

Heterogeneous Effects of Korea's National Employment Support Program*

Yong-seong Kim[†]

Abstract This paper investigates the heterogeneous treatment effects of employment services in Korea, the National Employment Support Program. The empirical analysis reveals substantial heterogeneity in program effects. In particular, positive employment effects are salient among younger individuals, women, and those with relatively limited work experience. Furthermore, the analysis of optimal treatment assignment demonstrates that selectively targeting individuals yields efficiency gains relative to uniform service provision. Overall, the findings highlight the limitations of a one-size-fits-all approach to employment services and provide empirical support for data-driven targeting based on heterogeneous treatment effects.

Keywords Treatment effect heterogeneity, individualized treatment rule.

JEL Classification C21, C45, J68, J64.

*I am grateful to the two anonymous reviewers for their comments and suggestions. This study was supported by the Education and Research Promotion Program of KOREATECH in 2024.

[†]Techno-HRD Graduate School, Korea University of Technology and Education, 18 Gwasuwon-gil, Seobuk-gu, Cheonan-si, Chungcheongnam-do, Republic of Korea 31079. E-mail: yongkim65@koreatech.ac.kr.

1. INTRODUCTION

Active Labor Market Policies (ALMPs), including employment services, are widely recognized as not having uniform effects across all participants. While the average treatment effect (ATE) provides important information regarding the overall impact of an active labor market program, increasing attention has been devoted to treatment effect heterogeneity, which captures how policy impacts vary across individuals. Understanding such heterogeneity is crucial because it reveals for whom a policy is effective, thereby enabling policymakers to design interventions that are tailored to individuals' characteristics and labor market status rather than applying a uniform one-size-fits-all approach.

This study applies causal machine learning methodologies to data from the National Employment Support Program (NESP)—Korea's representative employment service—in order to estimate the heterogeneous treatment effects of employment services. Based on these estimates, the study further examines treatment assignment rules.

This study differs from previous research evaluating employment service programs in Korea in two respects. First, to the best of my knowledge, it is among the first studies to apply causal machine learning methods to estimate the heterogeneous treatment effects of an employment service program using large-scale administrative data in Korea. Previous studies on employment service programs in Korea have mainly focused on average treatment effects or subgroup-specific average effects. While such estimates are useful for assessing overall program performance, they provide limited information on how treatment effects vary across individuals with different observed characteristics. Second, this study goes beyond documenting treatment effect heterogeneity by deriving individualized treatment rules based on the estimated CATE. In doing so, the analysis connects heterogeneity in program effects to the practical question of treatment assignment and provides a framework for informing the design of targeted employment services.

The main findings can be summarized as follows. The effects of the NESP exhibit clear heterogeneity depending on participants' characteristics. In general, positive effects are observed among youth, women, and job seekers with relatively short career interruption periods, whereas the effects are negligible or limited for older individuals, job seekers with work experience, men, and the less educated. A similar conclusion is obtained from the ITR.

This study is organized as follows. Section 2 reviews the existing literature on heterogeneity in the effects of employment services. Section 3 presents the analytical framework for the CATE, including the rationale for applying ML

methods, the challenges involved in estimation, and the discussion of ITR. Section 4 describes the data and construction process used in the empirical analysis, provides empirical evidence on the heterogeneous effects of employment services, and reports the estimation results for ITR. Section 5 concludes the study by summarizing the main findings and discussing their policy implications.

2. LITERATURE REVIEW

Numerous studies have shown that ALMPs have a considerable degree of heterogeneous effects. The impact of policy programs varies not only depending on the individual characteristics of participants—such as gender, age, and educational attainment—and their labor market status, including prior work experience, but also on program design and context, its delivery systems, and the timing and location of its implementation. Consequently, even the same program may produce different outcomes under different institutional and environmental conditions (Martin and Grubb, 2001).¹

Recent studies with more sophisticated empirical methods report that the effects of ALMPs differ substantially across target groups (e.g., women, individuals aged 50 and over, low-skilled workers, and those with low levels of education) in terms of both labor market outcomes, such as employment and job mobility, and social outcomes (OECD/European Commission, 2025). Focusing on employment services, some studies find positive effects on employment or wages (Graversen and van Ours, 2008; Card, Kluve and Weber, 2010, 2018; Crépon and van den Berg, 2016), whereas others report negative effects (Gerfin and Lechner, 2001; Lalive, van Ours and Zweimüller, 2008; Knaus, 2022). In addition, there are arguments that the effects of employment services are not uniform and depend on the methods of service delivery (Mühlböck *et al.*, 2022). Taken together, these findings illustrate that diverse and sometimes conflicting claims coexist in the literature regarding the policy effects of employment services.

Over the past decade, a growing body of research has systematically analyzed heterogeneity in policy effects using ML. Applications of ML were largely confined to prediction tasks. Recently, as ML has been extended and applied to causal inference, it has become possible to estimate CATE beyond the traditional focus on the ATE (Athey, 2018; Athey and Imbens, 2019; Wager and Athey,

¹Martin and Grubb (2001) argue that employment services are effective for unemployed women with career interruptions, while they show only modest effects for unemployed middle-aged and older men and no significant effects for the long-term unemployed.

2018; Kreif and DiazOrdaz, 2019; Lechner, 2019; Cockx, Lechner and Bollens, 2019; Curth and van der Schaar, 2021; Athey, Keleher and Spiess, 2025).

Several studies have applied causal ML methods to examine the heterogeneity of employment service effects. Using Swiss data, Mascolo *et al.* (2024) analyzed the effects of labor market policies—including employment subsidies, job search assistance, and training programs—and reported that the impacts of these policies are heterogeneous, with substantial differences depending on educational attainment and whether workers are domestic or foreign nationals. Similarly, an analysis of Swiss administrative data found that the effects of employment services within the first six months of unemployment are highly heterogeneous, with particularly large effects for individuals with little work experience and immigrants (Knaus, Lechner and Strittmatter, 2022).

Using German administrative data on jobseeker's benefit recipients, a study found that the effects of job placement services were relatively homogeneous for men but differed substantially for women depending on individual characteristics. More recently, in the United States, a study linking employment service data with tax records found that although employment services generate positive effects on average, the estimated effects vary considerably across individuals, with approximately 40 percent of the sample exhibiting negative estimated effects (Manoli, Michaelides and Patel, 2023).

Based on the estimated CATE, several studies have addressed the question of “to whom (which groups) a policy is effective,” that is, treatment assignment rules (Zhao *et al.*, 2012; Kitagawa and Tetenov, 2018; Athey and Wager, 2021). For example, Zhao *et al.* (2012) propose an approach in which individuals are assigned to the treatment group when the estimated policy effect takes a positive value, whereas individuals with a negative estimated effect are assigned to the control group, in which the treatment is not provided.

There are relatively few studies in the Korean context that analyze the heterogeneous effects of labor market policies. Existing studies include research examining differences in wage effects of vocational training (Keum, 2016), an analysis of the heterogeneity in the treatment effects of training programs (Kim, 2024), and a study investigating the employment effects of the youth internship program (Jeong, 2025). In addition, the *Employment Impact Evaluation of Fiscally Supported Programs* (Kim *et al.*, 2022) analyzed the heterogeneous effects of employment services by applying methods such as regression analysis, difference-in-differences (DiD), and propensity score matching. The study found that the net employment effects of employment service programs were positive for most programs, with the exception of the Psychological Stabiliza-

tion Program. The analysis also showed that differences in program effects by gender and age—namely, group average treatment effects—are not uniform and vary depending on the types of programs.²

Relative to these domestic studies, this study estimates individual-level treatment effect heterogeneity using causal machine learning rather than relying only on average effects or pre-specified subgroup comparisons. It then uses the estimated CATE to derive individualized treatment rules, thereby linking the estimation of heterogeneous effects to policy assignment.

