The Korean Journal of Applied Statistics
2023, Vol. 36, No. 4, 323–335

DOI: https://doi.org/10.5351/KJAS.2023.36.4.323
Print ISSN1225-066X / Online ISSN 2383-5818

Overview of estimating the average treatment effect using dimension reduction methods

Mijeong Kim[1,a]

[a]Department of Statistics, Ewha Womans University

Abstract

In causal analysis of high dimensional data, it is important to reduce the dimension of covariates and transform them appropriately to control confounders that affect treatment and potential outcomes. The augmented inverse probability weighting (AIPW) method is mainly used for estimation of average treatment effect (ATE). AIPW estimator can be obtained by using estimated propensity score and outcome model. ATE estimator can be inconsistent or have large asymptotic variance when using estimated propensity score and outcome model obtained by parametric methods that includes all covariates, especially for high dimensional data. For this reason, an ATE estimation using an appropriate dimension reduction method and semiparametric model for high dimensional data is attracting attention. Semiparametric method or sparse sufficient dimensionality reduction method can be uesd for dimension reduction for the estimation of propensity score and outcome model. Recently, another method has been proposed that does not use propensity score and outcome regression. After reducing dimension of covariates, ATE estimation can be performed using matching. Among the studies on ATE estimation methods for high dimensional data, four recently proposed studies will be introduced, and how to interpret the estimated ATE will be discussed.

Keywords: augmented inverse probability weight, average treatment effect, dimension reduction, inverse probability weight, propensity score

1. 서론

인과 추정에서 가장 중요한 것은 교란 변수를 통제하는 것이다. 교란 변수가 처리와 결과 사이에 허위 연관성(spurious association)을 나타나게 하여 인과 추정을 편향시킬 수 있기 때문이다. 무작위 통제 실험(randomized controlled trial; RCT)에서 적절하게 설계되고 실행된 무작위 배정은 교란 변수를 통제 하기 위한 가장 좋은 방법이지만, 현실적으로 RCT를 수행하는 것은 불가능하거나 비용이 많이 드는 경우가 많다. 이러한 이유로, 관측 데이터로부터 인과 관계를 추정하는 것이 필수적이며 중요한 문제로 인식되고 있다.

Adjustment set을 통제하는 것은 관측 데이터로부터 인과 효과 추정할 때 편향을 제거하는 효과적인 방법이다. Adjustment set은 교란 변수를 통제한다는 의미에서 deconfounding set이라고도 한다. Adjustment set의 크기는 인과 효과 추정량의 성능에 영향을 미칠 수 있으며, 작은 수의 변수로 편향을 제거할 수 있다면 adjustment set의 크기는 작은 것이 좋다. 적절한 adjustment set을 결정하기 위한 접근 방식으로는 다음과 같은 두 가지가 있다. (1) 모든 공변량을 이용하여 편향을 제거하는 방법, (2) backdoor 기준 또는 그 변형과 같은

This work was supported by National Research Foundation of Korea (NRF) grant funded by the Korean Government (NRF-2020R1F1A1A01074157).

[1] Department of Statistics, Ewha Womans University, 52 Ewhayeodae-gil, Seodaemun-gu, Seoul 03760, Korea. E-mail : m.kim@ewha.ac.kr

Published 31 August 2023 / journal homepage: http://kjas.or.kr

기준을 기반으로 한 공변량 변수의 하위 집합을 이용하는 것이다. 첫 번째 방법은 adjustment set에 속한 공변량이 많을 경우 편향이 더 생기기도 한다 (De Luna 등, 2011). 두 번째 방법은 backdoor 기준을 만족하는 변수를 찾기 위한 인과 관계 그래프가 필요한데, 그것은 일반적으로 알려져 있지 않고, 데이터에서 인과 관계 구조를 학습하는 것은 가능하지만 모든 관계를 정확히 알아내기는 어렵다. 최근에 이 두 가지 방법을 대체할 방법으로 다양한 차원 축소 방법이 제시되었다. 차원 축소 방법을 이용하여 교란 통제를 위한 새로운 축소된 공간에 작은 수의 변수 집합을 만들 수 있다. 예를 들어, 성향 점수는 개인이 처치(treatment)를 받을 확률인데, 여러 공변량을 이용하여 스칼라(1차원) 값을 갖는 성향점수를 계산하여 교란을 통제한다. 참고로, treatment는 처리, 처치 또는 처방으로 번역될 수 있는데, 이 중 문장 속에서 가장 자연스러운 단어를 선택하여 서술하고자 한다. 이 논문에서는 최근에 제시된 몇 가지 차원 축소 방법을 소개하고자 한다. 2장에서는 인과 추정의 기본 가정을 설명하고, 3장에서는 인과 추정에 대한 이해를 돕고자 역확률 가중치 방법을 설명하고, 고차원 데이터 분석에 적용 가능한 차원 축소 방법에 대해서 간단히 설명한다. 4장에서는 최근에 제시된 차원 축소 인과 추론 방법에 대해서 설명하고, 5장에서는 ATE 추정량 해석시 유의할 점을 언급하고자 한다. 6장에서는 소개한 인과 추론 방법을 이용하여 데이터 분석을 하고, 7장에서는 논문을 마무리하고자 한다.

2. 인과 추정의 가정

인과 추정은 관측 데이터를 이용하여 통제된 실험과 같은 효과를 얻는 추정 방법이다. 이해를 돕기 위해 인과 추정을 실험 계획의 랜덤화와 비교하여 설명하고자 한다. 실험 계획에서는 랜덤화를 통해 실험하고자 하는 인자 외에 실험 결과에 영향을 미칠 수 있는 외부 조건을 통제하도록 계획한다. 예를 들어, A라는 식품 또는 약이 혈압에 미치는 영향을 조사한다고 하자. A를 처방받는 사람들의 집단은 실험집단(또는 처치집단), A를 처방받지 않은 사람들의 집단은 통제집단이라고 한다. 실험 계획시, 랜덤화를 통해서 A 이외에 다른 조건은 실험집단과 통제집단이 동일하도록 한다. 만약에 실험집단은 당뇨 질환자(공복 혈당 126mg/dl 이상), 통제집단은 정상 혈당인 사람들로 구성되어 있다면 결과가 A로 인한 것인지, 혈당수치의 차이로 인한 것인지 구별하기가 어렵기 때문이다. 즉, 혈당수치는 결과에 영향을 줄 수 있는 교란 변수가 된다. A가 부작용이 거의 없는 식품이라면 A를 무작위로 할당하는 실험인 RCT가 가능하지만, A가 부작용 위험이 있는 약이라면 RCT를 진행할 수 없다. 많은 경우에 RCT가 불가능하기 때문에 관측 데이터를 활용해야 한다. A가 혈당에 효과가 있을 것이라는 기대감이 높은 상태라면, 많은 당뇨 질환자들이 이미 A를 섭취하고 있을 가능성이 높다. 이런 상황에서 관측 데이터를 살펴보면, A를 섭취하는 사람들의 대부분이 당뇨 질환자, A를 섭취하지 않는 사람들의 대부분은 정상 혈당인 사람들로 구성되어 있을 가능성이 있다. 이런 경우에는 당뇨 질환자 중에서 A를 섭취하는 사람들과 그렇지 않은 사람들을 나누어서 결과를 비교하고, 정상 혈당인 사람들 중에서 A를 섭취하는 사람들과 그렇지 않은 사람들을 나누어서 확인한다면, 혈당의 교란 효과를 통제하여 좀 더 개선된 결과를 얻을 수 있을 것이다. 다시 말하면, 혈당수치를 조건부로 하여 관측치를 살펴보는 것이다.

