고차원의 종속성을 측정하지 못한다. 이를 통해 우리는 가능한 모든 쿼리나 분석에 대해 재현데이터의 효용을 측정하는 것은 불가능함을 알 수 있다. 더 나아가 동일한 재현데이터를 사용하더라도 쿼리나 분석의 종류와 범위에 따라 품질도 다르게 측정될 것이다. 이런 이유로 재현 전후 데이터를 비교할 때는 특정한 분석 결과가 어떻게 달라지느냐를 비교하는 것도 필요하지만, 더 근본적으로는 데이터의 분포가 어떻게 달라지는지를 측정할 필요가 있다. 엄밀하게 말해 분석의 결과가 유사한 것과 두 데이터가 유사한 분포를 가지는 것은 매우 다른 이야기일 수 밖에 없기 때문이다. 결론적으로, 재현데이터의 품질을 측정할 때 분포 수준에서 원데이터와 비교하는 것이 합당한 방법이며 이를 확인하는 과정이 반드시 필요하다.

이러한 맥락에서 최근 개발된 재현데이터의 효용과 노출위험을 동시에 측정하는 품질척도인 DUPI(Data Utility and Privacy Index)[53]를 소개한다. PMSE 기법과 마찬가지로 특정한 쿼리(분석)에 의존하지 않는 글로벌척도로서 제안된 DUPI는 재현 전후 데이터 사이

53) Jeong, D., Kim, J. H., and Im, J.(2022)

의 거리를 분포에 기반해 확률적으로 측정하는 방식을 이용한다. 기본적인 개념은 원데이터가 재현데이터와 동일한 다변량 모집단에서 독립적으로 추출된 표본일 경우 가장 이상적이라 볼 수 있다는 사실에서 출발한다. DUPI는 비모수적 방법으로 모집단의 분포를 특정하지 않는다는 장점이 있고 연속형과 범주형을 둘 다 포함하는 데이터에 적용가능하다.

원론적인 수준에서 볼 때 재현 전후 데이터 사이의 거리가 지나치게 가까우면 두 데이터의 유사성이 충분히 강해 재현데이터가 원데이터의 효용을 유지하지만 동시에 노출위험 역시 원데이터와 비슷한 수준으로 높아진다. 반대로 거리가 멀수록 재현데이터는 원데이터와의 유사성이 줄어 효용 측면에서 정보손실이 커지지만 그만큼 노출위험이 감소하게 된다. DUPI는 이러한 확률적 거리 개념을 이용하여 재현데이터의 효용과 노출위험을 동시에 측정 및 수치화하고, 재현된 데이터가 상충관계 사이에서 어느 지점에 위치하는지 좌표평면에 그림으로도 시각화할 수 있다. 또한 원데이터와 재현데이터가 동일모집단에서 추출되었다는 가정을 이용하여 재현데이터가 가질 수 있는 최적 위치 역시 제공한다.

DUPI의 이론을 간략히 살펴보기 위해 먼저 $X_n = \{x_{1,} \cdots, x_n\}$를 크기가 n인 원데이터로, $Y_m = \{y_{1,} \cdots, y_m\}$을 크기가 m인 재현데이터로 두자. 여기서 m은 n과 다를 수 있다. 이제 재현 전후 데이터 사이의 거리를 측정하기 위해 거리함수가 필요한데, 이를 위해 특정한 두 점 x와 y 사이의 거리를 계산하는 함수를 $d(x,y)$라고 하겠다. 거리함수는 유클리드 거리 혹은 범주형을 포함하는 경우 HEOM(Heterogeneous Euclidean-Overlap Metric) 거리 등을 사용할 수 있다. 다음으로, 어떤 고정된 점 c과 임의의 데이터 A 사이의 거리를 재기 위해 c와 A의 개별원소들 사이의 거리들을 모두 계산하고, 그 중 k번째로 작은 값을 $d_A^{<k>}(c)$로 둔다. 다시 말해

$$d_A^{<k>}(c) = \{d(a,c) : a \in A\}^{<k>}$$

로 정의한다. 이를 이용하면 재현 전후 데이터가 동일한 모집단에서 추출되었다는 귀무가설하에서 다음의 식을 증명할 수 있다.