3. ANALYTICAL FRAMEWORK

3.1. CONDITIONAL AVERAGE TREATMENT EFFECT

Let T be a binary variable, taking the value of 1 if an individual receives employment services and 0 otherwise. Y is the observed outcome, indicating 1 if an individual finds a job and 0 otherwise. Denote the potential outcome by $Y(t)$ when $T = t$, $t \in \{0, 1\}$. Then the relationship between Y and $Y(t)$ can be expressed as follows (Rubin, 1974).³

$$Y = TY(1) + (1 - T)Y(0), \quad T \in \{0, 1\}. \quad (1)$$

In equation (1), the individual treatment effect (ITE, τ) of employment services is defined as the difference in the potential outcomes when $t = 1$ and $t = 0$, that is, $\tau = Y(1) - Y(0)$.⁴ Note that the two potential outcomes $Y(1)$ and $Y(0)$ for a given individual cannot be observed simultaneously. In practice, therefore, the conditional average treatment effect (CATE) is constructed for the group sharing the same characteristics $X = x$, as in equation (2):

$$\text{CATE: } \tau(x) = \mathbb{E}[\tau \mid X = x] = \mathbb{E}[Y(1) - Y(0) \mid X = x]. \quad (2)$$

The ways to identify CATE fall generally into two approaches. The first approach constructs mutually exclusive subgroup samples stratified by specific characteristics $X = x$ and then estimates an ATE for each subgroup. The second

²However, a closer examination of the study reveals that the control group is not clearly defined, making it difficult to assert the findings as results of causal inference in a strict sense.

³Equation (1) implicitly assumes the Stable Unit Treatment Value Assumption (SUTVA) and the Conditional Independence Assumption (CIA).

⁴From a policy perspective, estimating the effects of expanded outreach or reduced administrative barriers on program participation and outcomes is particularly important. Under independence and homogeneous treatment effects, this intent-to-treat effect depends on the average treatment effect and on the effect of these policy efforts on program take-up.

approach specifies a high-dimensional relationship between the treatment variable T and covariates X in the outcome regression equation and then identifies CATE through the estimated coefficients.

Both approaches have statistical limitations. The former is vulnerable to the problem that CATE may be influenced by arbitrary choices by the researcher, raising questions about the objectivity and reliability of the estimated treatment effects.⁵ In addition, this approach is susceptible to the multiple-testing problem, which increases the probability of false-positive findings that treatment effects differ across groups when in fact they do not. In the latter approach, there is a substantial risk that bias may arise in the estimated effects when the estimation function is arbitrarily specified in a high-dimensional and complex form. Prior studies show that treatment effect estimates from observational data can be highly model-dependent. LaLonde (1986), for example, reported difference-in-differences estimates ranging from $-\$1,325$ to $\$629$ across specifications, relative to an experimental benchmark of $\$886$. Similarly, Ho *et al.* (2007) demonstrated that alternative covariate sets and functional forms can generate substantially different average treatment effect estimates, implying that causal conclusions may depend critically on model specification choices.⁶

Under the common assumptions, suppose that the potential outcome $Y(t)$ takes the following functional form when $X = x$ and $T = t$ (Künzel *et al.*, 2019):⁷

$$Y(t) = \mu(x, t) + \varepsilon(t), \quad t \in \{0, 1\}, \quad (3)$$

where $\mu(x, t) = \mathbb{E}[Y(t) \mid X = x, T = t]$ is a response function and $\varepsilon(t)$ is a random error.

From equation (3), ITE is given by $\tau(x) = \mu(x, 1) - \mu(x, 0)$. One can estimate $\hat{\mu}(x, t)$ from the sample $\{X, Y\}_{T=t}$, and then calculate $\hat{\tau}(x) = \hat{\mu}(x, 1) - \hat{\mu}(x, 0)$. In practice, $\hat{\mu}(x, t)$ is chosen to minimize the mean squared error (MSE), $\min \mathbb{E}[(\tau - \hat{\tau}(x))^2 \mid X = x]$, so that $\hat{\tau}(x)$ is as close as possible to τ .

To illustrate why ML is well suited for causal inference, consider a simple linear model, $\mu(x, t) = \beta_t X$. Let $\hat{\mu}(x, t) = \hat{\beta}_t X$ be the estimated response function from a least-squares method applied to the in-sample data $\{X, Y\}_{T=t}$. When one

⁵This problem is often referred to by terms such as “data snooping,” “p-hacking,” or “phishing.”

⁶For LaLonde (1986), the reported range is based on columns (6)–(9) of the PSID-3 row in Table 5. For Ho *et al.* (2007), see Figure 2, which illustrates the extent of model dependence across alternative specifications.

⁷The function $\mu(x, t)$ denotes the conditional expectation of the potential outcome given $X = x$ and treatment status, while $Y(t)$ follows the standard Rubin (1974) potential-outcomes notation.

predicts a counterfactual $\widehat{\beta}_t X$ in the out-of-sample data $\{X\}_{T=1-t}$, the prediction error is

$$\mathbb{E}_{X \in \{X\}_{1-t}} \left[\{\mu(x, t) - \widehat{\mu}(x, t)\}^2 \mid X = x \right] = \underbrace{\left[(\beta_t - \mathbb{E}[\widehat{\beta}_t])x \right]^2}_{\text{Bias}^2} + \underbrace{\text{Var}(\widehat{\beta}_t)x^2}_{\text{Variance}} \quad (4)$$

The first term in equation (4) represents bias, measuring how much $\widehat{\beta}_t$ deviates from the true parameter β_t on average when the in-sample data are repeatedly drawn. The second term represents variance, the extent to which $\widehat{\beta}_t$ varies across repeated samples.

Total error, which combines the effects of bias and variance, tends to increase as the functional form of a model becomes more complex (Hastie, Tibshirani and Friedman, 2009). In this context, ML is well suited for CATE estimation not only because it does not require the functional form to be specified *a priori*, but also because it can maintain relatively accurate predictions of a counterfactual outcome even when the model takes a complex form. ML methods balance the bias–variance trade-off via regularization in the process of minimizing MSE.

When applying ML methods, particular attention should be paid to two issues: covariate-distribution imbalance and regularization bias between treatment and control groups. It is well known that the distribution of covariates typically differs between the treatment and control groups, which may lead to biased estimates of the treatment effect. This study, as shown in equation (5), attempts to correct the covariate imbalance by applying weights to the outcome variables (Knaus, Lechner and Strittmatter, 2021).⁸

$$\begin{aligned} \widehat{D}_t^* &= \frac{\widehat{w}(x)[1 - \widehat{w}(x)]}{t - \widehat{w}(x)} \left[Y - \widehat{\mu}(x, t) \right. \\ &\quad \left. + (2t - 1)\{\widehat{\mu}(x, 1) - \widehat{\mu}(x, 0)\} \right], \\ \widehat{w}(x) &= \mathbb{P}(T = 1 \mid X). \end{aligned} \quad (5)$$

A common feature of observational data is that the size of the treatment group (n_1) is substantially smaller than that of the control group (n_0). In such

⁸If the propensity score is denoted by $e(x) = \mathbb{P}(T = 1 \mid X)$, overlap weights are defined as $[w_1(x), w_0(x)] = [1 - e(x), e(x)]$. The widely used inverse probability weights, $[w_1(x), w_0(x)] = [1/e(x), 1/\{1 - e(x)\}]$, have the problem that $\widehat{D}_{it}^*$ in equation (5) may diverge when $e(x)$ is very small or close to one. In contrast, overlap weights control the divergence of the transformed outcome while focusing on common support (Li, Thomas and Li, 2018; Zeng *et al.*, 2021).

cases, response functions estimated using ML may capture not only true treatment effect heterogeneity but also bias induced by differential regularization resulting from unequal sample sizes. To alleviate the regularization-bias problem, this study uses the X-learner, which is known to exhibit relatively small regularization bias among various meta-algorithms. Table 1 shows the procedures of the X-learner.⁹

3.2. INDIVIDUALIZED TREATMENT RULE

This section discusses how to assign treatment–individualized treatment rules (ITR)–based on the estimated CATE. A straightforward approach might involve inspecting CATE and selecting subgroups with comparatively large treatment effects. Such ex-post data mining, however, tends to yield non-reproducible results upon replication (Deaton, 2010) and lacks a principled targeting strategy grounded in identified treatment effects. Arbitrary ex-post subgroup adjustment may therefore undermine the reliability of the inferences.