처방을 받으면 $T = 1$, 받지 않으면 $T = 0$으로, i 번째 사람에 대한 잠재적 결과(potential outcome)는 $T_i = 1$, $T_i = 0$에 대해서 각각 Y_{1i}와 Y_{0i}로 표시하도록 한다. 잠재적 결과 중 하나만 관측되기 때문에, 관측되지 않은 잠재적 결과를 반사실(counterfactual)이라고 한다. $Y_i \in \mathbb{R}$은 i 번째 사람에 대한 관측된 결과라고 하자. 관측된 결과 $Y_i \equiv T_i Y_{1i} + (1 - T_i) Y_{0i}$로 표시할 수 있다. 평균처리효과(average treatment effect; ATE)는 처리 여부에 따른 잠재적 결과의 차이의 기대값으로 다음과 같이 정의한다.

$$\tau \equiv E\left(Y_{1i} - Y_{0i}\right).$$

Assumption 1. *(강한 무시 가능성 가정(unconfoundedness/strong ignorability)) 처방 전 변수(pretreatment) X를 조건부로 했을 때, 잠재적 결과는 처리 T와 독립이다.*

$$(Y_{1i}, Y_{0i}) \perp\!\!\!\perp T_i \mid X_i.$$

X를 조건부로 했을 때, 잠재적 결과는 처치 여부과 상관이 없다는 의미이다. 위에서 든 예를 이용하여 부연 설명을 하고자 한다. A를 섭취하고 있다면 이미 건강 상태가 좋지 않을 가능성이 있으므로, A의 효과가 크지 않을 것으로 예상할 수 있다. 이 경우에는 잠재적 결과는 A 섭취 여부와 관련이 있다. 하지만, 정상 혈당인 사람들의 데이터(또는 당뇨질환자의 데이터)만 이용하는 경우, 즉 X를 조건부로 할 때, A 섭취 집단(처치집단)과 그렇지 않은 집단(통제집단)은 사람들의 건강 상태가 동일한(분포가 같은) 것으로 볼 수 있다. 즉, X가 deconfounding 역할을 하여 X를 조건부로 할 때에는 A 섭취 여부로 잠재적 결과를 예상할 수 없다. 관측 데이터에서 deconfounding set을 조건부로 하면, 실험계획에서 랜덤화를 통해 실험집단과 통제집단의 분포를 동일하게 하는 것과 같은 효과를 얻을 수 있다.

Assumption 2. *(Overlap/positivity) 처방 전 변수 X를 조건부로 했을 때, 처방 받을 확률은 0 또는 1이 될 수 없다.*

$$0 < P(T_i = 1 \mid X_i) < 1.$$

Glymour 등 (2016)에서는 X를 조건부로 한다는 것은 같은 X 값을 갖는 집단으로 필터링하는 것과 같다고 설명한다. 실제로는 수식 안에서 조건부를 통해 X를 고정시키는 것을 뜻하지만, 이해를 돕기 위해 Glymour 등 (2016)의 표현을 이용하고자 한다. 위에서 든 예에서는, 정상 혈당인 사람들(또는 당뇨 질환자)로 필터링하여 이용하는 것과 같다. 특정 $X = x$로 필터링된 데이터에 A를 섭취하는 사람이 한 사람도 없거나 모든 사람들이 A를 섭취하고 있다면, A 섭취 여부에 따른 결과의 차이를 조사할 수 없다. 필터링된 데이터에는 처방받은 사람들과 그렇지 않은 사람들이 섞여 있어야 한다는 의미이다.

3. 역확률 가중치 및 차원 축소 방법

3.1. 역확률 가중치와 확장 역확률 가중치

실험계획에서 2수준 요인배치법을 생각해보자. 하나의 요인에 대해서 두 가지 수준, 예를 들어 A를 처방하거나($T = 1$) 안 하는($T = 0$) 두 가지가 있는 경우이다. 반복적인 실험의 경우, 각각의 수준에 대해서 같은 횟수로 반복해야 한다. 만약 실험이 잘못 되어, 실험집단에서 30개의 데이터, 통제집단에서 10개의 데이터를 얻었다고 하자. 이 데이터로 회귀분석을 한다면 회귀식의 패턴은 통제집단의 특성보다는 주로 실험집단의 특성을 반영하는 형태, 즉 편향된 결과를 얻게 될 것이다. 이런 측면에서, 실험에서 실험집단과 통제집단의 균형을 맞추는 것은 중요하다. 관측 데이터의 경우에는 실험과 달리 균형된 데이터를 얻기 어렵다. 당뇨 질환자 중 A를 섭취한 사람(처치집단)과 그렇지 않은 사람들(통제집단)의 비율이 3 : 1이라고 하자. 각 집단의 데이터의 균형이 맞지 않으므로, 처치집단에 가중치를 1/4, 통제 집단에 가중치를 3/4을 할당하면 균형을 맞출 수 있다. 이런 관점에서 생성된 용어가 성향점수(propensity score; PS)인데, PS는 처방 받을 확률 $\pi(X) = P(T = 1|X)$을 뜻한다. PS 모형으로는 공변량을 이용한 로지스틱 회귀분석 모형이 주로 이용된다. 처치집단에는 1/PS로, 통제집단에는 1/(1-PS)로 가중치를 할당하게 되는데 이것을 역확률 가중치(inverse probability weighting; IPW)라고 한다. 역확률 가중치 방법은 Robins 등 (1994)에 의해 소개되었으며, ATE에 대해서 다음과 같은 수식으로 나타낼 수 있다.

$$\widehat{\tau} = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{T_i Y_i}{\widehat{\pi}(X_i)} - \frac{(1-T_i)\,Y_i}{1-\widehat{\pi}(X_i)}\right\}. \tag{3.1}$$

그러나 관측치의 수가 작을 경우, PS의 추정치 $\widehat{\pi}(X_i)$가 0 또는 1에 아주 가까운 값이 되고 추정치의 분산이 커진다. 이런 점을 보완하기 위해 개선된 추정량이 제시되었는데, Hahn (1998)이 제시한 semiparametric efficiency bound를 갖는 추정량과 확장 역확률 가중치(augmented inverse probability weighting; AIPW) 추정량 등이

있다. AIPW 추정량은 PS 모형과 결과 회귀(outcome regression; OR)모형을 이용한 ATE에 대한 추정치로써 다음과 같다.

$$\widehat{\tau} = \frac{1}{n}\sum_{i}^{n}\left\{\frac{T_i Y_i}{\widehat{\pi}(X_i)} - \frac{T_i - \widehat{\pi}(X_i)}{\widehat{\pi}(X_i)}\widehat{E}\left(Y_{1i} \mid X_i\right)\right\} - \frac{1}{n}\sum_{i}^{n}\left\{\frac{(1-T_i)\,Y_i}{1-\widehat{\pi}(X_i)} - \frac{T_i - \widehat{\pi}(X_i)}{1-\widehat{\pi}(X_i)}\widehat{E}\left(Y_{0i} \mid X_i\right)\right\}. \tag{3.2}$$

이론적으로는 PS와 OR 중 적어도 하나가 consistent하게 추정되면, ATE에 대한 consistent 추정치를 얻을 수 있다 (Glynn과 Quinn, 2010, Appendix A.1). 이런 이유로 AIPW 추정량은 이중 강건 추정량(doubly robust estimator)라고 불린다. AIPW 추정량은 (3.2)에서 모든 항이 consistent하게 구해진다면, 이론적으로 최소 분산을 갖기 때문에 IPW 추정량보다 efficient하다. 그러나 PS와 OR 중 하나만 consistent할 경우에는 AIPW 추정량의 분산은 비교적 큰 값을 갖는다. Kang과 Schafer (2007)에 따르면, PS와 OR 두 가지 모두 적절하게 추정되지 않은 경우에는 이중 강건 추정량은 편향이 상당히 커질 수 있다. Carpenter 등 (2006), Kang과 Schafer (2007), Vansteelandt 등 (2012)는 관측치가 적을때 PS와 OR 중 하나는 잘 추정되고 다른 하나의 편향이 크지 않을 때에도 ATE의 편향은 증폭될 수 있음을 지적하였다. 또한 PS와 OR에 대하여 모수 모형을 가정할 경우, 모수 모형은 유연하지 않고 제한적인 형태이므로 잘못된 추정으로 인해 큰 편향이 발생할 수 있다. Hahn (1998)은 PS와 OR에 대해서 특정 모수 모형을 가정하는 것보다 비모수 모형을 이용하는 것을 권하였다.

