한국경제포럼 제17권 제1호 47-72
DOI : 10.22841/kefdoi.2024.17.1.003

일자리사업의 향후 평가 방향: 인과적 머신러닝을 적용한 직업훈련사업의 평가*

김 용 성**

논문초록

본 연구는 최근 노동시장 분석에 활용되는 인과적 머신러닝 기법을 소개하고 우리나라 직업훈련 데이터에 적용한 결과를 제시하였다. 직업 훈련사업의 취업 효과를 분석한 다수의 연구 결과가 일치된 결론에 이르지 못하고 있다. 인과적 추론을 위해 소개하는 이중 머신러닝(double machine learning, DML)은 모형의 변수들 사이의 유연한(flexible) 관계를 허용함으로써 연구의 자의성을 줄일 수 있다. 향후 DML의 탈편의성과 강건성은 합리적인 효과의 범위를 제시하고, 이질적(heterogeneous) 처치효과의 추정은 평가 결과의 활용도를 높일 것으로 기대된다.

핵심 주제어: 인과적 머신러닝, 처치효과, 프로그램 평가
경제학문헌목록 주제분류: C18, C21, J08

투고 일자: 2024. 1. 2. 심사 및 수정 일자: 2024. 2. 5. 게재 확정 일자: 2024. 2. 15.

* 이 논문은 2021년도 한국기술교육대학교 교수 교육연구진흥과제 지원에 의하여 연구되었음. 본 연구는 한국고용정보원 2023년 재정지원 일자리사업 총괄평가 연구를 바탕으로 작성되었음.

** 한국기술교육대학교 테크노인력개발 전문대학원 교수, e-mail: yongkim65@koreatech.ac.kr

Ⅰ. 서 론

정부의 재정지원 일자리 사업의 예산은 2011년 약 8조8천억 원에서 2022년에는 약 31조1천억 원으로 급속히 확대되었으며, 직업훈련 예산도 같은 기간 약 1.1조 원에서 약 2.5조 원으로 2배 이상 증가하였다.[1] 정부의 일자리 사업의 규모가 커지면서 효과성에 대한 검토의 필요성이 제기되었고, 학계 및 연구기관 주도하에 많은 평가작업이 이루어졌다.

직업훈련의 효과에 대한 비교적 최근 국내 연구를 중심으로 살펴보면, 직업훈련이 취업(전직 또는 재취업 포함)에 긍정적인 효과를 가진다는 연구에서 통계적으로 뚜렷한 효과를 확인할 수 없다는 연구와 부정적 효과를 주장한 연구에 이르기까지 다양한 결과가 보고되었다. 한편 정부의 많은 재원이 투입되는 훈련 분야를 포함한 재정지원 일자리사업에 대해 정부와 다수의 연구기관이 성과지표를 바탕으로 사업의 효과를 살펴보고 있는데, 같거나 유사한 사업을 정기적으로 분석하는 평가 결과도 비교가능성과 타당성 확보에 어려움을 겪는 것으로 알려져 있다.

훈련사업의 효과를 연구한 결과가 엇갈리는 이유를 몇 가지로 나누어볼 수 있다. 첫째, 유사·동일한 디자인의 사업이더라도 훈련 참여자(교·강사 훈련생)의 속성, 훈련의 전달방식과 환경(때와 장소, 노동시장 상황) 등에서 서로 다르며, 각 사업의 특수성 차이는 상이한 결과로 나타나는 원인이 된다. 외국의 연구에서도 훈련사업 고유의 성격에 따라 효과의 차이가 크다는 사실을 밝히고 있다(Martin, 1998). 둘째, 자료가 가진 성격과 한계도 일관성 있게 훈련 효과를 파악하기 어렵게 하는 원인이 된다. 훈련 참여가 내생적으로 결정된 관측 자료(observational data)는 훈련 효과의 추정을 힘들게 하는 원인이 되는데, 실제 최근 연구에서 내생성을 고려하는 방법에 따라 훈련 효과가 민감하게 변하는 것이 확인된다(김용성, 2020). 셋째, 연구자에 따라 분석 결과가 영향을 받는다. 일반적으로 연구자는 선험적 가정에 따라 평가모형을 설정하고, 주어진 자료에 기초하여 변수를 선택한 후 특정한 분석 방법을 적용하는데, 이 모든 과정은 연구자의 주관에 크게 영향을 받는다. 가정은 얼마나 타당한지, 왜 평가모형이 연구자가 제시한 형태이어야만 하는지, 변수를 선택한 객관적 기준은 무엇이며 왜 다른 변수는 제외되었는지, 분석 방법은 실제 데이터가 생성되는 과정(data generating process)에 적합한 것인지 등의 많은 물음에 대부분의 연구는

1) 장신철 외(2021)의 표 2-4 참조.

설득력 있고 명쾌한 답을 제대로 주지 못한다. 결국 연구자 주관에 기반한 훈련 효과는 일치된 의견을 얻기 어려워 평가 결과가 사업의 추진 여부에 관한 판단과 개선 방향이라는 정책적 필요에 제한적으로 이바지하는 데 그치게 된다.

본 연구의 목적은 연구자의 주관을 최대한 줄이면서 사업의 효과를 평가하는 방법을 소개하는 데 있다. 본 연구가 다루는 인과적 머신러닝(Causal Machine Learning, CML)은 기존의 평가분석 방법과 크게 세 가지 점에서 차이가 있다. 첫째, 평가모형에 있어 매우 유연한(flexible) 함수 형태에 기초한 인과적 추론을 시도한다. 이론이나 관례에 따라 성과변수(Y)와 설명변수(X)의 구체적인 함수를 상정하는 대신 CML은 데이터에 기반(data-driven)하여 변수들 사이의 관계에 제약을 두지 않는 접근 방식을 채택하게 된다. 그 결과 CML은 모형 설정의 오류(model specification error)로부터 비교적 자유롭다는 장점이 있다. 둘째, 성과변수(Y)와 설명변수(X)의 관계가 추정(estimation)이 아닌 알고리즘의 조정(algorithm tuning)을 통해 이루어지면서 매우 정밀하게 근사(approximation)된다. 이는 인과관계 추론에 필요한 예측(prediction) 과정에서 강점을 가짐을 뜻한다(Athey, 2019). 셋째, CML은 프로그램의 평균 처치효과(Average Treatment Effect, ATE)와 함께 하위집단별 처치효과(Group Average Treatment Effect, GATE)와 조건부 처치효과(Conditional Average Treatment Effect, CATE)를 식별함으로써, 정책집행에 유용한 정보를 제공할 수 있다.[2] 전체 훈련대상의 평균적 효과인 ATE는 사업 전반의 성과를 가늠해 볼 때 중요한 지표이지만, 사업이 어떤 집단에 효과가 큰지, 또는 참여자의 특성에 따라 효과는 어떻게 달라지는지를 파악하는 데 필요한 정보를 제공하지 못한다. 따라서 참여자의 특성에 따른 이질적 처치효과(treatment effect heterogeneity)를 바탕으로 GATE와 CATE를 추정하는 작업은 평가 결과의 현실성과 정책 활용도를 높이는 데 도움이 된다.

본 연구의 구성과 내용을 간략히 요약하면 다음과 같다. 제Ⅱ장에서는 직업훈련 효과에 대한 국내·외 연구를 소개하였다. 제Ⅲ장에서는 널리 사용되는 프로그램 평가 방법을 정리하고, 두 가지 측면에서 비판적으로 검토하였다. 첫째, 정책의 효과를 측정하기 위해 알려지지 않은 성과변수와 관련 변수의 관계를 구체적인 함수 형태로 설정하면서 모형 설정의 오류로 인해 추정치가 취약할 수 있다는 점과 둘째, 정책대상

2) 물론 회귀분석에 기반을 둔 전통적인 평가방법도 GATE 또는 CATE를 파악할 수 있다. 다만 후술하는 Ⅲ장에서 밝히듯, 회귀분석 기반의 평가방법의 GATE와 CATE는 몇 가지 문제점을 가지고 있다.

과 개별 특성에 따라 다르게 나타나는 정책효과의 이질성을 제대로 반영하지 못하는 데 따른 문제점을 지적하였다. 본 연구의 핵심인 제Ⅳ장에서는 인과관계 추론을 위한 방법을 소개하고 성격(properties)을 살펴보았다. 경제학, 특히 노동시장 분석에 머신러닝을 적용한 연구가 활발하게 진행되었으며, 최근에는 단순한 상관분석과 예측을 넘어 인과적 추론에 머신러닝을 적극적으로 적용하는 방향으로 연구가 나아가고 있다.[3] 본 연구에서 소개하는 인과적 추론을 위해 제안된 '이중 머신러닝(Double machine learning, DML)'은 변수 간 유연한 형태의 관계를 기반으로 한 추정법으로서 탈편의성, 강건성 등 바람직한 성격을 가지며, 또한 처치효과의 이질성을 파악하는 데도 용이한 것으로 알려져 있다.

제Ⅴ장에서는 DML을 적용한 예시로 우리나라 직업훈련 데이터를 분석한 후 선행연구의 추정과 DML로 추정한 훈련 효과(ATE)를 비교하고, 나아가 GATE와 CATE 결과를 제시하였다. 최신의 자료와 새로운 방법을 적용하여 훈련 효과를 엄밀하게 추정하는 것도 의미 있는 작업이지만, 이 경우 연구 결과의 차이가 분석자료의 차이인지 혹은 방법론상의 차이인지를 분간하기 힘들어진다. 따라서 DML과 기존의 방법을 비교하려는 목적에 충실하기 위해 제Ⅴ장은 의도적으로 동일한 자료에 유사한 변수를 사용하는 추정 환경하에서 도출되는 결과를 비교하였다. 끝으로 제 Ⅵ장에서는 본 연구의 결과를 요약·정리하였다.