This study derives the ITR from the value of the policy. Let Z denote the vector of policy variables,¹⁰ and define the treatment assignment rule as $R(Z) = \text{sign}[f(Z)]$. Specifically, if $R(Z) = 1$ (i.e., $f(Z) > 0$), an individual is assigned to the treatment group; conversely, if $R(Z) = -1$ (i.e., $f(Z) \leq 0$), the individual is assigned to the control group. For policymakers seeking to assign individuals with $\hat{\tau}(X) > 0$ to the treatment group while leaving those with $\hat{\tau}(X) \leq 0$ untreated, the objective is to select $R(Z) = \text{sign}[f(Z)]$ that minimizes the misallocation defined in equation (6).¹¹

$$\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n |\hat{\tau}(X_i)| \mathbf{1}\{L_i \neq R(Z_i)\}, \quad L_i = \text{sign}[\hat{\tau}(X_i)], \quad R(Z_i) = \text{sign}[f(Z_i)]. \quad (6)$$

The interpretation of equation (6) is straightforward. Consider an individual with $\hat{\tau}(X) > 0$, i.e., $L = 1$, and $f(Z) < 0$, i.e., $R(Z) = -1$. Thus, $\mathbf{1}\{L \neq R(Z)\} = 1$, causing the objective function to increase by $|\hat{\tau}(X)|$. In other words, misclassifying an individual with a positive treatment effect into the untreated group adversely affects the minimization problem, with the misclassification penalty

⁹ $\hat{\mu}(x, 1)$, estimated from a small sample, is heavily regularized, whereas $\hat{\mu}(x, 0)$, estimated from a relatively large sample, is mildly regularized (Künzel *et al.*, 2019).

¹⁰ Z differs from X in terms of legitimacy and non-gameability, as explained later.

¹¹For detailed discussions, see Zhao *et al.* (2012), Zhou *et al.* (2017), Kitagawa and Tetenov (2018), and Athey and Wager (2021). Most studies share a common framework in which the problem is formulated as one of minimizing classification error.

1.	Estimate weight for pseudo-outcomes $w(X) = \mathbb{P}(T = 1 \mid X)$
2a.	Estimate response functions Treatment: $\mu(x, 1) = \mathbb{E}[Y(1) \mid X = x]$ Control: $\mu(x, 0) = \mathbb{E}[Y(0) \mid X = x]$
2b.	Impute pseudo-outcomes Treatment: $D_1(Y(1), \hat{\mu}(x, 0), \hat{w}(x))$ Control: $D_0(Y(0), \hat{\mu}(x, 1), \hat{w}(x))$
2c.	Estimate two CATE functions $\tau_1(x) = \mathbb{E}[D_1 \mid X = x]$, for all i $\tau_0(x) = \mathbb{E}[D_0 \mid X = x]$, for all i
2d.	Estimate final weight $g(x) = \mathbb{P}(T = 1 \mid X = x)$
3.	Calculate CATE $\hat{\tau}(x) = \hat{g}(x)\hat{\tau}_0(x) + [1 - \hat{g}(x)]\hat{\tau}_1(x)$, for all i

Table 1: PROCEDURES OF THE X-LEARNER. This table summarizes response-function estimation, pseudo-outcome imputation, treatment-specific CATE estimation, and propensity-score-weighted aggregation, following the Appendix in Künzel *et al.* (2019).

given by the magnitude of the treatment effect. For simplicity, this study assumes a linear specification, $f(Z) = \alpha + Z'\beta$, and optimizes equation (6) via a Support Vector Machine (SVM).¹²

Let $R^*(Z)$ denote the optimal rule and define $Q^*(Z) = [R^*(Z) + 1]/2$. The value of the policy can then be expressed as an inverse-probability-weighted average (Zhao *et al.*, 2012):

$$V[R^*(Z)] = \mathbb{E}[Y(Q^*(Z))] = \mathbb{E}\left[\frac{\mathbf{1}\{T = Q^*(Z)\}}{\mathbb{P}(T \mid Z)}Y\right], \quad T, Q^*(Z) \in \{0, 1\}. \quad (7)$$

In equation (7), the indicator $\mathbf{1}\{T = Q^*(Z)\}$ takes the value 1 if the recommended assignment $Q^*(Z)$ coincides with realized treatment T , and 0 otherwise.

¹²In SVM estimation, equation (6) is converted to

$$\min_f \frac{1}{n} \sum_{i=1}^n \left[|\hat{\tau}(X_i)| \zeta \{1 - \text{sign}[\hat{\tau}(X_i)]f(Z_i)\} + \lambda \|f\|^2 \right],$$

where $\zeta(u) = \max(0, u)$; see Zhao *et al.* (2012).

If $Q^*(Z) = 1$ for all Z , equation (7) reduces to the policy value under “treat all.” Conversely, if $Q^*(Z) = 0$, it corresponds to the policy value under “treat none.”

4. EMPIRICAL ANALYSIS

4.1. DATA AND DESCRIPTIVE STATISTICS

The analysis utilizes data from program participants who received employment services through NESP and non-participants, spanning 2021 through 2022.

Implemented in January 2021, the NESP offers both employment services and income support to job seekers who are not covered by Employment Insurance. Eligibility is restricted to individuals aged 15–69 whose income and assets fall below specified thresholds: income below 60 percent of median income (120 percent for youth), and assets below KRW 400 million (KRW 500 million for youth). The program targets job seekers who are vulnerable in the labor market, such as unemployed youth and women experiencing career interruptions.

Following an in-depth consultation and employability assessment, participants establish an Individual Action Plan (IAP). They then receive intensive, customized job-placement assistance. Participants are required to undertake specified job-search activities, which are systematically monitored. Employment centers administer these services to program participants for up to one year. Participants who fulfill job-search activities in accordance with their IAP receive a Job Search Promotion Allowance of KRW 600,000 per month for up to six months.

Table 2 presents the NESP’s annual budget and number of participants during 2021–2026. In its first year, the program had a budget of approximately KRW 1.4 trillion and served roughly 400,000 participants. Following a budget

	2021	2022	2023	2024	2025	2026
Budget (million KRW)	1,402,965	1,496,351	1,225,455	942,515	845,709	1,012,830
Participants	399,855	262,313	272,915	312,773	350,000	–

Table 2: ANNUAL BUDGET AND PARTICIPANTS IN THE NESP. This table reports the annual program budget and participant counts for 2021–2026. Data are drawn from the Ministry of Employment and Labor’s *Budget and Fund Operation Plan, 2022–2026* (Ministry of Employment and Labor, 2022–2026) and *Performance Evaluation Report on Fiscally Supported Employment Programs* (Ministry of Employment and Labor, 2022–2025); the participant count for 2025 is projected.

expansion in 2022, funding contracted sequentially through 2025. At the same time, participant numbers began to recover after the sharp decline in 2022.