3.2. 차원 축소 방법

공변량이 고차원일 때에는 충분 차원 축소(sufficient dimension reduction; SDR) 방법을 이용하여 효율적인 추정량을 구할 수 있다. 처리가 이진 변수일 때($T = 1$ 또는 $T = 0$), 다음을 만족하는 projection $\boldsymbol{B} \in \mathbb{R}^{p\times r}(r < p)$가 존재한다.

$$T \perp\!\!\!\perp X \mid \boldsymbol{B}^{\mathrm{T}}X.$$

이 때, $\boldsymbol{B}^{\mathrm{T}}X$는 $\boldsymbol{B}$의 column subspace에 X를 정사영(orthogonal projection)한 것이다. $\boldsymbol{B}$의 column subspace를 dimension reduction space (DRS)라고 한다 (Cook, 1996; Cook, 2009). 이러한 성질을 만족하는 $\boldsymbol{B}$는 여러 가지가 존재할 수 있는데, 그 중 차원이 가장 작은 $\boldsymbol{B}$의 column space를 central DRS라고 한다. Central DRS는 최소 차원이면서 유일한 공간이다 (Cook, 1996). $E(X|\boldsymbol{B}^{\mathrm{T}}X)$ 또는 $\mathrm{Var}(X|\boldsymbol{B}^{\mathrm{T}}X)$에 대해서 선형 모형을 적용한 방법으로는 sliced inverse regression (SIR) (Li, 1991), sliced average variance estimation (SAVE) (Cook과 Weisberg, 1991), directional regression (Li과 Wang, 2007), generalized directional regression (Li와 Dong, 2009; Dong과 Li, 2010)가 있다. SIR, SAVE와 directional regression 방법은 $E(X|\boldsymbol{B}^{\mathrm{T}}X)$를 X의 선형 모형으로 가정하였으나, Ma와 Zhu (2012)는 $E(X|\boldsymbol{B}^{\mathrm{T}}X)$에 대해서 특정 모수 가정을 하지 않고 준모수 방법을 이용하여 central DRS을 추정하는 방법을 제시하였다.

ATE를 추정하기 위해서 IPW 또는 AIPW 방법을 이용할 때에는 PS와 OR, 즉, 식 (3.1), (3.2)에서 $\pi(X_i)$, $E(Y_{1i}|X_i)$, $E(Y_{0i}|X_i)$를 추정해야 한다. 인과 추론에서 조건부 공변량 X가 고차원일 때에는 무시 가능성 가정을 만족하는 central DRS를 찾는 것은 중요한 문제이다. SDR 방법에 따라 다양한 차원축소 ATE 방법이 제시되었고, SDR을 이용한 ATE의 추정량은 모수모형을 이용한 이중강건 추정량보다 성능이 뛰어난 것으로 알려져있다.

3.3. 차원 선택

Ye와 Weiss (2003)가 제시한 Bootsrap 차원 선택 방법을 소개하고자 한다. 축소된 차원 d에 대해서 원래 데이터를 이용한 central DRS의 추정값을 $\widehat{\boldsymbol{B}}_d$, 새로 추출된 샘플을 이용하여 구한 central DRS의 추정값을 $\widehat{\boldsymbol{B}}_{\mathrm{d,b}}$

이라고 하자. Ye와 Weiss (2003)는 $\widehat{\boldsymbol{B}}_{d,b}$으로 span된 공간 $\widehat{\boldsymbol{B}}_{d,b}^{\mathrm{T}}X$와 $\widehat{\boldsymbol{B}}_{d}$으로 span된 공간 $\widehat{\boldsymbol{B}}_{d}^{\mathrm{T}}X$의 연관성을 계산하는 방법을 제안하였다.

$$\{\mathrm{var}\,(U)\}^{-\frac{1}{2}}\,\mathrm{cov}\,(U,V)\,\{\mathrm{var}\,(V)\}^{-1}\,\mathrm{cov}\,(V,U)\,\{\mathrm{var}\,(U)\}^{-\frac{1}{2}}\,.$$

위의 행렬의 양의 값을 갖는 eigenvalues를 $\lambda_1,\ldots,\lambda_d$라고 하면, squared trace correlation r^2는 다음과 같이 계산된다.

$$r^2\,(U,V) = d^{-1}\sum_{i=1}^{d}\lambda_i.$$

Bootstrap을 B번 반복하여, $r^2(U,V)$의 평균값 $\bar{r}_d^2$을 구할 수 있다.

$$\bar{r}_d^2 = \frac{1}{B}\sum_{b=1}^{B} r^2\left(\widehat{\boldsymbol{B}}_{d,b}^{\mathrm{T}}X, \widehat{\boldsymbol{B}}_{d}^{\mathrm{T}}X\right).$$

Ye와 Weiss (2003)는 가장 큰 $\bar{r}_d^2$을 갖는 d를 이용하는 것을 권장하였다. Dong과 Li (2010)과 Ma와 Zhu (2012)도 시뮬레이션을 통해 Ye와 Weiss (2003)이 제안한 차원 선택 방법으로 찾은 값이 실제로 서로 독립인 latent variable의 수와 일치하는 것을 보였다. 6장에서 데이터 분석시 이 방법을 이용하여 적합한 차원을 선택할 것이다.

4. 차원 축소 방법을 이용한 **ATE** 추정

이 장에서는 SDR를 이용한 ATE 추정에 대한 최근에 제시된 네 가지 연구를 소개한다.

4.1. Liu 등 (2018)

Liu 등 (2018)는 PS에 대하여 다음과 같은 준모수(semiparametric) 모형을 설정하였다.

$$P\,(T=1\mid X=x) = \frac{\exp\left\{\eta\left(\boldsymbol{B}^{\mathrm{T}}x\right)\right\}}{1+\exp\left\{\eta\left(\boldsymbol{B}^{\mathrm{T}}x\right)\right\}}, \tag{4.1}$$

여기서 $X\in\mathbb{R}^p$, $\boldsymbol{B}\in\mathbb{R}^{p\times d}$이고, η에 대해서는 특정 형태를 가정하지 않았다. Ma와 Zhu (2012)가 제시한 준모수 방법을 이용하여 X에 대한 central DRS $\boldsymbol{B}^{\mathrm{T}}X$를 구한 후 이 값을 이용하여 kernel estimation을 통해 η를 추정하는 것을 반복하여 PS를 구하였다. 준모수 방법으로 추정된 PS를 식 (3.1)에 대입하여, 즉 IPW를 이용하여 ATE를 추정하였다.