Ⅱ. 직업훈련의 효과에 관한 선행연구 소개[4]

직업훈련 프로그램은 '적극적 노동시장정책(Active Labor Market Policy, ALMP)'의 중요한 수단의 하나이다. OECD 국가의 추이를 보면 1980년대 후반 전체 ALMP 지출의 약 30%를 차지하던 직업훈련 지출은 2010~2018년에는 25.9% 수준까지 하락하였는데,[5] 직업훈련 사업이 다른 ALMP 수단과 비교해 비용이 많이 들고, 단기적 효과가 불확실하여 많은 국가가 직업훈련 분야에 소극적이었던 결과로 보인다.

직업훈련의 효과에 대해서 국·내외 많은 연구가 있었으나 뚜렷한 결론에 이르지 못하고 있다. 우선 외국의 연구를 보면, 훈련이 취업에 미치는 효과를 확인할 수 없

3) https://www.youtube.com/watch?v=L72E08QsyMc&list=PLLyRS3U0l8vHse0Nz-Bo9_vlDnsEwSBw2&index=1

4) 본 장은 김용성(2020)의 일부를 참조하여 작성하였다.

5) OECD https://stats.oecd.org/의 Labour market programmes

다는 연구(Koning and Peers, 2007)부터, 효과가 있더라도 매우 작을 것이라는 결과(Koning, 2007)가 있는 반면, 약간(modest)의 효과(Kluve, 2010), 또는 중·장기적으로 의미 있는 긍정적인 효과(Card et al., 2018; Vooren et al., 2019)를 주장한 연구가 있다. 미시자료에 평가기법(program evaluation)을 적용한 연구의 결과도 엇갈리고 있다. 표본 매칭을 적용해 프랑스 사례를 분석한 연구(Cavaco et al., 2013)와 루마니아의 직업훈련을 살펴본 연구(Rodriguez-Planas and Jacob, 2010)는 직업훈련 프로그램 미참여자에 비해 참여자의 취업확률이 높았음을 보고하였으며, Doerr(2022)는 독일의 훈련바우처제도를 통해 직업훈련에 참여한 경력단절 여성의 취업확률이 참여하지 않은 비교집단에 비해 현저히 높게 나타남을 밝히고 있으나, Nivorozhkin and Nivorozhkin(2007)은 훈련분야와 사업에 따라 프로그램의 효과가 일의적이지 않음을 주장하였다. 무작위 실험(random experiment)을 통해 분석에서도 직업훈련이 취업에 미치는 긍정적인 영향을 확인할 수 없었다는 연구(Hirshleifer et al., 2016)와 오히려 직업훈련이 취업에 부정적이라는 결과(Rosholm and Skipper, 2009)와 약간의 긍정적 효과가 확인된다는 연구(Yeyati et al., 2019)가 있다.

국내 연구 결과도 크게 다르지 않아, 직업훈련이 취업(전직 또는 재취업 포함)에 긍정적인 효과를 가진다는 연구(이병희, 2000; 유경준·이철인, 2008; 강순희 외, 2015)와 통계적으로 뚜렷한 효과를 확인할 수 없다는 연구(김용성·박우람, 2015) 나아가 오히려 부정적인 효과를 가진다는 주장(조준모 외 2010; 류기락 외 2014; 김용성, 2020)이 혼재하고 있다. 대체로 동일한 프로그램에 대해 정기적으로 일자리 사업을 모니터링 하는 한국노동연구원과 한국고용정보원의 작업에서도 평가 결과의 비교가능성과 일관성 확보에 어려움을 겪는 것으로 알려져 있다.[6] 몇몇 최근의 연구를 소개하면, Lee et al. (2019)은 OECD PIAAC(국제성인역량조사) 자료를 통해 분석한 결과 우리나라 직업훈련이 소득과 취업확률에 있어 특히 고령층에 대해 긍정적인 효과가 있음을 밝혔으며, 양용현 외(2019)는 고용보험자료를 통해 재직자 직업훈련은 훈련 이후 근로를 지속하는 확률이 높이는 효과가 있음을 주장하였다.[7]

---

6) 한국노동연구원은 '재정사업 고용영향평가' 사업의 테두리에서 일자리 사업의 효과를 살펴보고 있으며, 한국고용정보원은 매년 성과지표 중심의 '일자리사업 평가'를 실시하고 있다.

7) 훈련의 임금과 소득 효과는 유경준·강창희(2010), 금재호(2016), 오호영(2021), 김보배(2023)를 참조할 수 있다.

Ⅲ. 기존의 평가분석 방법에 대한 비판적 검토

1. 기존의 평가분석 방법론의 문제점

프로그램 참여자의 미시자료 분석을 이용해 사업의 효과를 측정하는 평가분석 방법은 크게 무작위 통제 시험법(randomized controlled trial, τ_{Rc}), 균형화법(balancing, τ_{Ipw}), 회귀분석 기반 추정법(regression-based, τ_{Rg}), 그리고 균형화와 회귀분석을 결합한 이중 강건 추정법(doubly-robust, τ_{Dr})으로 구분된다.[8] 아래 식 (1)에서 식 (4)는 성과변수를 Y, 특성 변수(covariates)를 X. 정책 참여 여부를 나타내는 더미변수를 $D \in \{0,1\}$이라 할 때, 각 추정법에 따른 ATE의 계산식을 보여준다.

$$\tau_{Rc} = \frac{1}{n}\sum [DY - (1-D)Y] \tag{1}$$

$$\tau_{Ipw} = \frac{1}{n}\sum \left[\frac{DY}{\pi(X)} - \frac{(1-D)Y}{1-\pi(X)}\right], \quad \pi(X) = E[D=1 \mid X] \tag{2}$$

$$\tau_{Re} = E[Y \mid X, D=1] - E[Y \mid X, D=0], \quad E[Y \mid X, D] = X\beta + \tau D \tag{3}$$

$$\tau_{Dr} = \frac{1}{n}\sum [\frac{DY - \{D-\pi(X)\}E[Y \mid X,D]}{\pi(X)} - \frac{(1-D)Y - \{D-\pi(X)\}E[Y \mid X,D]}{1-\pi(X)}] \tag{4}$$

식 (1)의 τ_{Rc}는 Y와 D만 의존하여 처치효과를 추정하는 직접적인 방법으로서, 무작위 배정을 통해서 사업의 성과와 참여가 관측 불가능한 변수(selection on unobservables)에 의한 영향을 배제함으로써 인과적 추론의 오류를 제거할 수 있는 장점이 있다. 반면 그 밖의 식 (2)-식 (4)의 추정법은 데이터와 관련하여 발생하는 문제(예를 들어, 표본 선택 편의, 확률적 독립성의 위배 등)를 교정하기 위해 추정 시에 '성향점수(propensity score)' 함수인 $\pi(X) = E[D=1 \mid X]$나 '성과 회귀식(outcome regression)'인 $E[Y \mid D, X]$의 구체적 형태를 가정한다.[9] 하지만 사전적

8) 대표적인 균형화 추정법으로는 성향점수 매칭, K-근거리(nearest) 매칭 등이 있으며, 회귀분석 기반 추정법으로는 회귀단절모형, 이중차분법 등이 있다.

9) $\pi(X)$의 형태로 로지스틱 함수가 널리 활용되며, 성과 회귀식의 대표적인 예로는 선형(linear)이 있다.

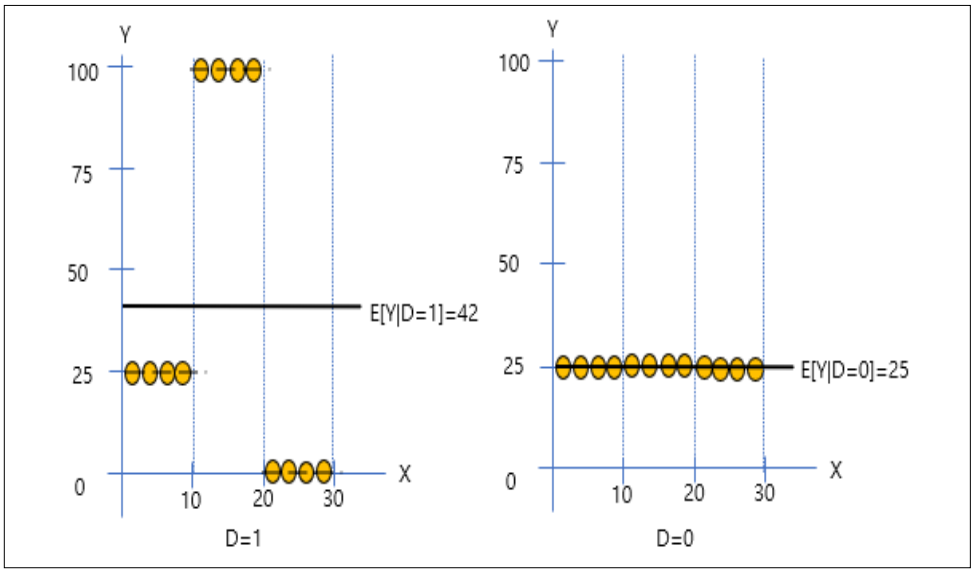
으로 알려지지 않은 D와 X의 관계인 $\pi(X)$와 Y와 D, X의 관계인 $E[Y \mid D, X]$를 특정한 형태의 함수로 대체하는 과정에서 발생하는 '설정의 오류(misspecification error)'는 인과적 추론의 결과에 신뢰성을 취약하게 만드는 원인이 된다.

2. 처치효과 동질성(treatment effect homogeneity)의 문제점

식 (1)에서 식 (4)의 ATE는 기본적으로 처치집단($D=1$)과 통제집단($D=0$) 사이의 Y의 평균값 차이로 구해지는데, 〈그림 1〉은 ATE의 산출을 시각적으로 보여준다. 먼저 처치집단의 효과를 보여주는 왼쪽 그림과 통제집단의 오른쪽 그림을 보면, $ATE = E[Y \mid D=1] - E[Y \mid D=0] = 42 - 25 = 17 > 0$이 되며, 이를 바탕으로 '프로그램은 효과가 있다.'라는 평가를 할 수 있다.