The analysis period overlaps with the outbreak of COVID-19. Since the pandemic substantially altered labor-market conditions (Aum, Lee and Shin, 2020; Lee and Yang, 2022), the effect of employment services on job placement may be conditional on the macroeconomic environment prevailing during this period. Therefore, the estimates are best interpreted as internally valid for the 2021–2022 policy environment, while their extrapolation to a non-pandemic labor market remains an issue of external validity.

Dataset construction proceeded as follows. The WorkNet Job Seeker Database was merged with the Employment Service Benefits Database to identify NESP participants ($n = 368,331$) and non-participants ($n = 7,199,920$). These records were then linked to the Employment Insurance Database to determine each individual's employment status based on eligibility information.¹³

For NESP participants, employment is defined as confirmed Employment Insurance eligibility at any point from job-search certification through six months post-program completion. For non-participants, it is defined as eligibility acquisition from job-search certification through six months after the search deadline.¹⁴

After excluding 2,223,575 observations with truncated employment-status information, the dataset comprised 5,344,676 participants and non-participants. From this, a 10 percent stratified random sample ($n = 534,468$) was drawn and subsequently split into three subsets: Train 1 ($n = 160,340$), Train 2 ($n = 160,340$), and Test ($n = 213,788$).

Table 3 presents the distribution of NESP participation and employment status in the test dataset. Among NESP participants, the employment rate is 68.23 percent, somewhat higher than the 62.77 percent observed among non-participants.¹⁵

Table 4 shows that participants and non-participants have substantial differ-

¹³From the Employment Service Benefits Database, cases classified under the following employment service categories were excluded: the Employment Success Package, SME Job Experience for Youth, Job Fair, Psychological Stabilization Program, and Group Counseling Program.

¹⁴Under Article 3 of the current Enforcement Decree of the Employment Insurance Act, short-hour workers—those working fewer than 60 hours per month or 15 hours per week—are exempted from mandatory employment insurance coverage. Therefore, even if an individual is actually employed, there may be no record of employment insurance eligibility, creating a potential risk of classifying that person as unemployed.

¹⁵When cases without confirmed Employment Insurance eligibility within the defined observation window were classified as non-employed and retained in the truncated sample, the total number of observations amounted to 7,568,251, with an employment rate of 58.8 percent among

National Employment Support Program ($n = 213,788$)		
	Non-participation	Participation
Unemployed	74,880 (37.2)	4,025 (31.8)
Employed	126,238 (62.8)	8,645 (68.2)
Total	201,118 (100.0)	12,670 (100.0)

Table 3: DISTRIBUTION BY NESP PARTICIPATION AND EMPLOYMENT STATUS. This table cross-tabulates employment status by NESP participation in the test sample ($n = 213,788$). Figures in parentheses are column percentages.

ences in some covariates but minor differences in others. The distributions of gender, academic major, and preferred contract type exhibit no substantial disparities between the groups. Desired wages are also similar, except for a slightly higher share of non-participants seeking wages over KRW 3 million.

Other variables display notable differences between the two groups. For example, the non-participant group exhibits a relatively higher proportion of individuals with a high-school education or less, whereas the participant group has a higher share of college graduates. The average age of the non-participant group is 43.1 years, approximately 8.6 years older than that of the participant group (34.5 years). The age distribution further reveals that individuals aged 45 and above account for nearly half of the non-participant group. In contrast, the participant group is characterized by a markedly higher proportion of younger individuals, particularly those under age 29.

Substantial differences are also observed in the distribution of work experience. Among participants, roughly 66 percent have no prior work experience or less than one year, whereas in the non-participant group, about 45 percent have three or more years of experience. With regard to preferred occupations, participants tend to favor research, engineering, and technical positions, while non-participants prefer occupations in beauty, travel, hospitality, security, cleaning, maintenance, and production. A majority of non-participants (56.9 percent) seek experienced positions, whereas most participants (71.1 percent) prefer entry-level positions. Finally, in terms of preferred workplace location, non-participants tend to prefer Gyeonggi Province, while participants prefer Seoul. However, when regions are broadly categorized into the Seoul metropolitan area

NESP participants. This rate is highly comparable to Kim *et al.* (2022), who report an employment rate of 59.9 percent over a similar period.

and non-metropolitan areas, the difference between the two groups is not substantial.

Covariate/category	Non-participation		Participation	
	<i>n</i>	(%)	<i>n</i>	(%)
Education				
Elementary	8,278	(4.1)	215	(1.7)
Middle	14,873	(7.4)	492	(3.9)
High	85,277	(42.4)	4,525	(35.7)
Some college	38,008	(18.9)	2,374	(18.7)
College	50,444	(25.1)	4,821	(38.1)
Graduate (master)	3,910	(1.9)	233	(1.8)
Graduate (doctorate)	328	(0.2)	10	(0.1)
Gender				
Men	87,449	(43.5)	5,647	(44.6)
Women	113,669	(56.5)	7,023	(55.4)
Major				
Not applicable	89,926	(44.7)	3,966	(31.3)
Humanities	12,333	(6.1)	1,077	(8.6)
Social science	31,620	(16.7)	2,505	(19.8)
Education	6,749	(3.4)	333	(2.8)
Engineering	31,281	(15.6)	2,429	(19.2)
Natural science	10,853	(5.4)	849	(6.7)
Medicine	5,620	(2.8)	369	(2.9)
Arts and sports	12,736	(6.3)	1,142	(9.0)
Preferred employment type				
Experienced	114,696	(57.0)	3,702	(29.2)
Entry-level	86,422	(43.0)	8,968	(70.8)
Preferred contract type				
Indefinite term	114,888	(57.1)	7,282	(57.5)
Fixed term	17,453	(8.7)	811	(6.4)
Any contract type	68,777	(34.2)	4,577	(36.1)
Wage expectation (monthly, KRW 10,000)				
Any amount	23,572	(11.7)	1,519	(12.0)
$w < 200$	52,144	(25.9)	3,142	(24.8)
$200 \leq w < 250$	68,515	(34.1)	4,705	(37.1)
$250 \leq w < 300$	30,072	(15.0)	2,287	(18.1)
$300 < w$	26,815	(13.3)	1,017	(8.0)
Work experience (years)				
None	78,373	(39.0)	8,046	(63.5)
$\text{exp} < 1$	6,786	(3.4)	359	(2.8)
$1 < \text{exp} \leq 2$	12,490	(6.2)	611	(4.8)
$2 < \text{exp} \leq 3$	11,550	(5.7)	512	(4.0)
$3 < \text{exp} \leq 5$	18,307	(9.1)	776	(6.1)
$5 < \text{exp} \leq 10$	28,374	(14.1)	1,024	(8.1)
$10 < \text{exp}$	45,238	(22.5)	1,342	(10.6)

Table 4: DESCRIPTIVE STATISTICS. This table compares the distributions of covariates by NESP participation. Numbers in parentheses are column percentages.