이 방법은 Ghasempour과 de Luna (2021)에 의해 R-패키지 `SDRcausal`로 구현되었고, `SDRcausal` 패키지는 `https://github.com/stat4reg` 에서 다운받을 수 있다. `ipw.ate` 함수와 `imp.var` 함수를 이용하여 각각 ATE와 ATE의 asymptotic variance를 추정할 수 있으나, 사용자가 $\boldsymbol{B}$의 초깃값을 설정해야 하며, $\boldsymbol{B}$의 초깃값에 민감한 결과를 도출하기도 한다. 초깃값 설정에 대한 연구가 필요할 것으로 생각된다.

4.2. Ma 등 (2019)

Ma 등 (2019)에서는 sparse SDR (SSDR)을 활용하여 PS과 OR을 추정하였다. 조건부 평균 $E(Y_i|X_i)$에 대해서는 공변량 $\boldsymbol{X}_i$의 d 선형 조합에 따라 달라지는 SSDR 모델을 고려하였다.

$$E\,(Y_i\mid X_i) = E\left(Y_i\mid \boldsymbol{B}^{\mathrm{T}}X_i\right) = E\left(Y_i\mid \boldsymbol{V}^{*\mathrm{T}}X_i\right).$$

이 때, $\boldsymbol{B}$과 $\boldsymbol{V}^*$는 $p \times d$인 행렬이다. $\boldsymbol{V}^*$의 추정치는 group Lasso penalized 방법 (Yuan과 Lin, 2006)에 의해 구해진다.

$$\boldsymbol{V}^* = \underset{\boldsymbol{V} \in \mathbb{R}^{p \times d}}{\operatorname{argmin}} \left(\frac{1}{2} \left\| \widetilde{\boldsymbol{W}} \boldsymbol{A}^* - \boldsymbol{X}\boldsymbol{V} \right\|^2 + \lambda \sum_{k=1}^{p} \|\boldsymbol{V}_k\| \right).$$

이 때, $\boldsymbol{X} = (X_1, \ldots, X_n)^{\mathrm{T}}$, $\widetilde{X}_i = (1, X_i^{\mathrm{T}})^{\mathrm{T}}$, $\widetilde{Y}_i = Y_i - E(Y_i)$, $\widetilde{W}_i = \widetilde{X}_i \widetilde{Y}_i$, $\widetilde{\boldsymbol{W}} = (\widetilde{W}_1, \ldots, \widetilde{W}_n)^{\mathrm{T}}$, $\boldsymbol{A}^{*\mathrm{T}} \boldsymbol{A}^* = \boldsymbol{I}$이고 λ는 tuning parameter이다.

임의의 변수 Z_i의 조건부 평균은 $\boldsymbol{V}^*$의 refitted unpenalized estimator $\widetilde{\boldsymbol{V}}$를 이용하여 kernel estimation을 통해 다음과 같이 추정할 수 있다.

$$\widehat{E}(Z_i \mid X_i = x) = \widehat{E}\left(Z_i \mid \widetilde{\boldsymbol{V}}^{\mathrm{T}} X_i = \widetilde{\boldsymbol{V}}^{\mathrm{T}} x\right) = \frac{\sum_{i=1}^{n} K_h\left(\widetilde{\boldsymbol{V}}^{\mathrm{T}} X_i - \widetilde{\boldsymbol{V}}^{\mathrm{T}} x\right) Z_i}{\sum_{i=1}^{n} K_h\left(\widetilde{\boldsymbol{V}}^{\mathrm{T}} X_i - \widetilde{\boldsymbol{V}}^{\mathrm{T}} x\right)}, \tag{4.2}$$

여기서 $\boldsymbol{u} = (u_1, \ldots, u_d)^{\mathrm{T}}$, 다변량 kernel density 함수 $K(u_1, \ldots, u_d)$이고 bandwidth 벡터 $\boldsymbol{h} = (h_1, \ldots, h_d)^{\mathrm{T}}$일 때, $K_h(\boldsymbol{u}) = h^{-d} K(u_1/h, \ldots, u_d/h)$이다. 식 (4.2)에 Z_i 대신 Y_{ji}를 대입하면, OR에 대한 모델링을 할 수 있다. 잠재적 결과는 두 가지 처리 중 한 가지 경우만 관측되기 때문에 처치집단 데이터로 $E(Y_{1i}|X_i = x)$, 통제집단 데이터로 $E(Y_{0i}|X_i = x)$를 추정한다. 이 때 위의 식 (4.2)에서 n 대신 각각 처치집단과 통제집단의 관측치 수로 대체해야 한다. 같은 방식으로, 식 (4.2)에 Z_i 대신 T_i를 대입하면 PS에 대한 추정량 $\widehat{\pi}(x) = \widehat{E}(T_i|x)$을 구할 수 있다. $\widehat{E}(Y_{ji}|X_i = x)(j = 0, 1)$와 $\widehat{\pi}(x)$를 식 (3.2)에 대입하여 ATE에 대한 이중 강건 추정량을 구한다.

4.3. Ghosh 등 (2021)

이 연구는 차원 축소에 대해 efficient semiparametric estimation을 이용한 방법으로 Liu 등 (2018) 방법과 차원 축소 방법은 유사하고, ATE 추정에 대해서는 개선된 방법을 제안한 연구이다. PS에 대해서는 Liu 등 (2018)에서 이용한 준모수 모형 (4.1)을 이용하여 추정한다.

$$\widehat{p}_i = \frac{\exp\left\{\widehat{\eta}\left(\widehat{\boldsymbol{B}}^{\mathrm{T}} x_i\right)\right\}}{\left[1 + \exp\left\{\widehat{\eta}\left(\widehat{\boldsymbol{B}}^{\mathrm{T}} x_i\right)\right\}\right]}.$$

OR에 대해서는 처방을 받을 때 ($T = 1$)에는 m_1으로, 처방을 받지 않을 때 ($T = 0$)에는 m_0로 모델링하였다.

$$Y_{1i} = m_1\left(\boldsymbol{B}_1^{\mathrm{T}} x_i\right) + \epsilon_1, \tag{4.3}$$

$$Y_{0i} = m_0\left(\boldsymbol{B}_0^{\mathrm{T}} x_i\right) + \epsilon_0, \tag{4.4}$$

여기서 $E(\epsilon_1|\boldsymbol{x}) = 0$, $E(\epsilon_0|\boldsymbol{x}) = 0$이며, $m_1(\cdot), m_0(\cdot)$에 대해서는 특정 형태를 가정하지 않는다. $\boldsymbol{B}_1, \boldsymbol{B}_0$는 각각 $p \times d_1$, $p \times d_0$인 벡터 또는 행렬이다 ($p > d_1, p > d_0$). (4.3)과 (4.4)에 대해서는 Ma와 Zhu (2014)의 준모수 방법을 이용하여 m_1, m_0, $\boldsymbol{B}_1$과 $\boldsymbol{B}_0$를 추정하였다. 잠재적 결과 중 하나의 값만 관측되므로, 만약 i가 처치집단에 속했다면, 통제집단에서 i는 결측값을 갖는다. x_i와 m_0, $\boldsymbol{B}_0$의 추정값을 이용하여 imputation 방법으로 Y_{0i}를 추정할 수 있다. 즉, 반사실 결과를 imputation 추정치로 대체하는 방식이다. ATE의 추정치 $\widehat{\tau}_{\mathrm{IMP}}$는 $T = 1$의 잠재적 결과의 기댓값과 $T = 0$의 잠재적 결과의 기댓값의 차이로 다음과 같이 계산된다.