하지만 X 값에 따른 프로그램 효과를 보면, $X < 10$일 때 ATE=0이며, $10 \leq X < 20$일 때 ATE=75, $20 \leq X$에서는 ATE=-25로서 X 값의 크기에 따라 평균 처치효과의 크기가 다름을 알 수 있다. 처치효과의 이질성을 고려할 때 프로그램의 효과는 $10 \leq X < 20$에 속한 대상자에 대해서만 확인될 뿐, 그 밖에는 효과가 없거나 오히려 부정적이라 할 수 있다. 〈그림 1〉과 같이 세분된 대상의 프로그램 효과는 정책의 실효성을 높이는 데 유용한 정보가 된다.

〈그림 1〉 처치효과의 동질성과 이질성



처치효과의 이질성을 고려하기 위해 i라는 개인의 처치효과(Individual Treatment Effect, ITE)를 τ_i라고 할 때, $X=x$인 개인들의 조건부 평균 처치효과(CATE)는 $\tau(x)=E[\tau_i \mid X=x]$가 된다. 이때 $\hat{\tau}(x)$를 CATE의 추정치라고 하면, $\tau(x)$와 $\hat{\tau}(x)$의 평균 제곱 오차(mean squared error)는 아래와 같이 표현된다.

$$E\left[\left\{\tau_i-\hat{\tau}(x)\right\}^2 \mid X=x\right]=\underbrace{E\left[\left\{\tau_i-\tau(x)\right\}^2 \mid X=x\right]}_{A} + \underbrace{E\left[\left\{\tau(x)-\hat{\tau}(x)\right\}^2 \mid X=x\right]}_{B} \quad (5)$$

식 (5)의 오른쪽에서 A는 $X=x$인 ITE의 조건부 분산(conditional variance)으로서, $X=x$인 개인들의 ITE가 어떤 모습의 분포를 보이는지에 따라 크기가 결정된다. 만약 X가 적절히 설정된다면(well-defined), 특성변수 $X=x$인 개인의 ITE 차이는 크지 않을 것이므로 A의 값은 작아지게 된다. 결국 A의 크기를 줄이는 것은 어떻게 X 선택함으로써 $X=x$인 개인의 ITE 차이를 최소화하는가의 문제로 귀결된다.

식 (5)의 오른쪽에서 B는 CATE($\tau(x)$) 추정을 위해 선택된 모형($\hat{\tau}(x)$)이 얼마나 적합하고 우수한지에 따라 크기가 결정된다. 따라서 B의 크기를 줄이는 것은 평가모형 선택의 문제로 귀결된다. 기존의 평가분석법에서 $\hat{\tau}(x)$는 식 (2)-식(4)의 조건부 형태를 통해 구해지는데, 조건부 평균을 구하는 과정에서 D와 X의 관계 또는 Y와 X의 관계에 의존할 수밖에 없어, ATE와 같이 '설정 오류'의 가능성에서 벗어날 수 없다. 이와는 달리 후술하는 인과적 머신러닝은 자료에 기반하여 유연한 형태의 변수들 사이의 관계를 허용한다는 점에서 CATE의 산출에 있어 장점이 있다.

Ⅳ. 인과적 머신러닝: 문헌 검토 및 소개

1. 머신러닝을 적용한 기존 연구

머신러닝을 적용한 연구가 최근 활발히 진행되고 있다. 2012년부터 2018년까지 약 7백만 개 오스트레일리아의 일자리 자료를 머신러닝(Xgboost)을 적용한 분석을 통해 숙련 부족(skill shortage)의 정도를 살펴본 연구(Dawson et al., 2020)와 실업자 본인과 사례관리자(caseworker)가 평가한 장기실업으로 이어질 확률을 머신러닝

(Random Forest)으로 예측·비교한 연구(van den Berg et al., 2023)가 있다.[10]

인과적 추론에 머신러닝을 적용한 사례로서 인도네시아의 의료보험 혜택이 출산에 미치는 영향을 다양한 머신러닝 방법을 통해 분석한 연구(Krief and DiazOrdaz, 2019), 스위스의 고용서비스 프로그램 자료에 머신러닝(LASSO)을 적용하여 직업탐색 프로그램(job search program)의 이질적 취업 효과를 분석한 연구(Knaus et al., 2020)가 있다.[11]

Cengiz et al. (2021)은 머신러닝을 이용해 최저임금에 노출될 확률이 높은 집단('high-probability' group)을 식별하고 이들에게 최저임금이 미친 영향을 살펴보았으며,[12] Goller et al. (2021)은 인과적 머신러닝 방법을 적용하여 직업훈련(job training), 취업 걸림돌 완화(reducing impedients), 알선·취업 서비스(placement service)가 장기실업자의 취업에 미치는 효과를 분석하였다.[13] 특히 해당 연구는 수정된 인과적 포레스트(Modified Causal Forest, MCF)를 사용하여 참여자 특성에 따른 효과성의 이질성을 파악하고, 프로그램의 실효성을 높이기 위한 데이터 기반(data driven) 최적 적격성 조건(optimal eligibility condition)을 제안하였다.

2. 이중 머신러닝(Double Machine Learning, DML)

본 절에서는 구체적인 함수 형태의 가정에서 자유로운(flexible) 평가분석 방법으로서 인과적 머신러닝의 하나인 '이중 머신러닝(Double Machine Learning, 이하 DML)'을 소개하고 DML 추정의 성격(properties of the DML estimator)에 대해 간략히 논의하고자 한다(Chernozhukov et al., 2018; Ahrens et al., 2023).

성과변수를 Y, 처치(treatment) 여부를 나타내는 더미변수(binary)를 $D(\in 0,1)$, 독립변수(covariates, 또는 features)를 $X=(X_1,X_2,\cdots,X_k)$, U와 V를 교란항(disturbances)이라 할 때, 변수들(Y, D, X, U, V)의 관계가 다음과 같이 표현된다고 하자.[14]

---

10) 두 연구는 머신러닝을 인과관계의 틀에서 적용한 분석은 아니다.

11) 직업탐색의 효과가 단기적으로 이질성(treatment effect heterogeneity)이 뚜렷하나, 장기적으로는 다소 동질적인(homogenous) 모습으로 나타남을 밝히고 있다.

12) 전체 최저임금근로자를 L, 각 하위 집단(sub-group)의 규모와 집단에 속한 최저임금 근로자를 $S_1,\cdots S_G$, $L_1,\cdots L_G$라 할 때, high-probability는 L_g/S_g를 기준으로 구분된다.

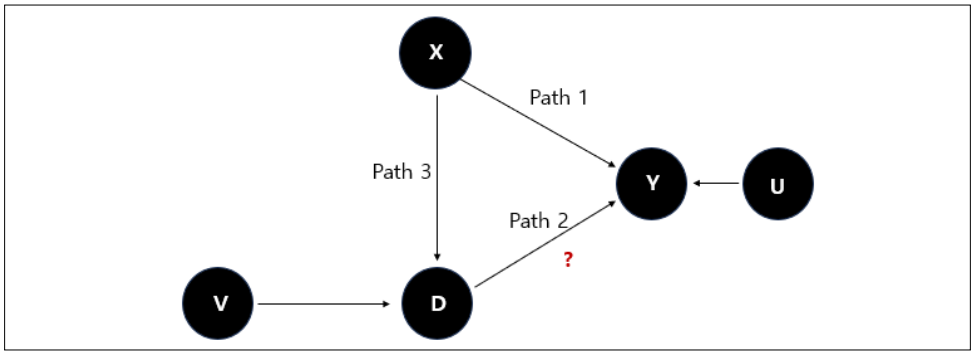
13) 연구는 알선 서비스가 장기실업자 취업에 가장 큰 도움이 되며, 그 효과가 가장 신속하게 일어나고 또한 오랫동안 지속된다는 점을 주장하였다.

$$Y = g_0(D,X) + U, \quad E[U \mid X, D] = 0 \tag{6}$$

$$D = m_0(X) + V, \quad E[V \mid X] = 0 \tag{7}$$

모형의 단순화를 위해 $g_0(D,X) = \theta_0 D + g_0(X)$라고 하자. 이때 D가 Y에 미치는 인과적 영향은 θ_0에 의해 포착되는데, θ_0를 정확히 추정하기 어려운 이유는 '방해함수(nuisance function)' m_0와 g_0에 혼재 요인(confounding factor) X가 존재하기 때문이다. 〈그림 2〉을 보면 'path 2'는 D가 Y에 직접 영향을 주는 통로가 되기도 하지만, 동시에 X가 D를 거쳐 Y에 영향을 주는 간접 통로(path 3→path 2)도 된다. 따라서 혼재 요인 X가 존재할 때, θ_0는 D와 Y의 '인과 효과(causal effect)'뿐만 아니라 X가 Y에 미치는 '중계 효과(mediation effect)'까지 포함하게 된다.

〈그림 2〉 혼재 요인(X)의 존재와 인과적 추론의 어려움



DML의 θ_0를 식별하기 위한 전략은 다음과 같이 직관적이며 간단하다. 첫째, $g_0(X) = E[Y \mid X]$와 $m_0(X) = E[D \mid X]$를 머신러닝을 이용해 $\hat{g}_0(X)$와 $\hat{m}_0(X)$로 예측하고, $\hat{U} = Y - \hat{g}_0(X)$, $\hat{V} = D - \hat{m}_0(X)$으로 식 (6)과 식 (7)을 잔차화(residualized) 한다.[15] 둘째, 두 식을 잔차화한 후, '잔차항에 대한 잔차항의 회귀분석(regressing the residual on the residuals)'인 아래 식 (8)에 따라 θ_0의 추정치 $\hat{\theta}_0$를 구하게 된다.[16]

14) 처치의 경우 $D=1$, 그렇지 않으며 0의 값을 가진다.