Covariate/category	Non-participation <i>n</i>	(%)	Participation <i>n</i>	(%)
Preferred region				
Any	6,607	(3.3)	1,057	(8.3)
Seoul	39,266	(19.5)	3,242	(25.6)
Busan	17,011	(8.5)	1,166	(9.2)
Daegu	9,332	(4.6)	677	(5.3)
Incheon	11,949	(5.9)	750	(5.9)
Gwangju	5,080	(2.5)	362	(2.9)
Daejeon	5,157	(2.6)	308	(2.4)
Ulsan	5,679	(2.8)	268	(2.1)
Sejong	994	(0.5)	46	(0.4)
Gyeonggi	46,671	(23.2)	1,958	(15.5)
Gangwon	5,322	(2.6)	289	(2.3)
Chungbuk	5,653	(2.8)	287	(2.3)
Chungnam	6,296	(3.1)	339	(2.7)
Jeonbuk	7,293	(3.6)	475	(3.7)
Jeonnam	5,019	(2.5)	226	(1.8)
Gyeongbuk	8,601	(4.3)	411	(3.2)
Gyeongnam	12,937	(6.4)	721	(5.7)
Jeju	2,251	(1.1)	88	(0.7)
Preferred occupation				
Management, administration, finance, and insurance	48,857	(24.3)	3,541	(27.9)
Research, engineering, and technicians	11,790	(5.9)	1,592	(12.6)
Education, legal, police, social welfare, firefighting, and military	15,939	(7.9)	754	(6.0)
Healthcare and medical	10,390	(5.2)	695	(5.5)
Arts, design, broadcasting, and sports	8,350	(4.2)	1,100	(8.7)
Beauty, travel, hospitality, food service, security, and cleaning	48,046	(23.9)	2,372	(18.7)
Sales, marketing, and transportation	16,716	(8.3)	893	(7.0)
Construction and mining	6,583	(3.3)	257	(2.0)
Installation, maintenance, and production	32,094	(16.0)	1,424	(11.2)
Agriculture, forestry, and fisheries	2,353	(1.2)	42	(0.3)
Age (years)				
Mean	43.1		34.5	
Standard deviation	14.0		12.8	
Minimum	15		15	
Maximum	70		69	

Table 4: DESCRIPTIVE STATISTICS (CONTINUED). This table compares the distributions of covariates by NESP participation. Numbers in parentheses are column percentages.

4.2. TREATMENT EFFECT HETEROGENEITY

CATE, $\tau(X)$, was estimated by applying the X-learner procedures specified in Table 1, where X is the vector of variables in Table 4. Propensity scores were estimated using logistic regression with five-fold cross-validation, while the response functions were estimated by gradient boosting (LightGBM). Does heterogeneity exist in the effects of employment services? Panel (a) of Figure 1 confirms this by revealing a wide range of CATE, from -0.5796 to 0.4607.¹⁶¹⁷

Manoli, Michaelides and Patel (2023) report that a substantial share (about 40 percent) of the sample exhibited negative treatment effects in their study of a U.S. job-search assistance program. They argue that these effects may reflect changes in individuals' subjective beliefs about the relative productivity of search strategies, which can lead some participants to switch to less effective approaches. This job-search strategy mechanism is consistent with the view that

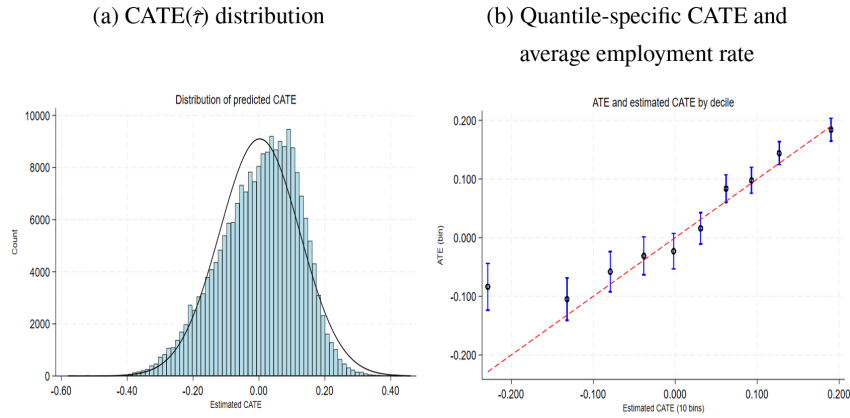


Figure 1: HETEROGENEOUS EFFECTS OF EMPLOYMENT SERVICES. Estimated CATEs ($\hat{\tau}$). Panel (a) shows the distribution of predicted CATEs. Panel (b) plots decile-specific empirical treatment effects against average predicted CATEs; vertical bars denote confidence intervals, and the dashed 45-degree line indicates equality.

¹⁶Previous studies comparing results across various ML methods have reported that differences in the results are generally not substantial but moderate (Künzel *et al.*, 2019).

¹⁷Variables such as preferred employment type and preferred contract type may be susceptible to endogeneity concerns. Therefore, I re-estimated an alternative CATE specification excluding these preference-related variables. The main findings are robust to this alternative specification, as shown in Appendix 4.



Figure 2: DISTRIBUTION OF INDIVIDUAL CHARACTERISTICS BY CATE QUINTILE. The heatmaps report within-quintile percentage distributions of education, age, work experience, and gender. Quintile 1 contains the lowest predicted CATEs, and quintile 5 contains the highest.

employment services affect not only the intensity of job search but also the allocation of search effort across different strategies.¹⁸

Panel (b) of Figure 1 partitions the test data into CATE-based deciles (estimated from the training sample) and plots the average treatment effect for each group. If the CATE captures meaningful heterogeneity, higher deciles should exhibit larger outcome differences between participants and non-participants.

¹⁸Evidence of negative effects from labor market programs is common. For example, Mascolo *et al.* (2024) found negative employment effects of the Basic Course (job application training) program in Switzerland using causal machine learning.

Consistent with the ideal pattern, average effects increase across deciles (Athey, Keleher and Spiess, 2025).

Figure 2 visualizes covariate distributions across CATE quintiles, from the bottom 20 percent (Group 1) to the top 20 percent (Group 5). Higher quintiles show larger shares of highly educated, younger, and less experienced individuals, and women, whereas lower quintiles are dominated by less-educated, older, and more experienced individuals, and men.¹⁹

4.3. INDIVIDUALIZED TREATMENT RULE

The covariates X used in estimating CATE may overlap with policy variables Z , but need not be identical. Suitability as a policy variable hinges on two key criteria:

- *Legitimacy*: Although individual characteristics often strongly influence CATE, their use for policy targeting warrants caution due to anti-discrimination principles.
- *Non-gameability*: Variables prone to low-cost manipulation—such as wage expectations—are unsuitable, as individuals could strategically self-select into or out of eligibility.

Age, gender, work experience, and education are selected as policy variables. Gender and age were chosen because employment service programs target women to boost female employment rates and incorporate age-based eligibility criteria (e.g., youth exemptions). Although not explicitly specified in the current program, work experience and education exhibit distinct distributional differences across CATE deciles and are thus included.

The ITR, $R(Z) = \text{sign}[f(Z)]$, where $f(Z) = \alpha + Z'\hat{\beta}$, is estimated via SVM, and the results are reported in Table 5. A policy variable Z_k with $\hat{\beta}_k > 0$ helps assign individuals to treatment, *ceteris paribus*. Based on Model 3, which includes all policy variables, several findings merit attention. First, age is negatively associated with treatment assignment. Relative to the reference group of individuals aged 15–29, the estimated coefficients for all older age groups, $\hat{\beta}_{\text{Age}}$, are negative, indicating that the probability of being assigned to the untreated group increases monotonically with age.

Second, work experience is associated with a high probability of untreated

¹⁹The patterns align with the findings of Knaus, Lechner and Strittmatter (2022), Martin and Grubb (2001), Pastore (2019), and others.

assignment ($\hat{\beta}_{\text{Career}} < 0$). In particular, individuals with moderate experience (5–9 years) exhibit the strongest negative effect on treatment assignment.

Third, higher educational attainment is generally associated with a greater likelihood of treatment assignment ($\hat{\beta}_{\text{Edu}} > 0$). In particular, having a college degree or higher appears to increase the probability of allocation to treatment. For highly educated individuals, the ITR suggests that employment services could enhance job-search productivity through professional counseling, ongoing monitoring of search strategies, and résumé and interview coaching, rather than addressing skill deficiencies. These services also reduce search frictions by motivating sustained effort, suggesting potential for further improvements in program effectiveness (Crépon *et al.*, 2013).