$$\widehat{\tau}_{\mathrm{IMP}} = \frac{1}{n} \sum_{i=1}^{n} \left\{ T_i Y_i + (1 - T_i)\, \widehat{m}_1\left(\widehat{\boldsymbol{B}}_1^{\mathrm{T}} x_i\right) \right\} - \frac{1}{n} \sum_{i=1}^{n} \left\{ (1 - T_i)\, Y_i + T_i \widehat{m}_0\left(\widehat{\boldsymbol{B}}_0^{\mathrm{T}} x_i\right) \right\}. \tag{4.5}$$

AIPW를 이용한 ATE의 추정치 $\widehat{\tau}_{\text{AIPW}}$는 다음과 같다.

$$\widehat{\tau}_{\text{AIPW}} = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{T_i Y_i}{\widehat{p}_i} + \left(1-\frac{T_i}{\widehat{p}_i}\right)\widehat{m}_1\left(\widehat{\boldsymbol{B}}_1^{\mathrm{T}} x_i\right)\right\} - \frac{1}{n}\sum_{i=1}^{n}\left[\frac{(1-T_i)\,Y_i}{1-\widehat{p}_i} + \left\{1-\frac{1-T_i}{1-\widehat{p}_i}\right\}\widehat{m}_0\left(\widehat{\boldsymbol{B}}_0^{\mathrm{T}} x_i\right)\right]. \tag{4.6}$$

추가로, Robins 등 (1995)에서 제시한 결측 데이터 분석시 working covariance를 이용하는 방법을 이용하여 좀 더 개선된(improved) AIPW 추정치로써 τ_{IAIPW}를 제시하였다.

$$\widehat{\tau}_{\text{IAIPW}} = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{T_i Y_i}{\widehat{p}_i} + \widehat{\gamma}_1\left(1-\frac{T_i}{\widehat{p}_i}\right)\widehat{m}_1\left(\widehat{\boldsymbol{B}}_1^{\mathrm{T}} x_i\right)\right\} - \frac{1}{n}\sum_{i=1}^{n}\left[\frac{(1-T_i)\,Y_i}{1-\widehat{p}_i} + \widehat{\gamma}_0\left\{1-\frac{1-T_i}{1-\widehat{p}_i}\right\}\widehat{m}_0\left(\widehat{\boldsymbol{B}}_0^{\mathrm{T}} x_i\right)\right]. \tag{4.7}$$

이 때, $\widehat{\gamma}_1$과 $\widehat{\gamma}_0$는 다음과 같다.

$$\widehat{\gamma}_1 = \operatorname{cov}\left\{\widehat{m}_1\left(\widehat{\boldsymbol{B}}_1^{\mathrm{T}} x_i\right)\frac{T_i}{\widehat{p}_i}, \left(1-\frac{T_i}{\widehat{p}_i}\right)\widehat{m}_1\left(\widehat{\boldsymbol{B}}_1^{\mathrm{T}} x_i\right)\right\}^{-1} \operatorname{cov}\left\{\frac{T_i Y_i}{\widehat{p}_i}, \left(1-\frac{T_i}{\widehat{p}_i}\right)\widehat{m}_1\left(\widehat{\boldsymbol{B}}_1^{\mathrm{T}} x_i\right)\right\},$$

$$\widehat{\gamma}_0 = \operatorname{cov}\left[\frac{(1-T_i)}{(1-\widehat{p}_i)}\widehat{m}_0\left(\widehat{\boldsymbol{B}}_0^{\mathrm{T}} x_i\right), \left\{1-\frac{1-T_i}{1-\widehat{p}_i}\right\}\widehat{m}_0\left(\widehat{\boldsymbol{B}}_0^{\mathrm{T}} x_i\right)\right]^{-1} \operatorname{cov}\left[\frac{(1-T_i)\,Y_i}{1-\widehat{p}_i}, \left\{1-\frac{1-T_i}{1-\widehat{p}_i}\right\}\widehat{m}_0\left(\widehat{\boldsymbol{B}}_0^{\mathrm{T}} x_i\right)\right].$$

Mukherjee와 Chatterjee (2008)에 따라, unbiased estimator $\widehat{\tau}_{\text{IMP}}$와 efficient estimator $\widehat{\tau}_{\text{AIPW}}$의 가중평균치인 shrinkage estimator도 이용할 수 있다.

Ghosh 등 (2021)의 3.4절에는 ATE에 대한 asymptotic property를 증명하였다. $\widehat{\tau}_{\text{IMP}}$는 m_1와 m_0 중 적어도 하나가 mis-specified일 때 inconsistent이고, $\widehat{\tau}_{\text{AIPW}}$는 m_1와 m_0가 mis-specified일 때에도 consistent이다.

이 방법은 R-패키지 `SDRcausal`에 구현되어 있으며, 각각의 방법에 따라 다음과 같은 R 함수를 이용할 수 있다.

1. imputation 방법을 이용한 ATE의 추정치 $\widehat{\tau}_{\text{IMP}}$ (식 (4.5))

   추정치: `imp.ate`, 추정치의 분산: `vimp.ate`.

2. AIPW 추정치 $\widehat{\tau}_{\text{AIPW}}$ (식 (4.6))

   추정치: `aipw.ate`, 추정치의 분산: `aipw.var`.

3. IAIPW 추정치 $\widehat{\tau}_{\text{IAIPW}}$ (식 (4.7))

   추정치: `aipw2.ate`, IAIPW 추정치의 분산만 따로 구하는 R 함수는 구현되어 있지 않음.

`SDRcausal` 패키지의 `inf.ate` 함수를 이용하면 $\widehat{\tau}_{\text{IMP}}, \widehat{\tau}_{\text{AIPW}}, \widehat{\tau}_{\text{IAIPW}}$ 의 추정값 및 Liu 등 (2018)의 IPW 추정치 (3.1)와 각각의 추정치의 asymptotic variance를 추정할 수 있다. `SDRcausal` 패키지 이용시, 함수 이용시 설정해야 하는 초깃값에 따라 민감한 결과가 도출되기도 하고, 특히 차원이 커질수록 추정값이 수렴하지 않는 경우가 종종 발생한다. 따라서, 초깃값 설정시 주의가 필요하고, 추정값의 수렴 여부를 확인해야 한다.

4.4. Cheng 등 (2022)

Cheng 등 (2022)는 PS와 OR을 이용하지 않고, matching 방법을 이용하여 반사실 결과를 imputation하는 방법을 제시하였다. Cheng 등 (2022)는 그들이 제안한 연구 방법을 causal effect estimator by using sufficient dimension reduction (CESD)로 명명하였다. 우선, central DRS $\boldsymbol{B}$를 찾는다. SDR에 의해 X는 $Z = \boldsymbol{B}^{\mathrm{T}}X$ 와 $Q = (1-\boldsymbol{B}^{\mathrm{T}})X$로 분해할 수 있다. Z는 T와 연관되어 있고, Z를 조건부로 하면 처리 T는 X로부터 분해된 Q와 독립이다. 즉 $T \perp\!\!\!\perp Q|Z$이다. Cheng 등 (2022)의 Fig 1에서 제시한 Z, T와 Q의 관계를 표현한 그래프 모형을 참고하기 바란다. 이 때 Z는 Y에 대한 T의 평균 인과 효과를 추정하기 위한 적절한 deconfounding set이다.