15) 식 (6)과 식 (7)의 g_0와 m_0를 머신러닝으로 두 번 예측한다는 의미에서 '이중(double) ML'이라 불린다.

16) 이를 'Frisch-Waugh-Lovell theorem'이라 한다.

$$\hat{\theta}_0^{DML} = \left(n^{-1}\sum_i \hat{V}^2\right)^{-1}\left(n^{-1}\sum_i \hat{U}\hat{V}\right) \tag{8}$$

DML로 구한 $\hat{\theta}_0^{DML}$는 다음과 같은 성격(properties)을 가지는 것으로 알려져 있다. 첫째, $\hat{\theta}_0^{DML}$는 탈편의(debiased)의 성격을 가진다. Chernozhukov et al. (2018)에 따르면, $\sqrt{n}(\hat{\theta}_0^{DML} - \theta_0)$은 아래와 같이 분해된다.

$$\begin{aligned}
&\sqrt{n}(\hat{\theta}_0^{DML} - \theta_0) \simeq O_1 + O_2,\\
&\quad \text{여기서 } O_1 = (E[V^2])^{-1}\frac{1}{\sqrt{n}}\sum\left(\hat{m}_0(X) - m_0(X)\right)\left(\hat{g}_0(X) - g_0(X)\right),\\
&\qquad \lim_{n\to\infty} O_2 = 0
\end{aligned} \tag{9}$$

식 (9)에서 $n\to\infty$일 때 $\sqrt{n}(\hat{\theta}_0^{DML} - \theta_0)$는 O_1에 의해 결정되는데, 만약 $\hat{m}_0(X) - m_0(X)$와 $\hat{g}_0(X) - g_0(X)$ 중 하나라도 0이면, $\sqrt{n}(\hat{\theta}_0 - \theta_0) = 0$, 즉 $plim_{n\to\infty}\hat{\theta}_0 = \theta_0$이 성립한다. 나아가 $\hat{m}_0(X) - m_0(X) \neq 0$와 $\hat{g}_0(X) - g_0(X) \neq 0$일 경우, $\hat{m}_0$와 $\hat{g}_0$가 m_0와 g_0에 각각 $n^{-\phi_1}$, $n^{-\phi_2}$의 비교적 느린 속도로 접근하더라도, 두 항의 곱(product)인 O_1은 $n^{-(\psi_1+\psi_2)}$의 빠른 속도로 0에 수렴하여 $plim_{n\to\infty}\hat{\theta}_0 \simeq \theta_0$가 성립하게 된다.

둘째, $\hat{\theta}_0$는 머신러닝 추정값 $\hat{g}_0$와 $\hat{m}_0$에 대해 강건성(robustness)을 가진다. 앞의 식 (8)의 '모멘트 조건(moment condition)'은 아래 식으로 표현된다.

$$E[\psi(Y,D,X:\theta_0,m_0,g_0)] = 0 \text{ 혹은 } \frac{1}{n}\sum\psi(Y,D,X;\hat{\theta}_0,\hat{g}_0,\hat{m}_0) = 0 \tag{10}$$

여기서 $\psi(Y,D,X:\theta_0,g_0,m_0) = [\,Y - g_0(X) - \theta_0(D - m_0(X))\,]\,\lfloor\, D - m_0(X)\,\rfloor$은 '점수함수'로서 식 (11)의 '네이먼 직교성(Neyman orthogonality)'이 성립하는 것으로 알려져 있다(Chernozhukov et al., 2018).

$$\begin{aligned}
&\partial_\eta E[\psi(Y,D,X:\theta_0,g_0,m_0)]_{\hat{\eta}_0 \simeq \eta_0} = 0,\\
&\text{여기서 } \hat{\eta} = (\hat{g}_0,\hat{m}_0),\ \eta_0 = (g_0,m_0)
\end{aligned} \tag{11}$$

식 (11)이 의미하는 바는 명확하다. 방해함수의 추정값 $\hat{\eta}=(\hat{g}_0,\hat{m}_0)$가 실제(true) 방해함수 $\eta_0=(g_0,m_0)$와 비록 같지 않더라도, $\hat{\eta}_0$가 η_0와 서로 충분히 가까운 근방(neighbor)에 위치하게 되면, 식 (10)의 '모멘트 조건'은 여전히 성립하며, 그 결과 $\hat{\theta}_0^{DML}$은 여전히 유효한 θ_0 추정치(valid method of moment estimator)가 됨을 뜻한다. 실증분석에서 $\hat{\eta}_0 \simeq \eta_0$의 조건을 충족하기 위해서 우수한 예측력(predictive power)을 가지는 것으로 알려진 머신러닝의 방법을 적용하게 된다.

DML을 장점을 정리·요약하면 다음과 같다. 첫째, 성과변수(Y), 처치변수(D), 설명변수(X) 사이의 관계에 대해 사전적으로 알려지지 않은 함수를 설정하는 대신 유연한(flexible) 형태의 관계를 가정함으로써 설정의 오류로부터 발생하는 문제를 상당 부분 해소할 수 있다. 둘째, DML은 잔차화 하는 과정에서 데이터 기반 머신러닝을 통해 방해함수의 예측 정확성을 높임으로써 '$\sqrt{n}$-일치성($\sqrt{n}$ consistency)'를 확보하고 편의를 해소(debiased)하는 이점을 가지게 된다. 셋째, 비록 방해함수의 추정이 완전하지 못하더라도 '네이먼 직교성'에 의해 DML은 '모멘트 조건'을 충족하는 강건하고 유효한 추정치의 성격을 가지게 된다.

DML의 장점과 함께 한계도 인식할 필요가 있다. DML은 〈그림 2〉 X와 같이 관측 가능한(observable) 혼재 요인인 존재할 때, 처치효과를 추정하는 방법이므로, 만약 X가 관측 불가능한(unobservables) 경우, 인과추론의 식별에 필요한 '무혼재(unconfoundedness) 조건', 즉 $(Y \perp D) \mid X$를 확보할 수 없게 되어 DML의 결과는 신뢰를 잃게 된다. 결국 DML은 관측 가능한 혼재 변수의 문제가 존재할 때, 가장 유연한 형태의 모형을 통해 처치효과를 추정하는 방법일 뿐이며, 관측자료(observational data)의 한계로부터 오는 식별의 문제(selection on unobservables)까지 해결하는 방법은 아니다.

3. 처치효과의 이질성(Treatment effect heterogeneity)

프로그램의 동질적 효과($\tau=\theta_0$, for $\forall i$)를 가정한 식 (6)을 $Y=\theta_0(X)D+g_0(X)+U$로 수정하면 처치효과 $\theta_0(X)$는 참여자의 특성($X=x$)에 따라 이질적인 값을 가지게 된다. '무혼재 (unconfoundedness)' 조건에서 이질적 처치효과인 CATE는 아래의 식으로 주어진다.

$$\theta_0(x) = E[\,Y_1 - Y_0 \mid D=1, X=x\,]$$
$$= E[\,Y \mid D=1, X=x\,] - E[\,Y \mid D=0, X=x\,] \quad (12)$$

식 (12)의 이질적 효과를 추정하는 가장 직접적인 방법은 $\mu_d(x) = E[\,Y \mid D=d, X=x\,]$를 머신러닝으로 추정한 $\hat{\mu}_1(x)$와 $\hat{\mu}_0(x)$의 값을 대입하여 $\hat{\theta}_0(X) = \hat{\mu}_1(X) - \hat{\mu}_0(X)$를 계산하는 방법이다.[17] 하지만 $\hat{\mu}_1(x)$와 $\hat{\mu}_0(x)$을 머신러닝으로 예측하는 과정에서 발생하는 정규화 편의(regularization bias)로 $\hat{\theta}_0(X)$가 제대로 추정되지 못할 수가 있다.

$\theta_0(x)$가 x의 근방(neighborhood, $N(x)$)에서 상대적으로 일정한 값을 가진다는 가정 하에, 식 (8)의 DML을 $N(x)$ 영역에 적용하여 $\hat{\theta}_0(x)$를 식 (13)으로부터 구하게 된다(Athey and Wager, 2019).

$$\hat{\theta}_o(x) = \frac{\sum_{X \in N(x)} (Y - \hat{g}_0(X)(D - \hat{m}_0(X))}{\sum_{X \in N(x)} (D - \hat{m}_0(X))^2} \quad (13)$$

이때 식 (13)의 $N(x)$는 랜덤 포레스트 기반 근방(Random forest based neighborhood)이며 $\hat{g}_0(X)$와 $\hat{m}_0(X)$는 아래와 계산된다.

$$\hat{g}_0(X) = \frac{1}{B}\sum_b^B \sum_i^n Y \frac{1[X_i \in L_b(x)]}{|L_b(x)|} = \sum_i^n Y w(x) \quad (14)$$

$$\hat{m}_0(X) = \frac{1}{B}\sum_b^B \sum_i^n D \frac{1[X_i \in L_b(x)]}{|L_b(x)|} = \sum_i^n D w(x) \quad (15)$$

여기서 B는 배깅(bagging)의 수와, $L_b(x)$는 나무(tree) b의 잎(leaf)를 말하며, $\frac{1[X_i \in L_b(x)]}{|L_b(x)|}$은 나무 b에서 같은 잎에 속한 관측치는 동일한 가중치를 가지게 됨을 말한다. 식 (13)과 식 (14)에 따르면 $N(x)$에서 정의되는 방해함수 $\hat{g}_0(X)$와

17) 머신러닝을 적용하는 구체적인 방법(S-learner, T-learner, X-learner)에 대해서는 Künzel, Sekhon, Bickel and Yu(2019)를 참조할 수 있다.

$\hat{m}_0(X)$는 Y 및 D의 $X=x$에서 $w(x)$를 가중치로 사용한 포레스트 커널(forest kernel)로 볼 수 있다.