Fourth, being female appears to increase the probability of treatment assignment ($\hat{\beta}_{\text{Sex}} > 0$). Bergemann and van den Berg (2008) and Knaus, Lechner and Strittmatter (2022) provide evidence that employment services benefit individuals with fewer employment opportunities and weaker labor market attachment, such as women, supporting the results of the ITR. Overall, these findings are broadly consistent with previous studies identifying characteristics associated with more favorable employment outcomes from employment services (Pastore, 2019; Knaus, Lechner and Strittmatter, 2022).

Table 5 shows that the ITR using policy variables raises the expected employment rate, $\mathbb{E}[Y(Q^*(Z))]$, by approximately 1.9–2.0 percentage points in Model 2 and 1.3–1.6 percentage points in Model 3, relative to the benchmarks of treat all or treat none.

Figure 3 illustrates the ITR implied by the results in Table 5. For the youth group (aged 15–29), most individuals—except low-educated men—qualify for NESP employment services (“Treat”). In particular, those with little work experience (less than five years) are assigned to the policy target group regardless of gender or education. In contrast, among older individuals (aged 50–59), only highly educated women with short work experience are assigned treatment.

Meanwhile, the ITR for middle-aged and older individuals varies. Those with 5–9 years of work experience are classified as untreated, regardless of gender or education, whereas individuals with no work experience or short experience (less than five years) are assigned to treatment or non-treatment depending on their gender and education. For example, women are classified as eligible for treatment regardless of educational attainment, while among men, those with a high-school diploma or less are classified as untreated.

In summary, the ITR favors untreated status for older individuals, those with longer work experience, and men, while assigning treatment to relatively

younger individuals, those with little or short experience, women, and highly educated groups.

	Model 1	Model 2	Model 3
Age (reference: 15–29 years)			
$\hat{\beta}_{\text{Age}}$ 30–39	0.01830 (0.00728)	-0.45048 (0.00667)	-0.49911 (0.00754)
40–49	-0.05527 (0.00689)	-0.49860 (0.00649)	-0.46182 (0.00761)
50–59	-0.55253 (0.00630)	-1.02318 (0.00671)	-0.88403 (0.00774)
60+	-0.08609 (0.01614)	-1.08986 (0.00715)	-0.93387 (0.00848)
Experience (reference: less than 5 years)			
$\hat{\beta}_{\text{Career}}$ 5–9		-0.67729 (0.00648)	-0.69871 (0.00637)
10+		-0.22311 (0.00658)	-0.25620 (0.00633)
Gender (reference: men)			
$\hat{\beta}_{\text{Sex}}$ Women	0.31458 (0.00524)	0.25381 (0.00464)	0.25960 (0.00506)
Education (reference: less than high school)			
$\hat{\beta}_{\text{Edu}}$ High			-0.11080 (0.00839)
Some college			0.39723 (0.00959)
College+			0.39521 (0.00959)
Constant	-0.03639 (0.00501)	0.63460 (0.00427)	0.44854 (0.00957)
Expected employment rate			
Treat all	0.65058	0.62814	0.63185
Treat none	0.62830	0.62871	0.62877
$\mathbb{E}[Y(Q^*(Z))]$	0.64724	0.64771	0.64450
[95% CI]	[0.64102, 0.65362]	[0.64128, 0.65374]	[0.63768, 0.65106]

Table 5: INDIVIDUALIZED TREATMENT RULES ESTIMATED BY A LINEAR SUPPORT VECTOR MACHINE (SVM). This table reports SVM coefficients and expected employment rates under treat-all, treat-none, and estimated-rule policies for three model specifications. Bootstrap standard deviations and 95 percent confidence intervals are based on 1,000 replications; $n = 213,788$.

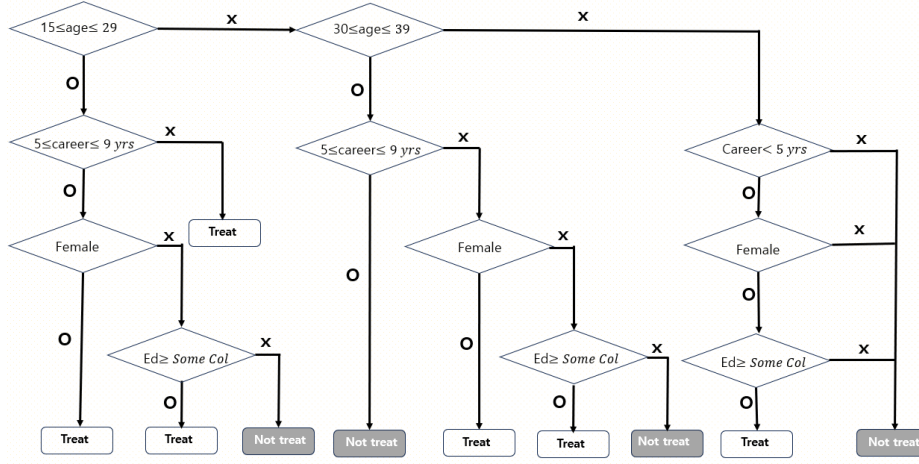


Figure 3: TREATMENT ASSIGNMENT TREE. This figure translates the estimated individualized treatment rule into “Treat” and “Not treat” decisions by age, work experience, gender, and education. Circles indicate that a branch condition is satisfied, whereas crosses indicate that it is not satisfied.

5. CONCLUSION

This study applies causal machine learning—a machine-learning-based causal inference framework—to estimate the heterogeneous treatment effects of the National Employment Support Program. Specifically, it uses the X-learner meta-algorithm to estimate the conditional average treatment effect (CATE), addressing the sample imbalance between treatment and control groups, and attempts to derive an individualized treatment rule (ITR).

The main findings are summarized as follows. First, the effects of employment services exhibit substantial heterogeneity across individual characteristics. Estimated CATE spans both negative and positive values, with a slightly heavier left tail than the normal distribution. Moreover, the difference in treatment effects between the upper and lower deciles is statistically significant, providing clear evidence of heterogeneity.

Second, treatment effects vary with individual characteristics and labor market activities. The effects of the employment service program are positive for younger individuals, women, and those with higher levels of education and shorter work experience. In contrast, for middle-aged individuals, men, and those with relatively greater work experience, the effects are either statistically

insignificant or even negative. These findings are consistent with evidence reported in international studies.

Third, the ITR indicates that selective provision of services based on policy variables is more efficient. When optimal assignment based on age, work experience, gender, and education is applied, the expected employment rate may increase by approximately one to two percentage points.

This study contributes by providing empirical evidence for a shift from one-size-fits-all employment services toward data-driven, customized policy design. It is desirable that future employment service policies focus on enhancing the efficiency of labor market interventions by moving beyond mere quantitative expansion of services. In particular, policies should aim to identify program effects while accounting for the characteristics of target groups and, based on these insights, redesign service content or provide alternative types of programs tailored to differential effectiveness.

REFERENCES

- Athey, S. (2018). “The impact of machine learning on economics,” in Agrawal, A., Gans, J. and Goldfarb, A. (eds.), *The economics of artificial intelligence: An agenda*, NBER.
- Athey, S. and Imbens, G. W. (2019). “Machine learning methods economists should know about,” arXiv:1903.10075.
- Athey, S. and Wager, S. (2021). “Policy learning with observational data,” *Econometrica* 89, 122–161.
- Athey, S., Keleher, N. and Spiess, J. (2025). “Machine learning who to nudge: Causal vs. predictive targeting in a field experiment on student financial aid renewal,” *Journal of Econometrics* 249, 105945.
- Aum, S., Lee, S. Y. and Shin, Y. (2020). “COVID-19 doesn’t need lockdowns to destroy jobs: The effect of local outbreaks in Korea,” NBER Working Paper No. 27264.
- Bergemann, A. and van den Berg, G. J. (2008). “Active labor market policy effects for women in Europe—A survey,” *Annales d’Économie et de Statistique* 91/92, 385–408.
- Card, D., Kluve, J. and Weber, A. (2010). “Active labour market policy evaluations: A meta-analysis,” *Economic Journal* 120, F452–F477.