데이터로부터 적절한 deconfounding set을 찾기 위해 kernel-based SDR 방법 (Aronszajn, 1950)을 활용하였다. 우선, reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ 상에서 cross-covariance operators를 찾는다. 다음과 같은 Gaussian kernel을 이용하였다.

$$k\left(x_i, x_j\right) = \exp\left(-\frac{\left\|x_i - x_j\right\|^2}{2\delta^2}\right),$$

여기서 δ는 bandwidth이다. 두 개의 RKHS $(\mathcal{H}_1, k_1)$과 $(\mathcal{H}_2, k_2)$가 주어졌을 때, 모든 $f \in \mathcal{H}_1$과, $g \in \mathcal{H}_2$에 대해서, $\mathcal{H}_1$부터 $\mathcal{H}_2$로의 cross-covariance operator는 다음과 같다.

$$\langle g, \Sigma_{TX} f\rangle_{\mathcal{H}_2} = E_{XT}\left[f(X)\, g(T)\right] - E_X\left[f(X)\right] E\left[g(T)\right].$$

$\mathcal{H}_1$에서 conditional covariance operator $\Sigma_{TT|Z}$는 다음과 같이 정의된다.

$$\Sigma_{TT|Z} \equiv \Sigma_{TT} - \Sigma_{TZ}\Sigma_{ZZ}^{-1}\Sigma_{ZT}.$$

Fukumizu 등 (2004)에 따르면, 모든 Z에 대하여 $\Sigma_{TT|Z} \geq \Sigma_{TT|X}$이고, $\Sigma_{TT|X} - \Sigma_{TT|Z} = 0 \iff T \perp\!\!\!\perp Q|Z$이다. 따라서, $\widehat{\Sigma}_{TT|Z}$의 determinant를 최소로 만드는 Z을 구하는 것이 central DRS을 구하는 것과 같은 문제가 된다. 다음과 같이 반복적으로 $\boldsymbol{B}$를 업데이트하여 추정한다.

$$\boldsymbol{B}^{s+1} = \boldsymbol{B}^s - \beta\frac{\partial \log\det\widehat{\Sigma}_{TT|Z}}{\partial \boldsymbol{B}} = \boldsymbol{B}^s - \beta Tr\left(\widehat{\Sigma}_{TT|Z}^{-1}\frac{\partial\widehat{\Sigma}_{TT|Z}}{\partial \boldsymbol{B}}\right),$$

여기서 β는 golden section search (Fukumizu 등, 2004)에 의해 구해진다. Deconfounding set Z를 구한 후 nearest neighbour matching (NNM) (Abadie와 Imbens, 2006; Rubin, 1973)과 같은 mathcing 방법을 이용하여 i 개체에 대한 반사실 결과 $Y_i^*(t_i)$를 추정한다. 이 때 i 개체가 받은 처리 $T_i = t_i$이라면, 반사실 결과는 처리가 $1 - t_i$일 때의 결과이다. i의 반사실 결과는 다음과 같이 공변량을 축소한 값 (Z)의 Mahalanobis distance를 최소로 하는 상대 집단에 있는 개체 k에 해당하는 결과 값으로 대체(imputation)한다.

$$\begin{aligned}\mathrm{Dist}\left(z_i, z_j\right) &= \left\{\left(z_i - z_j\right)^{\mathrm{T}}\widehat{\Sigma}_z^{-1}\right\}^{\frac{1}{2}},\\ Y_i^*(t_i) &= Y_k(1 - t_i), \quad k = \underset{j\in \boldsymbol{D}_{(1-t_i)}}{\operatorname{argmin}}\,\mathrm{Dist}\left(z_i, z_j\right),\end{aligned}$$

여기서 $\boldsymbol{D}_{(1-t_i)}$는 $1 - t_i$에 해당하는 데이터셋이다. 관측된 결과값과 추정된 반사실 결과값의 차이의 평균으로 ATE를 추정한다.

Central DRS를 구하는 방법은 R-패키지 `KDRcpp`로 구현되어 있고, `https://github.com/aschmu/KDRcpp`에서 다운받을 수 있다. 이 패키지는 초깃값 설정 필요 없이 고차원 데이터에서도 안정적으로 작동한다. `MatchIt` 패키지를 이용하여 matching을 할 수 있으며, `Zelig` 패키지를 이용하여 ATE에 대한 통계적 추론을 할 수 있다. `https://github.com/chengdb2016/CESD`을 참고하기 바란다.

5. ATE 추정량 해석시 유의할 점

2장에서 인과 추론의 두 가지 가정에 대해 다루었다. 잠재적 결과와 공변량에 동시에 영향을 줄 수 있는 교란 변수의 존재 여부를 데이터를 통해 검정하기 어렵기 때문에, 현재까지 진행된 연구로는 강한 무시 가능성 가정을 검정하는 것은 가능하지 않다. 이런 이유로 민감도 분석(sensitivity test)을 통해 가정이 위배될 때의 결과의 민감도를 파악하는 방법이 연구되고 있다.

Table 1: Variables of bike sharing dataset

Variables	Description
Season	1: winter, 2:spring, 3:summer, 4:fall
Year	0:2011, 1:2012
Workingday	If day is neither weekend nor holiday is 1, otherwise is 0.
Weathersit	1: clear/ few clouds/ partly cloudy/ partly cloudy 2: mist and cloudy/ mist and broken clouds/ mist and few clouds/ mist 3: light snow/ light rain, thunderstorm and scattered clouds/ light rain and scattered clouds
Temp	Normalized temperature in Celsius (only in hourly scale).
Atemp	Normalized feeling temperature in Celsius (only in hourly scale).
Hum	Normalized humidity. The values are divided to 100 (max).
Windspeed	Normalized wind speed. The values are divided to 67 (max).
Cnt	Count of total rental bikes including both casual and registered

Cheng 등 (2022)의 5.1절에서는 인과 추론에서 자주 이용되는 두 가지 합성 데이터를 Ghosh 등 (2021)의 shrinkage 방법, causal forest (Wager과 Athey, 2018) 방법, Cheng 등 (2022)의 CESD의 방법 등을 이용하여 분석하였다. Cheng 등 (2022)의 Fig 3과 Fig 4에 ATE의 참값과 제안된 방법들로 구한 추정량의 95% 신뢰구간이 표시되어 있는데, 어느 추정치도 참값과 일치성을 보이지 않는다. Cheng 등 (2022)는 CESD 방법으로 구한 추정량이 ATE의 참값과 가장 근사한 값임을 강조하고 있다. 하지만, Liu 등 (2018)의 4장, Ma 등 (2019)의 7장과 Ghosh 등 (2021)의 4장에서 시뮬레이션 결과는 각각의 이론과 일치하는 추정치를 갖는 것을 보였다. 시뮬레이션에서는 강한 무시 가능성 가정을 만족하도록 데이터를 발생할 수 있었으므로 일치성을 갖는 추정치를 얻는 것이 가능했을 것이다. 합성 데이터가 강한 무시 가능성 가정을 만족하지 않을 가능성을 배제할 수 없으며, 몇몇 방법의 경우에는 데이터 분석시 설정한 초깃값이 부적절했을 가능성도 있다.