V. 실증분석과 결과 비교: 직업훈련의 효과

본 장에서는 직업훈련 자료에 인과관계 추론 모형인 DML을 적용한 후 기존 연구 결과와 비교하였다. 본 연구의 주된 목적에 비추어, 과거 연구에서 사용한 동일한 자료에 동일한 변수를 사용하여 DML을 적용하였다. 다만 DML추정에 있어 소수의 제한된 변수를 사용하면서 초래될 수 있는 DML의 불안정성을 줄이기 위해 '잠재적 고용가능성의 향상 정도(h)'인 $h_i = E[Y_i = 1 \mid D=1] - E[Y_i = 0 \mid D=0]$를 포함하였다.[18]

본 장에서 각 추정치의 차이는 자료나 변수선택이 아닌 방법론에 기인하는 것으로 해석될 수 있으며, 이때 설정의 오류로부터 비교적 자유롭고, 추정치의 탈편의성과 강건성을 특징으로 하는 DML의 결과는 훈련효과 추정범위의 방향성을 제시하는 데 유용한 정보가 된다.[19]

1. 분석자료

본 연구는 2013년 구직자 정보(Work-Net), 훈련정보(HRD-NET), 고용보험을 연결하여 파악된 구직정보-훈련정보-취업 자료를 사용하였다.[20] 2000년대 중반 훈련계좌제(2008~2010년)의 시작에서 내일배움카드제(2011~2019년)를 거쳐 국민내일

18) h_i는 반응함수(response function)의 차이로, 훈련의 평균처치효과 $\tau_{ATET} = E[Y_i^1 - Y_i^0 \mid D=1] = E[Y_i \mid D=1] - E[Y_i^0 \mid D=1]$와는 구분된다. 유사한 방법을 적용한 연구로 Klauber & Koch (2023)는 무더위가 고령층의 병원 입원율에 미치는 영향을 분석할 때 개인의 더위와 관련된 잠재적 취약성(heat-related vulnerability to hospitalization)을 산출하였으며 최저임금 분석에 적용한 연구로는 Cengiz et al. (2021)이 있다. 한편, 본 연구의 표적 계수(target parameter) θ_0는 h_i 포함 여부에 큰 영향을 받지 않는 것으로 나타났다(〈표 2〉 참고).

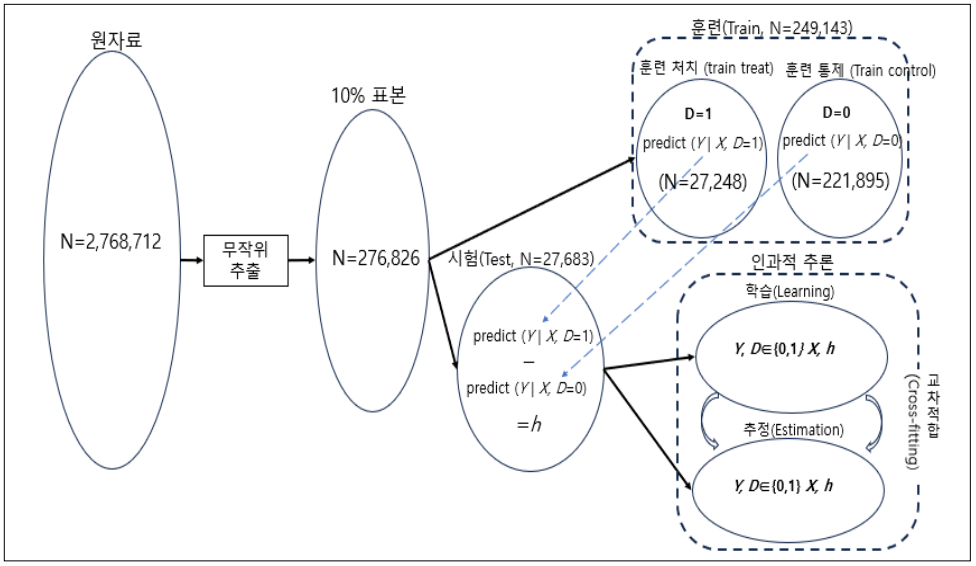
19) 일자리 사업을 평가한 결과가 일률적인 나타나는 것은 좋은지는 의문이 있다. 사업의 특수성(훈련 내용의 차이), 참여자들이 이질성(훈련 교·강사 교습 방법과 수강생의 학습능력 차이), 정책환경의 차이(훈련 시간과 장소)에 따라 효과는 서로 다를 수밖에 없기 때문이다. 본 연구의 규칙성은 일률적인 훈련효과를 의미하는 것이 아닌 하기보 수합리적 범위의 수용 가능한 훈련효과이다.

20) 본 연구의 자료에 대한 보다 상세한 내용은 김용성(2020)을 참조할 수 있다.

배움카드제(2020년 이후)까지 훈련제도가 거쳐온 큰 변화에 비추어 10년이 지난 과거 자료를 분석하는 것이 과연 적절한지 의문이 있을 수 있다. 사용 가능한 가장 최신의 자료를 활용하여 현재 시행되는 훈련 프로그램의 효과를 평가하는 것은 매우 중요한 작업임이 틀림없다. 그럼에도 불구하고 본 연구가 의도적으로 10년이 지난 자료를 재분석하는 이유는 같은 자료에 상이한 추정법을 통해 얻은 다양한 연구 결과를 서로 비교할 수 있기 때문이다. 구체적으로 본 연구의 중요한 목적이 DML 방법을 적용하여 정밀하게 훈련 효과를 측정하는 데 있다는 점에서 동일한 자료를 분석하여 추정한 효과가 기존의 연구인 결과와 다를 때, 그 차이는 추정법의 차이로 해석될 수 있다.

분석자료는 개인의 특성 변수(학력, 나이, 성별, 지역, 취약계층 여부)와 훈련과 관련된 정보(참여 여부, 참여 훈련과정, 훈련비, 과정 이수 여부) 등 다양한 변수를 포함하고 있다. 한편 자료가 가진 한계를 보면, 분석대상으로 적절하지만, 구직자 정보, 훈련정보, 고용보험 자료에 포착되지 않아 누락되는 경우가 있다. 예를 들어, 고용센터에 구직등록을 하지 않은 실업자, 실제 훈련에 참여하였으나 HRD-Net에 훈련정보가 없는 사람, 고용보험 미가입 사업장의 취업자 등은 분석자료에서 제외된다.

〈그림 3〉 분석자료의 구축



〈그림 3〉은 분석자료의 구축 과정을 보여준다. 먼저 김용성・박우람(2015)과 김용성(2020)이 분석에 사용한 원자료(Raw data) 관측치(N=2, 768, 712)의 약 10% 표본

(N=276, 826)을 무작위 추출하였다.[21] 본 연구는 일반적인 머신러닝을 기법에 따라 10% 표본을 '훈련 데이터(Train, N=249, 143)'와 '시험 데이터(Test, N=27, 683)'로 분할하였으며,[22] 인과적 추론을 위한 DML을 적용하기 위해 시험 데이터를 다시 '학습(Learning) 데이터'와 '추정(Estimation) 데이터'로 나누고, 두 데이터에 DML 알고리즘을 교차적합(cross-fitting)하여 $\hat{g}_0$, $\hat{m}_0$를 구하였다.

〈표 1〉 기초통계량: 원자료와 분석자료의 비교

변수		원자료 (Raw data)	10% 표본 (10% Random sample)		
				훈련(Train) 데이터	시험(Test) 데이터
취업 여부(y) (취업=1)		0.32	0.32	0.32	0.32
훈련 여부(D) (훈련=1)		0.11	0.11	0.11	0.11
나이 (세)		38.69	38.71	38.71	38.72
성별 (남성=1)		0.44	0.44	0.44	0.43
경력 유무 (있음=1)		0.48	0.48	0.48	0.48
학력 (%)					
	고졸미만	12.47	12.47	12.52	12.01
	고졸	36.55	36.55	36.50	37.01
	전문대	20.90	20.90	20.90	20.92
	대졸 이상	30.07	30.07	30.08	30.05
구직건수 (N)		2, 768, 712	276, 826	249, 143	27, 683

추정에 앞서 무작위 추출과 데이터 분할 과정에서 원자료가 왜곡되었을 가능성을 점검할 필요가 있다. 〈표 1〉은 원자료와 본 연구의 분석자료의 기초통계량을 보여주

21) 원자료(Raw data)를 분석의 전산상 부담(computational burden)을 줄이기 위해 무작위 표본추출(random sampling)을 하였다.

22) 통상적으로 훈련 데이터와 시험 데이터는 8:2 정도로 비율로 나누지만, 본 분석에서는 그 비율을 9:1로 하였는데, 이는 낮은 취업률과 훈련 참여율 상황에서 훈련 데이터를 이용한 충분한 머신러닝이 일어나도록 하기 위함이다.

는데, 표에 따르면 취업 여부(y)와 훈련 여부(D)의 평균은 모든 자료에서 일치하며, 나이, 성별, 학력, 교육 수준 등 특성변수도 대부분 같거나 아주 작은 차이만 보인다.

분석자료('시험 데이터')를 간략히 소개하면, 구직건수의 중 약 32%가 취업을 보였으며, 직업훈련에 참여한 경우는 약 11%로 나타났다. 구직자의 평균 연령은 38.7세이며, 남성의 비율이 43% 나머지는 여성으로 구성되어 있다. 전체 구직자의 48%가 경력이 있는 것으로 응답하였으며, 학력분포를 보면 고졸이 약 37%로 가장 많으며, 이어서 대졸 이상(약 31%), 전문대(약 21%), 고졸미만(약 12%)의 순서를 보였다.[23]

2. 평균 처치효과(Average Treatment Effect, ATE) 추정의 결과 비교

〈표 2〉는 추정 방법을 차이에 따른 결과를 보여주고 있다. 통제변수의 추정 결과를 해석하는 대신, 본 연구의 표적 계수인 '훈련 여부'에 초점을 두고 설명하고자 한다. 우선 (모형 1)~(모형 4)의 원자료, 10% 표본, 시험 데이터에 OLS를 적용한 결과에서 ATE의 값이 -0.101~-0.103으로 매우 비슷한 것을 확인할 수 있으며, OLS 결과와 〈표 1〉의 기초통계량으로부터 본 연구의 분석자료인 '시험 데이터'와 기존 연구에서 사용한 '원자료' 사이에 자료의 일관성은 충분히 유지되는 것으로 판단된다.