- Card, D., Kluve, J. and Weber, A. (2018). "What works? A meta-analysis of recent active labor market program evaluations," *Journal of the European Economic Association* 16, 894–931.
- Cockx, B., Lechner, M. and Bollens, J. (2019). "Priority to unemployed immigrants? A causal machine learning evaluation of training in Belgium," IZA Discussion Paper No. 12875.
- Crépon, B., Duflo, E., Gurgand, M., Rathelot, R. and Zamora, P. (2013). "Do labor market policies have displacement effects? Evidence from a clustered randomized experiment," *Quarterly Journal of Economics* 128, 531–580.
- Crépon, B. and van den Berg, G. J. (2016). "Active labor market policies," IZA Discussion Paper No. 10321.
- Curth, A. and van der Schaar, M. (2021). "Nonparametric estimation of heterogeneous treatment effects: From theory to learning algorithms," arXiv:2101.10943.
- Deaton, A. (2010). "Instruments, randomization, and learning about development," *Journal of Economic Literature* 48, 424–455.
- Gerfin, M. and Lechner, M. (2001). "A microeconomic evaluation of the active labour market policy in Switzerland," CEPR Discussion Paper No. 2993.
- Graversen, B. K. and van Ours, J. C. (2008). "How to help unemployed find jobs quickly: Experimental evidence from a mandatory activation program," *Journal of Public Economics* 92, 2020–2035.
- Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction*, 2nd ed., Springer, New York.
- Ho, D. E., Imai, K., King, G. and Stuart, E. A. (2007). "Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference," *Political Analysis* 15, 199–236.
- Jeong, Y. (2025). "Heterogeneous effects of the youth employment internship program on labor market outcomes of college graduates: An X-learner-based treatment effect estimation," *KLI Quarterly Journal of Labor Policy* 25, 99–129.

- Keum, J. (2016). “A study on the wage effect of vocational training by training supplier,” *Journal of Korean Economic Studies* 34, 121–151.
- Kim, K.-A., Kim, K.-D., Jeon, W., Kim, H.-A., Park, S.-G. and Shin, U. (2022). *Research on improving methodologies to measure the effect of employment services*, Korea Labor Institute.
- Kim, Y. (2024). “New direction of evaluation on labor market programs: Application of causal machine learning to vocational training in Korea,” *Korean Economic Forum* 17, 47–72.
- Kitagawa, T. and Tetenov, A. (2018). “Who should be treated? Empirical welfare maximization methods for treatment choice,” *Econometrica* 86, 591–616.
- Knaus, M. C. (2022). “Double machine learning-based program evaluation under unconfoundedness,” arXiv:2003.03191v5.
- Knaus, M. C., Lechner, M. and Strittmatter, A. (2021). “Machine learning estimation of heterogeneous causal effects: Empirical Monte Carlo evidence,” *Econometrics Journal* 24, 134–161.
- Knaus, M. C., Lechner, M. and Strittmatter, A. (2022). “Heterogeneous employment effects of job search programs: A machine learning approach,” *Journal of Human Resources* 57, 597–636.
- Kreif, N. and DiazOrdaz, K. (2019). “Machine learning in policy evaluation: New tools for causal inference,” arXiv:1903.00402v1.
- Künzel, S. R., Sekhon, J. S., Bickel, P. J. and Yu, B. (2019). “Metalearners for estimating heterogeneous treatment effects using machine learning,” *Proceedings of the National Academy of Sciences* 116, 4156–4165.
- Lalive, R., van Ours, J. C. and Zweimüller, J. (2008). “The impact of active labour market programmes on the duration of unemployment in Switzerland,” *Economic Journal* 118, 235–257.
- LaLonde, R. J. (1986). “Evaluating the econometric evaluations of training programs with experimental data,” *American Economic Review* 76, 604–620.
- Lechner, M. (2019). “Modified causal forests for estimating heterogeneous causal effects,” arXiv:1812.09487.

- Lee, J. and Yang, H.-S. (2022). "Pandemic and employment: Evidence from COVID-19 in South Korea," *Journal of Asian Economics* 78, 101432.
- Li, F., Thomas, L. E. and Li, F. (2018). "Addressing extreme propensity scores via the overlap weights," *American Journal of Epidemiology* 188, 250–257.
- Manoli, D. S., Michaelides, M. and Patel, A. (2023). "Long-term and heterogeneous effects of job-search assistance," NBER Working Paper No. 24422.
- Martin, J. P. and Grubb, D. (2001). "What works for whom: A review of OECD countries' experiences with active labour market policies," *Swedish Economic Policy Review* 8, 9–56.
- Mascolo, F., Bearth, N., Muny, F., Lechner, M. and Mareckova, J. (2024). "The heterogeneous effects of active labour market policies in Switzerland," arXiv:2410.23322.
- Ministry of Employment and Labor (2022–2025). *Performance evaluation report on fiscally supported employment programs*.
- Ministry of Employment and Labor (2022–2026). *Budget and fund operation plan*.
- Mühlböck, M., Kalleitner, F., Steiber, N. and Kittel, B. (2022). "Information, reflection, and successful job search: A labor market policy experiment," *Social Policy & Administration* 56, 48–72.
- OECD/European Commission (2025). *Counterfactual impact evaluations of active labour market policies: Lessons from using linked administrative data*, OECD Publishing, Paris.
- Pastore, F. (2019). "The quasi-market of employment services in Italy," IZA Discussion Paper No. 12662.
- Rubin, D. B. (1974). "Estimating causal effects of treatments in randomized and nonrandomized studies," *Journal of Educational Psychology* 66, 688–701.
- 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.
- Zeng, S., Li, F., Wang, R. and Li, F. (2021). "Propensity score weighting for covariate adjustment in randomized clinical trials," *Statistics in Medicine* 40, 842–858.

Zhao, Y., Zeng, D., Rush, A. J. and Kosorok, M. R. (2012). “Estimating individualized treatment rules using outcome weighted learning,” *Journal of the American Statistical Association* 107, 1106–1118.

Zhou, X., Meyer-Hamblett, N., Khan, U. and Kosorok, M. R. (2017). “Residual weighted learning for estimating individualized treatment rules,” *Journal of the American Statistical Association* 112, 169–187.

APPENDIX 1. PROCEDURE OF THE X-LEARNER

Figure A1 illustrates the X-learner algorithm. It proceeds in five steps. First, the response functions for treated (μ_1) and untreated (μ_0) individuals are estimated separately. Second, missing potential outcomes are imputed using the estimated response functions. Third, pseudo-treatment effects are constructed separately for treated and untreated individuals (D_1^*, D_0^*). Fourth, these pseudo-outcomes are used to estimate two CATE functions ($\hat{\tau}_1(X), \hat{\tau}_0(X)$). Finally, the two CATE estimates are combined using propensity-score-based weights.