Liu 등 (2018), Ma 등 (2019)와 Ghosh 등 (2021)의 방법은 PS와 OR에 근거한 방법이고, Cheng 등 (2022)의 CESD는 PS와 OR를 추정하지 않고 matching을 이용한 방법이다. 단순히 CESD의 matching에 근거한 방법이 PS와 OR에 근거한 방법보다 더 우월하다고 보기는 어렵다. 인과 추론은 실제로 일어나지 않은 상황을 관측 데이터를 통해 최대한 잘 설명하고자 하는 것을 목표로 하고 있기 때문에, 각각의 방법이 가정하고 있는 것을 파악하고 그 가정에 비추어 예상할 수 있는 결과를 적절하게 설명해야 한다. 2장에서 언급한 인과 추론의 가정을 만족하고, 주어진 데이터로부터 추정한 PS와 OR이 적절하다면, Liu 등 (2018), Ma 등 (2019)와 Ghosh 등 (2021)의 방법으로 구한 ATE의 추정값은 참값에 가까울 것이다. 이 추정량들 중에서 추정량의 분산이 상대적으로 작고, PS와 OR이 약간 잘못 추정될 때(mis-specified)에도 robust한 추정치로는 Ghosh 등 (2021)의 shrinkage 방법으로 구한 추정치를 들 수 있다. 하지만, Ghosh 등 (2021)에 PS와 OR이 잘못 추정된 정도에 따른 결과의 민감도는 구체적으로 연구되어 있지 않다. Matching을 하기에 충분한 데이터가 있고, 처치집단과 통제집단의 사람들이 최대한 비슷한 성향의 상대 집단의 사람들과 matching이 가능하다는 가정이 적절하다면 CESD의 방법으로 구한 ATE의 추정값은 실제 값과 가까울 것이다.

6. 데이터 분석

UCI machine learning repository 사이트에 공유된 bike sharing 데이터를 이용하여 (`https://archive.ics.uci.edu/ml/datasets/bike+sharing+dataset`) 포르투갈의 평일(working day)과 공휴일(weekend or holiday)에 따른 자전거 대여 수의 차이를 인과 추론 방법으로 분석하고자 한다. Table 1에 변수에 관한 설명을 기재하였다. 변수 workingday가 처리에 해당하고, 변수 cnt가 반응 변수에 해당한다. 참고로, UCI machine learning repository 사이트에서 weathersit은 1,2,3,4의 값을 갖는 것으로 설명되어 있었으나, 실제로 weather-

Table 2: Calculated $\bar{r}_d^2$ using bootstrap method

d	3	4	**5**	6
$\bar{r}_d^2$	0.821	0.755	**0.922**	0.891

Table 3: Estimation of ATE

Matching method	Estimate	s.e.	95% CI
Full	207.19	88.96	(32.84, 381.55)

sit = 4에 해당되는 데이터는 존재하지 않아 Table 1에 weathersit = 4에 대한 설명은 제외하였다. Season과 weathersit은 범주형 변수이므로, 다음과 같이 dummy variable로 변형하였다.

$$\text{season1} = \begin{cases} 1, & \text{if season} = 1, \\ 0, & \text{otherwise}, \end{cases} \quad \text{season2} = \begin{cases} 1, & \text{if season} = 2, \\ 0, & \text{otherwise}, \end{cases} \quad \text{season3} = \begin{cases} 1, & \text{if season} = 3, \\ 0, & \text{otherwise}, \end{cases}$$

$$\text{weathersit1} = \begin{cases} 1, & \text{if weathersit} = 1, \\ 0, & \text{otherwise}, \end{cases} \quad \text{weathersit2} = \begin{cases} 1, & \text{if weathersit} = 2, \\ 0, & \text{otherwise}. \end{cases}$$

범주형 변수 대신에 dummy variable을 이용하면 처리를 제외한 공변량은 10개가 된다. 이 10개의 공변량은 서로 연관성이 높다. 예를 들면, 일반적으로 기온 temp와 체감 온도 atemp는 비슷한 값을 갖고, 계절 season에 따라 기온 temp과 습도 hum가 특정 패턴을 갖게 된다. 차원 축소 방법을 이용하여 10개의 공변량을 적절하게 설명할 수 있는 차원 d를 선택하기 위하여 다음과 같은 절차를 따랐다. 우선, 차원 $d = 3, 4, 5, 6$에 대해서 4.4절에서 다루었던 Cheng 등 (2022)의 방법을 이용하여 central DRS $\boldsymbol{B}^{\mathrm{T}}X$를 추정한다. Ye와 Weiss (2003)이 제안한 Bootstrap 방법을 이용하여 가장 적합한 d를 선택할 수있다. Bootstrap 반복을 $B = 100$번 하여 계산한 결과를 Table 2에 기재하였다. $d = 5$일 때 가장 $\bar{r}_d^2$이 가장 크므로 10개의 공변량을 5차원으로 축소하여 분석을 진행하였다.

Cheng 등 (2022)의 방법을 이용하여 ATE에 대한 추정을 하고자 한다. Central DRS에 의해 구해진 adjusting set Z를 이용하여 matching을 한다. R-패키지 `MatchIt`의 `matchit` 함수에는 matching 방법에 대한 다양한 옵션을 제공하고 있다. 평일(처리집단)과 공휴일(통제집단)은 각각 500과 231개의 관측치를 갖고 있다. Mahalanobis distance에 의한 matching 방법은 "nearest", "optimal", "full"과 "genetic"이 있는데, 이 중 "nearest", "optimal"와 "genetic"를 이용할 경우 몇몇 관측치는 matching이 되지 않아, "full"을 이용하여 matching을 하였다. Table 3에 ATE의 추정값, 표준 오차와 95% 신뢰구간을 기재하였다. 이 값은 R-패키지 `Zelig`를 이용하여 얻은 값으로 계산 과정에서 시뮬레이션을 통해 구한 값이다. Seed를 지정하지 않으면 매번 약간씩 다른 값이 나온다. 강한 무시 가능성 가정을 만족하고, 처리집단과 통제집단의 관측치가 상대집단의 유사한 공변량을 갖는 값으로 matching이 잘되었다는 가정 하에서, 모든 환경이 랜덤하게 주어졌을 때 평일에는 공휴일보다 자전거 대여 건수가 207.19건 정도 많고 이 값은 5% 유의수준에서 통계적으로 유의하다.

7. 결론

지금까지 준모수 방법, SSDR 및 reproducing kernel Hilbert space 상에서 cross-covariance operators을 이용한 차원 축소 방법을 활용한 ATE 추정 방법에 대해 알아 보았다. 실험을 통해 처리의 효과를 알아볼 수 있다면 가장 이상적이지만, 대부분의 경우에는 실험을 할 수 없기 때문에 관측 데이터를 통해 처리의 효과를 추정해야 한다. 수집된 관측 데이터에서는 처치집단과 실험집단간에 균형을 이루는 경우가 거의 없으므로, 처치를 받을 확률인 PS를 계산하여 불균형한 데이터의 가중치를 조정하는 방법이 주로 이용되었다. PS만 이용한

IPW 방법에 의한 ATE의 추정치는 분산이 크기 때문에, 이를 보완한 여러 가지 방법이 제시되었다. 대표적인 추정방법으로는 이중 강건 추정량으로 불리는 AIPW 추정량이다. AIPW 추정량을 구하기 위해서는 PS와 OR에 대한 consistent한 추정치가 필요한데, 이 때 특정 모수 모형을 가정하는 것보다 좀 더 유연한 준모수 모형을 이용하는 것이 권장된다. 모든 공변량을 차원 축소 없이 이용하여 PS와 OR를 구하는 것은 간편하지만, 공변량이 고차원일 때에는 이런 방식으로 구한 PS와 OR를 ATE의 추정에 이용할 경우 편향이 커질 수 있다. 따라서 고차원 데이터를 효율적으로 차원 축소하는 것은 PS, OR 및 ATE를 추정하는 데 있어서 중요한 문제이다. 이 논문에서는 준모수 방법와 SSDR를 이용하여 차원 축소를 하여 PS와 OR을 구하는 연구를 소개하였다. 또한, PS와 OR를 이용하지 않고 고차원 데이터의 ATE를 추정하는 방법도 제시되었다. 고차원의 공변량을 reproducing kernel Hilbert space 상에서 cross-covariance operators을 이용하여 차원 축소를 한 후, deconfounding set을 구하고, 처리 집단과 통제집단의 deconfounding set과 상대집단의 유사한 deconfounding set을 찾아 matching하여 ATE를 추정하는 방법이다. 기존 논문을 살펴보면, 인과추론의 방법에 따라 ATE의 추정값이 차이가 많이 나기도 한다. 인과추론은 실제로 발생하지 않은 사건을 관측치를 통해 추론하는 연구이므로, 다소 강한 가정이 필요하다. 기본적으로 강한 무시 가능성 가정과 overlap으로 불리는 두 가정을 전제로하고, 인과 추론 방법에 따라 반응변수에 대한 모형을 가정하기도 하고, matching이 적절하다는 가정하에 분석을 하기도 한다. 따라서, 선택한 모형이 취하고 있는 가정을 고려하여 적절하게 해석을 해야 한다.