$$\frac{1}{n}\sum_{i=1}^{n} P\left(d_{Y_m}^{<k>}(x_i) \le d_{X_{n\setminus i}}^{<k>}(x_i)\right) = \sum_{s=k}^{2k-1} \frac{\binom{s-1}{k-1}\binom{n+m-s-1}{m-k}}{\binom{n+m-1}{m}}$$

여기서 $X_{n \setminus i}$는 원데이터 X_n에서 i번째 관측값 x_i이 빠진 것이다. 만약 i번째 관측값이 포함되면 x_i는 자신과의 거리인 0을 산출하게 되기 때문이다. 즉, $d_{X_n}^{<1>}(x_i) = 0$이 되는 상황을 피하기 위한 장치이다. 위의 이론적 결괏값을 $DUPI_0^{<k>}$로 정의하고, 이 값의 표본 추정량을 $DUPI^{<k>}$로 두면 지시함수 $I(\cdot)$를 이용해

$$DUPI^{<k>} = \frac{1}{n}\sum_{i=1}^{n} I\left(d_{Y_m}^{<k>}(X_i) \le d_{X_{n \setminus i}}^{<k>}(X_i)\right)$$

를 얻는다. 논문에서는 k값을 1로 두고 있는데 이는 큰 k값을 선택할 경우 데이터의 극단치 혹은 이상치에 민감하게 되고 $k = 1$만으로도 재현 전후 데이터의 특징을 충분히 비교할 수 있기 때문이다. 따라서, 위의 식에 의해 $DUPI_0 = DUPI_0^{<1>}$, $DUPI = DUPI^{<1>}$가 된다. 이렇게 계산된 $DUPI$ 값은 0과 1 사이의 값을 지니며 다음의 해석이 가능하다.

- $DUPI$값이 1에 가까우면 재현 전후 데이터가 지나치게 유사해 생성된 재현데이터가 효용은 높지만 노출위험 또한 크다.
- 반대로 $DUPI$값이 0에 가까우면 재현 전후 데이터가 지나치게 달라서 생성된 재현데이터의 노출위험은 최소화할 수 있지만 원데이터의 정보를 거의 보존하지 못한다.
- 이상적인 $DUPI$값은 $DUPI_0$로서 이 경우 귀무가설을 만족해 재현데이터가 원데이터와 동일한 모집단에서 독립적으로 생성된 표본으로 볼 수 있다.

이러한 성질을 이용해 논문은 재현데이터의 $DUPI$값이 효용과 노출위험의 관점에서 최적점인 $DUPI_0$에서 얼마나 떨어져 있는지에 대한 시각화 기법을 제안한다. 시각화를 위해 먼저

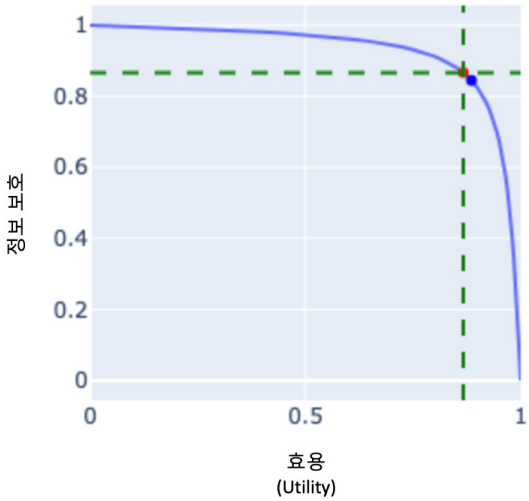
$$g(DUPI) = \begin{cases} \dfrac{DUPI}{2DUPI_0}, & 0 \le DUPI \le DUPI_0, \\ \dfrac{DUPI - DUPI_0}{2(1 - DUPI_0)}, & DUPI_0 < DUPI \le 1. \end{cases}$$

를 정의하고, nuisance 모수인 τ를 이용해

$$UI(DUPI) = \frac{\arctan[\tau \times g(DUPI)]}{\arctan(\tau)}$$
$$PI(DUPI) = \frac{\arctan[\tau \ - \ \tau \times g(DUPI)]}{\arctan(\tau)}$$

로 변환을 하고, 이로부터 주어진 재현데이터의 (UI, PI) 좌표값을 도출해 낸다. 여기서 디폴트 τ값은 5로 지정되었다.