(모형 1)~(모형 5)까지 회귀분석에 기반한 훈련의 평균 처치효과를 보면 추정계수가 모두 통계적으로 유의미한 음의 값을 가지는 것으로 나타났는데, 특히 종속변수(취업 여부)의 성격을 반영한 프로빗 (모형 5)의 추정계수는 -0.306이며, 이를 바탕으로 계산한 훈련의 평균 한계효과(average marginal effect)는 약 -10%p에 이르러 상대적으로 부정적 효과가 크게 나타났다. 한편 내생성을 완화하기 위해 '잠재적 고용가능성의 향상 정도(h)'를 포함한 (모형 4)의 추정결과가 (모형 3)의 결과보다 0에 가깝다는 점에서 부정적 효과가 소폭 완화되는 것을 알 수 있다.[24] 다만 이를 바

23) 성별 특성변수의 요약은 김용성(2020)을 참조할 수 있다.

24) 훈련 여부(D) 및 취업 여부(y)와 관계를 각각 $\theta(=cov(h,D)/var(D))$와 $\lambda(=cov(y,h)/var(h))$라 할 때, h를 포함하지 않은 모형 3의 추정계수 $\hat{\beta}^{(3)}$과 h를 포함한 모형 4의 추정계수 $\hat{\beta}^{(4)}$의 관계는 $\hat{\beta}^{(3)}=\hat{\beta}^{(4)}+\lambda\theta$로 표현된다. 분석에서 $\theta>0$, $\lambda<0$로 나타나면서, $\hat{\beta}^{(3)}=-0.103<\hat{\beta}^{(4)}=-0.101$의 관계가 성립하게 되는데, 이는 h의 누락이 편향된 추정치로 이어질 가능성을 시사한다. 즉 훈련을 통한 잠재적 고용가능성 개선 효과는 크지만, 실제 취업으로

〈표 2〉 훈련이 취업에 미치는 평균 처치효과(Average Treatment Effect) 비교

		회귀분석 기반(regression-based)					인과적 머신러닝(DML)			
		OLS (모형 1)	OLS (모형 2)	OLS (모형 3)	OLS (모형 4)	Probit (모형 5)	Lasso (모형 6)	R-Forest (모형 7)	Tree (모형 8)	Xgboost (모형 9)
변수		원자료	10% 표본	시험 데이터	시험 데이터	시험 데이터	시험 데이터	시험 데이터	시험 데이터	시험 데이터
훈련 여부 (훈련=1)		-0.102** (0.001)	-0.102** (0.003)	-0.103** (0.008)	-0.101** (0.008)	-0.305** (0.027)	-0.092**	-0.074**	-0.085**	-0.070**
나이		-0.012** (0.000)	-0.012** (0.001)	-0.011** (0.002)	-0.001 (0.002)	-0.002 (0.005)	○	○	○	○
$나이^2$		0.000** (0.000)	0.000** (0.000)	0.000** (0.000)	0.000 (0.000)	0.000 (0.000)	○	○	○	○
성 (남성=1)		0.021** (0.001)	0.020** (0.002)	0.015** (0.006)	0.001 (0.006)	0.004 (0.017)	○	○	○	○
경력 (있음=1)		0.030** (0.001)	0.033** (0.002)	0.026** (0.006)	-0.027** (0.008)	-0.075** (0.023)	○	○	○	○
학력 (기준=고졸 미만)										
	고졸	0.031** (0.001)	0.030** (0.003)	0.015 (0.010)	0.019 (0.010)	0.056 (0.029)	○	○	○	○
	전문대졸	0.050** (0.001)	0.050** (0.004)	0.030* (0.012)	0.028* (0.012)	0.080* (0.034)	○	○	○	○
	대졸 이상	0.028** (0.001)	0.028** (0.003)	0.023* (0.011)	0.019 (0.011)	0.056 (0.032)	○	○	○	○
잠재적 고용가능성 정도 (h)					-0.493** (0.048)	-1.395** (0.138)	○	○	○	○
상수항		0.557** (0.003)	0.554** (0.010)	0.553** (0.030)	0.557** (0.030)	0.182* (0.084)				
관측 수 (N)		2,768,712	276,826	27,683	27,683	27,683	27,683	27,683	27,683	27,683

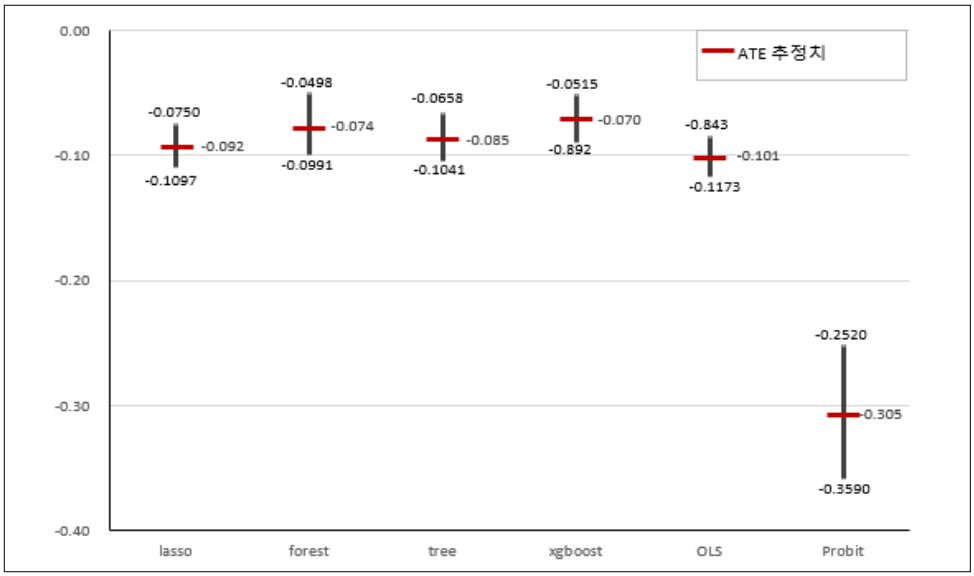
주: 1) **=5% 통계적 유의성, *=10% 통계적 유의성.
2) 괄호 안의 숫자는 표준오차(standard error)임.
3) R-forest는 Random-forest, Tree는 Decision tree임.

이어질 확률은 높지 않은 구직자로 표본 선택의 편의(sample selection bias)가 있음을 의미한다.

탕으로 훈련이 취업에 별다른 도움이 되지 않다고 할 수는 단언하기는 어렵다. 그 이유는 훈련은 단기보다는 장기에 긍정적인 효과를 가지는 것으로 알려져 있는데, 분석 자료가 과연 훈련의 효과를 관찰하기에 충분한 기간을 포함하고 있는지에 대해 의문이 있기 때문이다.[25]

(모형 6)~(모형 9)는 방해함수 $g_0(X)$와 $m_0(X)$의 조건부 기댓값을 추정하는 데 사용한 다양한 머신러닝 알고리즘에 기반하여 구한 DML의 추정 결과를 보여주고 있다. 언급한 바와 같이 DML은 y, D, X의 관계에 있어 유연한 형태를 취함으로써 설정의 오류로부터 자유롭고, 나아가 탈편의성과 강건성으로 유효한 모멘트 추정치(valid MOM estimator)의 성질을 가진다는 점에서 훈련 효과의 방향성을 제시할 수 있다. 우선 DML의 ATE 추정 결과는 -0.070~-0.092의 범위에서 통계적으로 유의미한 음의 값을 가지는 것으로 나타났다.

〈그림 4〉 추정방법에 따른 평균 처치효과(ATE) 비교



〈그림 4〉는 (모형 6)~(모형 9)의 DML 추정 결과와 OLS(모형 4)와 프로빗(모형 5)으로 추정한 ATE의 값과 95% 신뢰구간(confidence interval)을 비교하고 있다. 〈그림 4〉에서 DML과 OLS의 ATE 값은 비슷하며, 프로빗 추정치는 다소 차이를 보

25) Card et al. (2018)의 Table 3(a)에 따르면 훈련은 단기(프로그램 종료 후 1년 이내)보다 장기(프로그램 종료 후 2년 초과)에 효과가 나타남을 밝히고 있다.

이며, 특히 프로빗은 다른 추정법에 비해 넓은 95%의 신뢰구간을 가짐을 알 수 있다. 한편 DML 중 랜덤 포레스트(random forest)와 Xgboost의 추정계수의 값은 비슷하며, LASSO와 의사결정나무(decision tree)는 훈련 효과에 조금 더 부정적인 것으로 나타났다.

결과를 종합하면 프로빗 모형은 ATE의 음의 효과를 과대 추정할 가능성과 추정값의 신뢰구간도 가장 넓게 나타나는 등 문제를 보였다. 반면 OLS와 유사한 값을 보인 DML의 탈편의성과 강건성을 감안할 때 훈련 효과 추정범위의 방향성을 제시하고 있으며, 특히 랜덤 포레스트와 Xgboost는 매우 유사한 추정계수 값과 좁은 신뢰구간을 보인다는 점에서 우수한 것으로 판단된다.