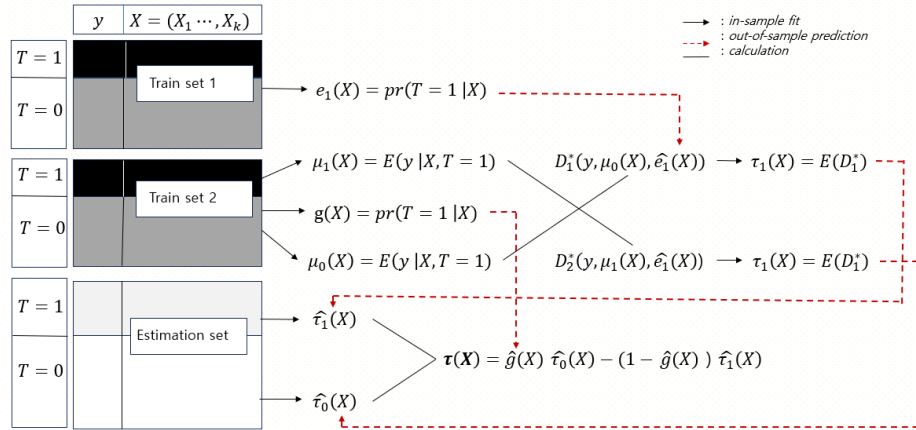


Figure A1: SCHEMATIC ILLUSTRATION OF CATE ESTIMATION WITH THE X-LEARNER. This figure summarizes sample splitting, response-function and propensity-score estimation, pseudo-outcome construction, and propensity-score-weighted aggregation of treatment-specific CATE estimates. Adapted from the Appendix in Künzel *et al.* (2019).

APPENDIX 2. DATA ATTRITION

Around 2.22 million out of 7.5 million observations are excluded due to a lack of information on employment status. This attrition may raise concern about the validity of results from the remaining sample if attrition exhibits systematic covariate differences. In response, the covariate distributions of the attrition sample and the remaining sample are compared and found to be very similar (Figure A2).



Figure A2: COMPARISON OF COVARIATE DISTRIBUTIONS BETWEEN TRUNCATED AND NON-TRUNCATED SAMPLES. Each panel compares the percentage distribution of a covariate within NESP participant and non-participant groups. The similar distributions indicate that incomplete employment-status follow-up does not materially alter observed sample composition.

APPENDIX 3. COMPARISON OF SUBGROUP ATEs AND AVERAGE CATES BY SUBGROUP

Let $G = g$ denote subgroup g . A subgroup average treatment effect ATE_g is the observed treated–control outcome difference and is calculated by

$$ATE_g = \frac{1}{N_{g,T=1}} \sum_{i:G_i=g} \mathbf{1}(T_i = 1)Y_i - \frac{1}{N_{g,T=0}} \sum_{i:G_i=g} \mathbf{1}(T_i = 0)Y_i,$$

where $N_{g,T=1}$ and $N_{g,T=0}$ are the numbers of observations with $T = 1$ and $T = 0$ in group g , respectively. Define the average of CATEs within subgroup g as

$$\bar{\tau}_g = \frac{1}{N_g} \sum_{i:G_i=g} \tau(X_i), \quad \tau(X) = \mathbb{E}[Y(1) - Y(0) | X],$$

where N_g is the number of observations in group g .

The results are broadly consistent with the main findings: the program is more effective for college graduates, younger individuals, and those with shorter work experience. Although subgroup ATEs are positive for both genders, the average estimated CATE is negative for men and positive for women. This discrep-

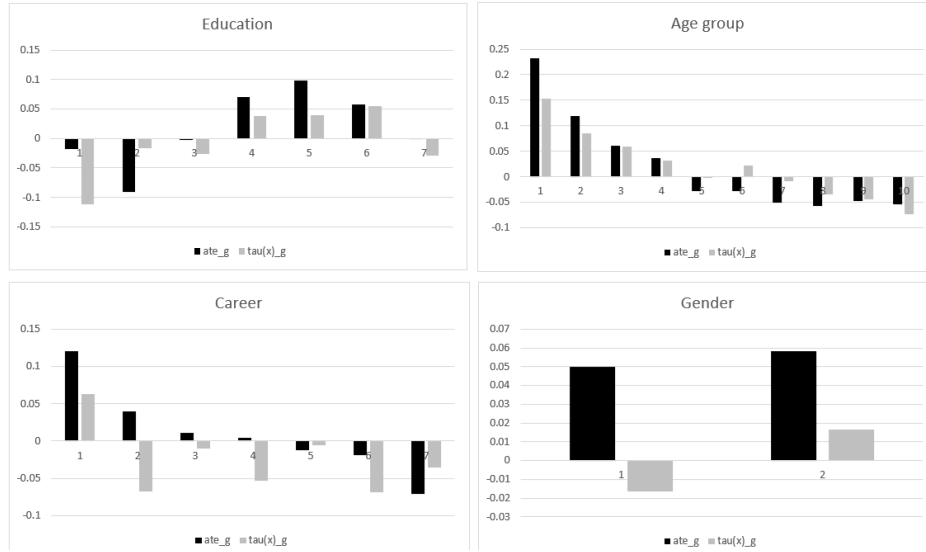


Figure A3: SUBGROUP ATEs AND AVERAGE CATES BY SUBGROUP. The panels compare raw subgroup treated–control employment-rate differences with mean estimated CATEs across education, age, work-experience, and gender categories. Differences between the two measures may reflect covariate composition within subgroups.

ancy may reflect compositional differences, as treated men are younger, better educated, and less experienced than untreated men.

APPENDIX 4. RE-ESTIMATION OF CATE EXCLUDING PREFERENCE-RELATED VARIABLES

Variables such as preferred employment type and preferred contract type may raise endogeneity concerns. An alternative specification without these preference-related variables is estimated, and the results are presented in Figures A4 and A5. The distribution of the estimated CATE in Figure A4 remains heterogeneous and resembles Figure 1. Moreover, differences in the mean, standard deviation, minimum, and maximum of the two distributions are negligible.

Figure A5 shows the CATE quintile distributions excluding preference-related variables. The results are similar to those in Figure 2.

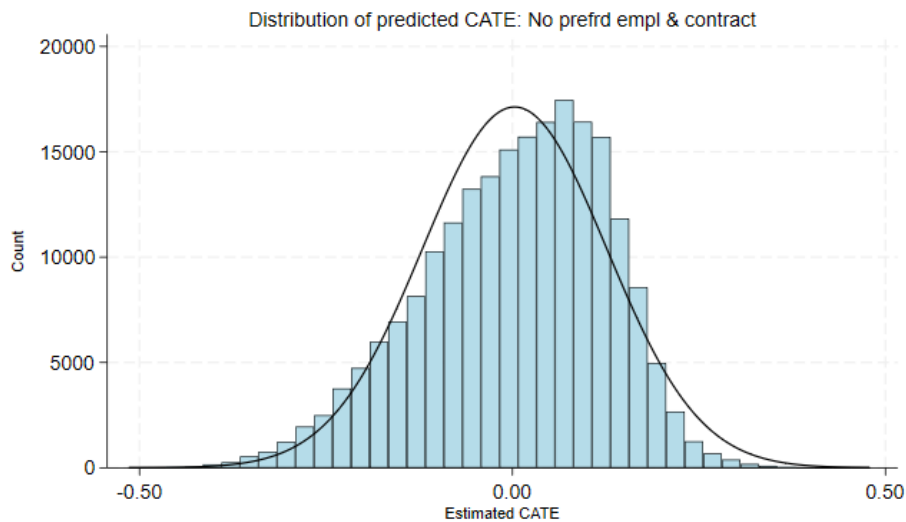


Figure A4: CATE ESTIMATION EXCLUDING PREFERRED EMPLOYMENT AND CONTRACT TYPE. This figure shows the predicted CATE distribution from a robustness specification that omits preference-related variables; its shape and summary statistics remain close to those in Figure 1.



Figure A5: DISTRIBUTION OF INDIVIDUAL CHARACTERISTICS BY CATE QUINTILE, EXCLUDING PREFERENCE-RELATED VARIABLES. The heatmaps report within-quintile percentage distributions of education, age, work experience, and gender under the robustness specification.