〈그림 Ⅳ-3〉 DUPI 시각화 예시



자료: Jeong, D., Kim, J. H., and Im, J.(2022)

〈그림 Ⅳ-3〉은 앞서 예시로 보인 보험데이터에 DUPI를 적용해 시각화한 결과이다. 그림에서 가로축은 데이터의 효용(Utility Index; UI)을, 세로축은 노출보호(Privacy Index; PI)를 뜻하며 두 축 모두 [0,1] 구간에 존재한다. 위로 볼록한 곡선은 주어진 데이터에 대해 가능한 모든 (UI, PI) 조합의 매핑인데, 곡선의 우측하단으로 갈수록 효용은 크지만 노출보호가 낮은 재현데이터를, 반대로 좌측상단은 높은 노출보호가 보장되지만 낮은 효용을 가지는 재현데이터를 나타낸다. 이 그림은 앞서 소개한 〈그림 Ⅳ-2〉와 비슷한 정보를 제공하지만 세로 축의 의미가 반대이므로 곡선 역시 서로 대칭적으로 보여진다.

위 그림에서 최적의 재현데이터는 점선으로 표시된 두 직선이 교차하는 지점에 있는 빨간

색 점에 위치하며 효용과 노출위험이 가장 적절히 상쇄되는 기준점이다. 곡선 위에 찍혀 있는 검은 점은 앞장에서 살펴본 보험데이터를 CART 알고리즘으로 재현한 데이터의 위치를 나타낸다. 이 점은 최적점에 비해 우하단에 위치해 있어 재현된 데이터가 최적데이터보다 비교적 높은 수준의 효용을 가지고 있지만 노출보호는 상대적으로 약하다는 것을 알 수 있다. 그럼에도 불구하고 재현데이터는 최적점과 가까운 곳에 위치해 있어 재현이 잘되었다고 평가할 수 있다. 그림에서 표현된 재현데이터와 기준점의 좌표를 도출하고 수치적 비교를 하는 과정은 원논문[54]을 참고하면 되겠다.

그러나 데이터의 크기가 커지면서 계산량이 지수적으로 증가하고, 재현데이터 품질을 표현하는 척도가 점 추정으로만 되어있어 이 추정량의 불확실성이 어느 정도가 되는지는 알 수 없다는 점에서 DUPI 방법론은 개선될 여지가 있다고 하겠다.

또 다른 한계점으로 최적점에서 어느 정도로 멀거나 가까울 때 노출위험이 없다고 할 수 있는지에 대한 수치적인 가이드라인이 없다는 점을 지적할 수 있다. 그러나 이는 현재 존재하는 모든 계량적인 기법과 척도에 공통적으로 해당되는 사항이다. 재현데이터를 원데이터와 비교했을 때 원데이터와 완전히 동일한 관측값이 없다면 일차적으로는 노출위험이 없다고 할 수 있다. 그러나 완전히 동일하지는 않지만 유사한 관측값은 존재할 것이고 이를 다양한 외부 데이터와 연결한다면 개인정보의 일부라도 노출이 될 수도 있다는 점에서, 모든 가능한 시나리오하에서 노출위험을 수리적으로 0으로 만드는 것은 불가능하다. 따라서 노출위험은 없애는 것이 아니라 최소화하는 것이며 이를 위해 전문가나 규제당국이 토론과 경험을 통해 허용 가능한 노출위험의 수준을 정하는 논의가 필요하다. 비슷한 예로서 금융사의 자산건전성을 관리·감독하는 국제적 규제인 바젤Ⅱ 어코드의 경우 위험자본의 양을 계산할 때 Value-at-Risk(VaR) 99.9% 수준을 요구하고 있다. 이 수치 역시 이론적인 근거로부터 도출된 것이 아니라 특정 위험을 100%의 확률로 차단하거나 피하는 것이 현실적으로 불가능하다는 것을 받아들여 바젤위원회가 금융사의 이윤추구와 소비자·투자자의 공익이라는 상충관계를 고려해 결정한 위험허용의 수준으로 볼 수 있다.

54) Jeong, D., Kim, J. H., and Im, J.(2022)