3. 집단별 평균 처치효과(Group Average Treatment Effect, GATE)와 조건부 평균 처치효과(Conditional Average Treatment Effect, CATE)

현실에서 일정한 특성을 가진 집단(가령 미취업 청년층, 빈곤층 등)을 집중적으로 배려하는 정책을 시행할 때는 전체(population)를 대상으로 하는 ATE보다 해당 집단의 평균 효과(Group ATE)를 살펴봄이 적절하다. 〈그림 5〉는 예시적으로 성과 학력 수준별로 구분되는 8개의 집단에 대해 DML로 추정한 GATE의 결과이다.[26]

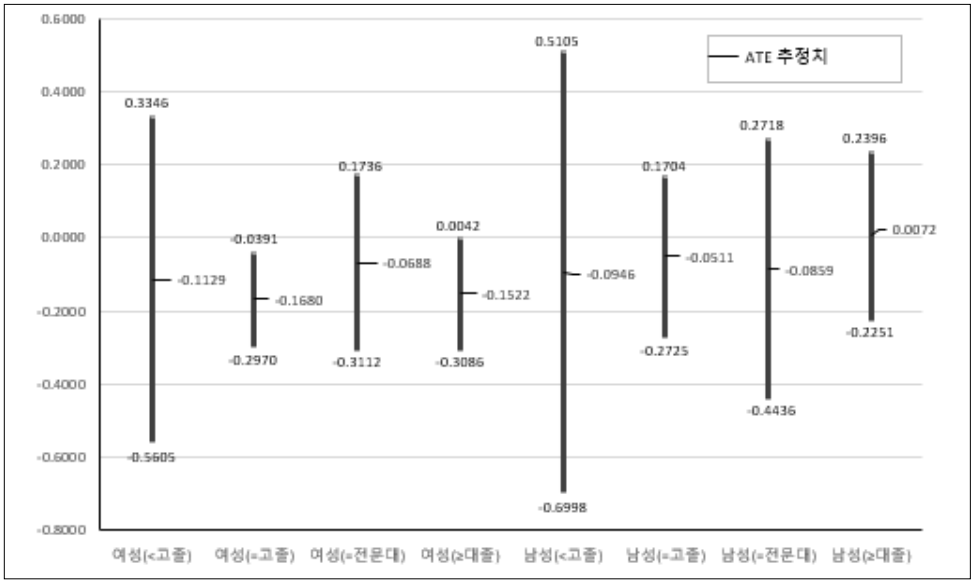
〈그림 5〉로부터 다음과 같은 점이 관찰된다. 첫째, 여성 고졸 학력층은 통계적으로 유의미한 부정적인 훈련 효과를 보이고 있을 뿐, 그 밖에 대부분의 집단에서는 추정치의 95% 신뢰구간이 0을 포함하고 있어 통계적으로 의미가 있는 훈련 효과를 보이지 않는다. 둘째 비록 통계적이로 유의미하지 않으나, 남성과 여성을 비교하면 대체로 남성(전문대졸 제외)의 훈련 효과가 여성보다 약간 양호한 것으로 나타나고 있다. 이와 같은 결과는 훈련이 취업에 미치는 영향에 있어 남녀 간 체계적인 차이가 존재하지 않는다는 다수의 기존 연구와 궤를 같이한다(Card et al., 2010; Kluve, 2010; Card et al., 2018). 셋째, 교육 수준별 훈련 효과에 있어 남녀 모두 고졸 마만 학력층에서 추정치의 95% 신뢰구간이 가장 넓은 특징을 보이며, 이는 저학력층에 대한 훈련 효과의 불확실성이 클 가능성을 시사한다. 특히 남성 고졸미만 학력층의 훈련 효과의 불확실성이 가장 높게 나타난다는 데, 이는 제한된 변수의 선택이나 계량기법

---

26) $\tau_{GATE}(X) = E[\,Y(1) - Y(0) \mid X \in G]$ 이며 $g_0(X)$와 $m_0(X)$의 조건부 기댓값은 random forest로 추정하였다.

의 불완전성에 기인하거나 또는 해당 계층이 속한 개인이 매우 이질적인 경우, 추후 연구를 통해 더욱 밝혀져야겠지만 만약 후자의 경우 보다 정교하고 세심한 정책적인 설계가 필요함을 시사한다.

〈그림 5〉 집단별 평균 처치효과(Group ATE) 비교: 성×학력수준



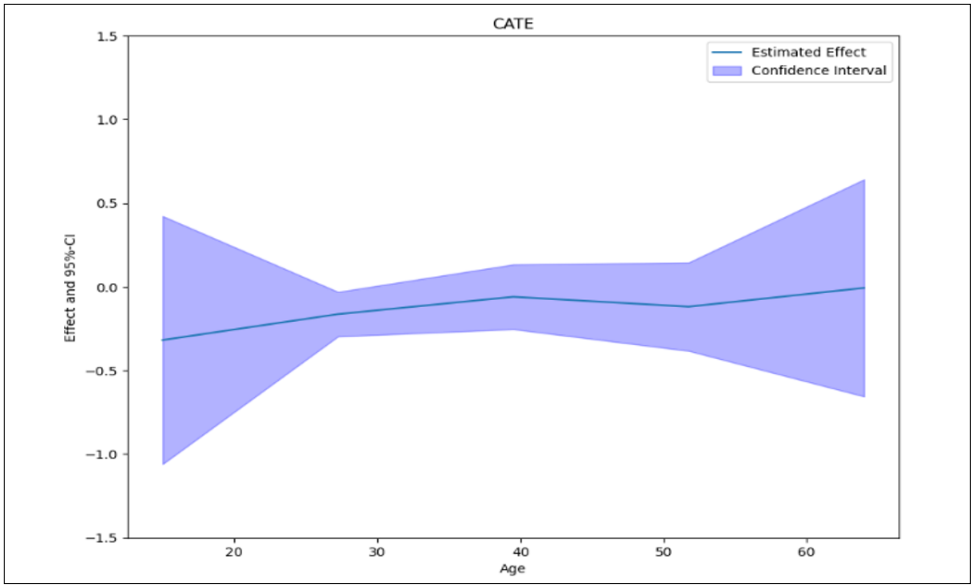
〈그림 6〉은 DML을 적용해 추정한 연령에 따른 훈련의 조건부 평균 효과(Conditional ATE)와 95% 신뢰구간을 보여준다. GATE가 주로 교육 수준 등 범주형(categorical) 특성에 의해 구분되는 집단을 분석하는 데 효과적이라면, CATE는 연령, 소득 등 연속적(continuous) 변수에 따른 처치효과를 파악하는 데 활용된다.[27]

〈그림 6〉에 따르면 40세 이전까지 훈련 효과는 완만히 증가하다가, 40대에는 정체 또는 감소하는 추세를 보인 후, 50대 초반부터 다시 증가하는 비선형적인 모습을 보인다. 다만 연령별 추정치의 95% 신뢰구간이 0을 포함하고 있어, 연령에 따른 통계적으로 의미가 있는 훈련 효과와 특징을 말하기 힘들다. 이러한 결과는 연령에 따른 훈련 효과가 불명확하다는 일련의 기존연구의 결과를 뒷받침한다(Fay, 1996; Koning, 2007; Kluve, 2016; Card et al., 2018). 한편 연령별 추정치의 95% 신뢰구간의 모

27) $\tau_{CATE}(X=x_i)=E[Y(1)-Y(0) \mid X=x_i]$이며, GATE와 같이 $g_0(X)$와 $m_0(X)$의 조건부 기댓값은 random forest로 추정하였다.

습을 보면 30대까지 청년기와 50대 이후 장년기에 넓게 나타나고, 30~50대 연령층에서는 좁은 모습을 보인다. 이는 훈련대상의 측면에서 볼 때 청년층과 장년 및 고령층이 매우 이질적이며, 따라서 훈련 효과의 불확실성을 줄이는 방향의 정책적 노력(예를 들어, 수준별 훈련 교육, 적절한 수강 규모 등)이 필요함을 시사한다.

〈그림 6〉 연령에 따른 조건부 평균 처치효과(Conditional ATE)



Ⅵ. 결 론

본 연구는 최근 노동시장 분석에 활발히 사용되는 인과적 머신러닝 기법을 소개하고 훈련 데이터에 적용한 결과를 제시하였다. 일자리사업의 효과를 살펴본 다수의 연구가 있었음에도 평가는 일치된 결론에 이르지 못하고 있다. 상당한 정부의 재원이 투입되는 훈련 분야의 취업 효과를 분석한 연구의 결과도 비교가능성과 타당성 확보에 어려움을 겪는 것으로 알려져, 정책적 판단에 필요한 유용한 정보를 제공하는 데 제한적인 역할에 머물고 있다.

본 연구가 인과적 추론을 위해 소개하는 이중 머신러닝(double machine learning, DML)은 평가모형의 변수들 관계를 구체적 형태로 설정하는 강한 가정(strong assumption)을 완화함으로써 연구모형의 자의성을 줄일 수 있으며, DML이 가지는 탈편의성과 강건성은 합리적이고 수용 가능한 정책효과 추정값의 범위를 제공하는데

방향성을 제시할 수 있다. 또한 처치효과의 동질성(treatment effect homogeneity)에서 이질성(heterogeneity)을 적극 고려하면서 평균 처치효과(Average Treatment Effect, ATE)뿐만 아니라 집단별, 인적 특성 변수에 따른 집단의 평균 처치효과(Group ATE)와 조건부 평균 처치효과(Conditional ATE)를 정교하게 파악하는 방법을 시도하였다.

기존 연구가 다양한 방법으로 추정·분석한 2013년 구직·훈련·고용 자료에 DML을 적용하여 훈련 효과를 살펴보았다. 과거 자료를 사용하여 분석한 이유는 동일한 추정환경에서 얻은 DML의 결과를 기존의 연구 결과와 직접 비교할 수 있기 때문이다. 평균 처치효과에서 DML과 OLS는 매우 유사한 값을 보였으며, 프로빗 모형과는 다소 차이가 있는 것으로 나타났다. 또한 교육 수준별 훈련 효과에 있어 남녀 모두 고졸 마만 학력층에서 추정치의 95% 신뢰구간이 가장 넓은 특징을 보였으며, 연령별 평균 처치효과의 특별한 패턴은 확인되지 않으며, 추정치의 신뢰구간이 30대까지 청년기와 50대 이후 장년기에 넓게 나타나는 특징을 보였다.