References

Abadie A and Imbens GW (2006). Large sample properties of matching estimators for average treatment effects, *Econometrica*, **74**, 235–267.

Aronszajn N (1950). Theory of reproducing kernels, *Transactions of the American Mathematical Society*, **68**, 337–404.

Carpenter JR, Kenward MG, and Vansteelandt S (2006). A comparison of multiple imputation and doubly robust estimation for analyses with missing data, *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **169**, 571–584.

Cheng D, Li J, Liu L, Le TD, Liu J, and Yu K (2022). Sufficient dimension reduction for average causal effect estimation, *Data Mining and Knowledge Discovery*, **36**, 1174–1196.

Cook RD (1996). Graphics for regressions with a binary response, *Journal of the American Statistical Association*, **91**, 983–992.

Cook RD (2009). *Regression Graphics: Ideas for Studying Regressions through Graphics*, John Wiley & Sons, New York.

Cook RD and Weisberg S (1991). Sliced inverse regression for dimension reduction: Comment, *Journal of the American Statistical Association*, **86**, 328–332.

De Luna X, Waernbaum I, and Richardson TS (2011). Covariate selection for the nonparametric estimation of an average treatment effect, *Biometrika*, **98**, 861–875.

Dong Y and Li B (2010). Dimension reduction for non-elliptically distributed predictors: Second-order methods, *Biometrika*, **97**, 279–294.

Fukumizu K, Bach FR, and Jordan MI (2004). Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces, *Journal of Machine Learning Research*, **5**, 73–99.

Ghasempour M and de Luna X (2021). SDRcausal: An R package for causal inference based on sufficient dimension reduction, Available from: *arXiv preprint arXiv:2105.02499*

Ghosh T, Ma Y, and De Luna X (2021). Sufficient dimension reduction for feasible and robust estimation of

average causal effect, *Statistica Sinica*, **31**, 821–842.

Glymour M, Pearl J, and Jewell NP (2016). *Causal Inference in Statistics: A Primer*, John Wiley & Sons, United Kingdom.

Glynn AN and Quinn KM (2010). An introduction to the augmented inverse propensity weighted estimator, *Political Analysis*, **18**, 36–56.

Hahn J (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects, *Econometrica*, **66**, 315–331.

Kang JD and Schafer JL (2007). Demystifying double robust-ness: A comparison of alternative strategies for estimating a population mean from incomplete data, *Statistical Science,* **22**, 523–539.

Li B and Dong Y (2009). Dimension reduction for nonelliptically distributed predictors, *The Annals of Statistics*, **37**, 1272–1298.

Li B and Wang S (2007). On directional regression for dimension reduction, *Journal of the American Statistical Association*, **102**, 997–1008.

Li KC (1991). Sliced inverse regression for dimension reduction, *Journal of the American Statistical Association*, **86**, 316–327.

Liu J, Ma Y, and Wang L (2018). An alternative robust estimator of average treatment effect in causal inference, *Biometrics,* **74**, 910–923.

Ma Y and Zhu L (2012). A semiparametric approach to dimension reduction, *Journal of the American Statistical Association,* **107**, 168–179.

Ma Y and Zhu L (2014). On estimation efficiency of the central mean subspace, *Journal of the Royal Statistical Society: Series B: Statistical Methodology*, **76**, 885–901.

Ma S, Zhu L, Zhang Z, Tsai CL, and Carroll RJ (2019). A robust and efficient approach to causal inference based on sparse sufficient dimension reduction, *Annals of Statistics*, **47**, 1505–1535.

Mukherjee B and Chatterjee N (2008). Exploiting gene-environment independence for analysis of case–control studies: An empirical Bayes-type shrinkage estimator to trade-off between bias and efficiency, *Biometrics*, **64**, 685–694.

Robins JM, Rotnitzky A, and Zhao LP (1994). Estimation of regression coefficients when some regressors are not always observed, *Journal of the American Statistical Association*, **89**, 846–866.

Robins JM, Rotnitzky A, and Zhao LP (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data, *Journal of the American Statistical Association*, **90**, 106–121.

Rubin DB (1973). Matching to remove bias in observational studies, *Biometrics*, **29**, 159–183.

Vansteelandt S, Bekaert M, and Claeskens G (2012). On model selection and model misspecification in causal inference, *Statistical Methods in Medical Research*, **21**, 7–30.

Wager S and Athey S (2018). Estimation and inference of heterogeneous treatment effects using random forests, *Journal of the American Statistical Association*, **113**, 1228–1242.

Ye Z and Weiss RE (2003). Using the bootstrap to select one of a new class of dimension reduction methods, *Journal of the American Statistical Association*, **98**, 968–979.

Yuan M and Lin Y (2006). Model selection and estimation in regression with grouped variables, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **68**, 49–67.

Received February 24, 2023; Revised May 6, 2023; Accepted May 8, 2023

차원축소 방법을 이용한 평균처리효과 추정에 대한 개요

김미정[1,a]

[a]이화여자대학교 통계학과

요 약

고차원 데이터의 인과 추론에서 고차원 공변량의 차원을 축소하고 적절히 변형하여 처리와 잠재 결과에 영향을 줄 수 있는 교란을 통제하는 것은 중요한 문제이다. 평균 처리 효과(average treatment effect; ATE) 추정에 있어서, 성향점수와 결과 모형 추정을 이용한 확장된 역확률 가중치 방법이 주로 사용된다. 고차원 데이터의 분석시 모든 공변량을 포함한 모수 모형을 이용하여 성향 점수와 결과 모형 추정을 할 경우, ATE 추정량이 일치성을 갖지 않거나 추정량의 분산이 큰 값을 가질 수 있다. 이런 이유로 고차원 데이터에 대한 적절한 차원 축소 방법과 준모수 모형을 이용한 ATE 방법이 주목 받고 있다. 이와 관련된 연구로는 차원 축소 부분에 준모수 모형과 희소 충분 차원 축소 방법을 활용한 연구가 있다. 최근에는 성향점수와 결과 모형을 추정하지 않고, 차원 축소 후 매칭을 활용한 ATE 추정 방법도 제시되었다. 고차원 데이터의 ATE 추정 방법 연구 중 최근에 제시된 네 가지 연구에 대해 소개하고, 추정치 해석시 유의할 점에 대하여 논하기로 한다.

주요용어: 성향점수, 역확률 가중치, 차원 축소, 평균처리효과, 확장 역확률 가중치

이 논문은 연구재단 연구 과제 NRF-2020R1F1A1A01074157에 의하여 수행되었음.

[1](03760) 서울시 서대문구 이화여대길 52, 이화여자대학교 통계학과. E-mail: m.kim@ewha.ac.kr