머신러닝을 통한 프로그램 효과를 추정하는 방법을 제안하는 본 연구의 목적에서 나아가 추후 최신의 자료와 풍부한 정보를 활용해 일자리사업의 효과를 엄밀하게 추정하고 다양한 차원에서 바라보는 연구로 이어질 필요가 있다.

■ 참 고 문 헌

1. 강순희·어수봉·최기성, “미취업자의 직업훈련 참가 결정요인과 고용성과 분석,”『HRD 연구』, 제17권 제2호, 2015, pp. 267-298.
2. 금재호, “시행주체에 따른 직업훈련의 임금효과 연구,”『한국경제연구』, 제34권 제2호, 2016, pp. 121-151.
3. 김보배, “가구 소득분위에 따른 직업훈련의 효과 추정,” *Asia-pacific Journal of Convergent Reserch Interchange*, Vol. 9, No. 8, 2023, pp. 365-377.
4. 김용성·박우람,『실업지속의 원인분석과 직업훈련의 효과 및 개선방안에 관한 연구』, 한국개발연구원, 2015.
5. 김용성, “내일배움카드제 훈련이 취업성과에 미치는 영향,”『노동경제논집』, 제43권 제1호, 2020, pp. 1-34.
6. 류기락·나영선·이수경·김미란·정재호,『직업능력개발훈련 이수자 실태조사』, 한국직업능력개발원, 2014.
7. 양용현·최광성·최 충, “재직자 직업훈련이 취업 및 이직에 미치는 영향,”『노동경제논집』, 제42권 제3호, 2019, pp. 75-98.
8. 오호영, “사업주 직업훈련 참여와 보상의 성별격차,”『노동정책연구』, 제21권 제1호, 2021, pp. 99-127.

9. 유경준 · 강창희, “직업훈련의 임금효과 분석: 경제활동인구조사를 중심으로,”『한국개발연구』, 제32권 제2호, 2010, pp. 29-53.
10. 유경준 · 이철인, “실업자 직업훈련의 효과 추정,”『노동경제논집』, 제31권 제1호 2008, pp. 59-103.
11. 이병희, “실업자재취직훈련의 재취업 성과에 관한 준실험적 평가,”『노동경제논집』제23권 제2호, 2000, pp. 107-126.
12. 장신철 · 박종성 · 최윤정 · 조한진 · 이영란 · 오문환,『고용서비스 전문자격 신설에 대한 타당성 연구』, 한국기술교육대학교 고용직업능력개발센터 연구보고서 2021-06.
13. 조준모 · 박상일 · 나영선 · 안준기 · 이재성,『직업능력개발계좌제 시범사업 평가 및 개선방안 연구』, 뉴거버넌스연구센터, 2010.
14. Ahrens, A., C. B. Hansen, M. E. Schaffer, and T. Wiemann, “ddml: Double/Debiased Machine Learning in Stata,” IZA DP No. 15963, 2023.
15. Athey, S., The Impact of Machine Learning on Economics, in Ajay Agrawal, Joshua Gans, Avi Goldfarb eds, The Economics of Artificial Intelligence: An Agenda, University of Chicago Press, 2019, pp. 507-547.
16. Athey, S., and S. Wager, “Estimating Treatment Effects with Causal Forests: An Application,” *Observational Studies*, Vol. 5, Issue 2, 2019, pp. 37-51.
17. Card, D., J. Kluve, and A. Weber, “What Works? A Meta Analysis of Recent Active Labor Market Program Evaluations,” *Journal of the European Economic Association*, Vol. 16, No. 3, 2018, pp. 894-931.
18. _____, “Active Labour Market Policy Evaluations: A Meta-analysis,” *The Economic Journal*, Vol. 120, 2010, F452-F477.
19. Cavaco, S., D. Fougère, and J. Pouget, “Estimating the Effect of a Retraining Program on the Re-Employment Rate of Displaced Workers,” *Empirical Economics*, Vol. 44, No. 1, 2013, pp. 261-287.
20. Cengiz, D., A. Dube, A. S. Lindner, and D. Zentler-Munro, “Seeing Beyond the Trees: Using Machine Learning to Estimate the Impact of Minimum Wages on Labor Market Outcomes,” NBER Working Paper No. 28399, 2021.
21. Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins, “Double/debiased Machine Learning Treatment for Treatment and Structural Parameters,” *Econometrics Journal*, Vol. 21, 2018, pp. C1-C68.
22. Dawson, N.,, M-A. Rizoiu, B. Johnston, and M-A. Williams, “Predicting Skill Shortages in Labor Markets: A Machine Learning Approach,” 2020 IEEE International Conference on Big Data (Big Data).
23. Doerr, A., “Vocational Training for Female Job Returners - Effects on Employment, Earnings and Job Quality,” *Labour Economics*, Vol. 75, 2022, pp. 102-139.
24. Fay, Robert, G., “Enhancing the Effectiveness of Active Labour Market Policies: Evidence from Programme Evaluations in OECD Countries,” OECD Labour Market ans Social Policy Occasional Papers No. 18, 1996.
25. Goller, D., T. Harrer, M. Lechner, and J. Wolff, “Active Labour Market Policies for the Long-Term Unemployed: New Evidence from Causal Machine Learning,” IZA DP No. 14486, 2021.
26. Hirshleifer, S., D. McKenzie, R. Almeida, and C. Ridao-Cano, “The Impact of Vocational

Training for the Unemployed: Experimental Evidence from Turkey," *The Economic Journal*, Vol. 126, 2016, pp. 2115-2146.

27. Klauber, H., and N. Koch, Determinants of Heat Risk in an Aging Population: A Machine Learning Approach, IZA DP No. 15996, 2023.
28. Kluve, J., "The effectiveness of European Active Labor Market Programs," *Labour Economics*, Vol. 17, No. 4, 2010, pp. 904-918.
29. _____, "A Review of the Effectiveness of Active Labour Market Programmes with a Focus on Latin America and the Caribbean," Research Department Working Paper No. 9, International Labour Office, 2016.
30. Knaus, M. C., M. Lechner, and A. Strittmatter, "Heterogeneous Employment Effects of Job Search Programs: A Machine Learning Approach," *Journal of Human Resources*, Vol. 57, No. 2, 2022, pp. 597-636.
31. Koning, J., and Y. Peers, "Evaluating ALMP Evaluations," SEOR Working Paper No. 2007/2, 2007.
32. Koning, J. eds, The Evaluation of Active Labour Market Policies, Measure, Public Private Partnerships and Benchmarking, Edward Elgar Publishing, 2007.
33. Kreif, N., and K. DiazOrdaz, "Machine Learning in Policy Evaluation: New Tools for Causal Inference," Paper Submitted to Arxiv on March 2019.
34. Künzel, S. R., J. S. Sekhon, P. J. Bickel, and B. Yu, "Metalearners for Estimating Heterogeneous Treatment Effects using Machine Learning," *PANAS*, Vol. 116, No. 10, 2019, pp. 4156-4165.
35. Lee, J-W., J-S. Han, and E. Song, "The Effects and Challenges of Vocational Training in Korea," *International Journal of Training Research*, Vol. 17, 2019, supl. 96-111.
36. Martin, J. P., "What Works Among Active Labour Market Policies: Evidence From OECD Countries' Experience," OECD Labour Market and Social Policy Occasional Papers, No. 35, OECD, 1998.
37. Nivorozhkin, A., and E. Nivorozhkin, "Do Government Sponsored Vocational Training Programs Help the Unemployed Find Jobs? Evidence from Russia," *Applied Economics Letters*, Vol. 14, No. 1, 2007, pp. 5-10.
38. Rodriguez-Planas, N., and J. Benus, "Evaluating Active Labor Market Programs in Romania," *Empirical Economics*, Vol. 38, 2010, pp. 65-84.
39. Rosholm, M., and L. Skipper, "Is Labour Market Training a Curse for the Unemployed? Evidence from a Social Experiment," *Journal of Applied Econometrics*, Vol. 24, 2009, pp. 338-365.
40. van den Berg, G. J., M. Junaschk, J. Lang, G. Stephan, and A. Uhlendorff, "Predicting Re-Emplotment: Machine Learning versus Assessments by Unemployed Workers and by their Caseworkers," IZA Discussion paper series, DP. No. 16426, 2023.
41. Vooren, M., C. Haelermans, W. Groot, and H. van den Brink, "The Effectiveness of Active Labor Market Policies: A Meta-analysis," *Journal of Economic Survey*, Vol. 33, No. 1, 2019, pp. 125-149.
42. Yeyati E., M. Montane, and L. Sartorio, "What Works for Active Labor Market Policies?" CID Faculty Working Paper No. 358, Center for International Development at Harvard University, 2019.

New Direction of Evaluation on Labor Market Program Application of Causal Machine Learning to Vocational Training in Korea*

Yong-seong Kim**

Abstract

This study introduces a machine learning method (DML) recently used for causal analyses on labor market issues and presents the results of applying DML to Korea's vocational training data. Voluminous studies have estimated the effect of training on employment and they have come up short of consensus. The DML's flexibility in setting relationships of variables may help to avoid problems related to model specification. The DML's properties of the debiasedness and robustness may also provide what the reasonable range of training effects should be. In addition, the paper attempt heterogeneous treatment effects, which will be informative to policy implementation.

Key Words: causal machine learning, treatment effect, program evaluation
JEL Classification: C18, C21, J08

Received: Jan. 2, 2024. Revised: Feb. 5, 2024. Accepted: Feb. 15, 2024.

* This paper was supported by the Education and Research Program of Korea Tech in 2021. This work is based on the KEIS's "Evaluation on job market program in Korea (2023)."

** Professor, Graduate School of Techno HRD, Korea University of Technology and Education, 1600 Chungjeol-ro, Dongnam-gu, Cheonan, Chungnam 31253, Korea, Phone: 82-41-560-1101, e-mail: yongkim65@koreatech.ac.kr