Statistica Sinica Preprint No: SS-2025-0267	
Title	Semi-Parametric Estimation of Potential Outcome Distributions and General Causal Estimands by Borrowing Information from Both Treatments and Controls
Manuscript ID	SS-2025-0267
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202025.0267
Complete List of Authors	Manli Cheng and Yukun Liu
Corresponding Authors	Yukun Liu
E-mails	ykliu@sfs.ecnu.edu.cn
Notice: Accepted author version.	

Statistica Sinica

Semi-parametric estimation of potential outcome distributions and general causal estimands by borrowing information from both treatments and controls

Manli Cheng[1,2] and Yukun Liu[2]

[1] *University of Waterloo and* [2] *East China Normal University*

Abstract: Estimation of treatment effects is of great importance and has drawn much attention in various areas. In the potential outcomes framework, since potential outcomes refer to the same characteristic under different treatment assignments, they usually share certain similarities. Such similarities are largely overlooked in the literature, leading to potential efficiency loss. In this paper, we introduce a semi-parametric proportional likelihood ratio model (SPLRM) to jointly model the conditional distributions of potential outcomes through a shared baseline distribution, fully utilizing information from both treatment and control groups. We estimate all underlying parameters and general causal estimands by maximum empirical likelihood estimation. An iterative empirical likelihood algorithm is developed for parameter estimation, and a simple likelihood ratio test is introduced to assess the distributional treatment effect. We show that the proposed estimators for various treatment effects are asymptotically normal, and the likelihood ratio test statistic follows asymptotically a central Chi-square distribution when there is no distributional treatment effect. This approach improves efficiency and robustness compared to traditional methods that separately estimate treatment effects. Simulations and an analysis of the National Supported Work Demonstration dataset demonstrate the practical applicability and advantages of the pro-

*Author's ORCID ID:
Yukun Liu: orcid.org/0000-0002-9743-9276

posed SPLRM-based method.

Key words and phrases: Causal inference, Distributional treatment effect, Likelihood ratio test, Semiparametric proportional likelihood ratio model.

1. Introduction

Accurately estimating treatment effects in observational studies, where treatment assignment is not randomized, is a fundamental task in empirical research across fields such as social sciences, educational policy research, clinical epidemiology, and econometrics. In observational studies, the absence of randomization frequently introduces confounding bias due to systematic differences in baseline covariates between treated and control groups. This makes it impossible to identify treatment effects without assumptions, let alone estimate them.

We adopt the Neyman–Rubin potential outcomes framework (Neyman, 1990; Rubin, 1974) to study treatment effects with a binary treatment variable. Let X denote a vector of covariates, D denote the treatment status with $D = 1$ a treatment and $D = 0$ a control, and $Y(0)$ and $Y(1)$ denote potential outcomes. To identify causal effects, a commonly-used assumption is the unconfoundedness assumption (Rosenbaum and Rubin, 1983), i.e. given X, the treatment assignment D is independent of the potential outcomes $Y(0)$ and $Y(1)$. Under this assumption, standard methods such as regression adjustment, inverse probability weighting (IPW), augmented IPW (AIPW), and matching can be used to estimate causal effects (Imbens and Rubin,

2015). In particular, the AIPW estimator is appealing due to its double robustness property: it remains consistent if either the propensity score model (Rosenbaum and Rubin, 1983) or the outcome model is correctly specified (Bang and Robins, 2005; Lunceford and Davidian, 2004). Most research focuses on estimating mean treatment effects, such as the average treatment effect (ATE), average treatment effect on the treated (ATT), average treatment effect on controls (ATC) (e.g., Hirano et al., 2003,?; Chan et al., 2016; Cattaneo, 2010; Yu et al., 2022), and quantile treatment effects (Firpo, 2007; Frölich and Melly, 2013, QTE), which assess treatment effects at specific quantiles of the outcome distribution and provide additional insights beyond mean effects.

A more comprehensive evaluation of treatment effects requires estimating the full distribution of potential outcomes, although there is limited research on this topic. Chernozhukov et al. (2013) proposed using conditional quantile regression or distribution regression to estimate the conditional distribution of potential outcomes given covariates, and developed a general asymptotic theory for distributional and quantile effects. Donald and Hsu (2014) introduced IPW estimators for the distribution of potential outcomes and obtained quantile functions through inverse mapping in binary treatment settings. Ai et al. (2022) extended this framework to continuous treatments using a weighted regression approach, where the weights are estimated via covariate balancing conditions.

It is worth noting that potential outcomes generally refer to the same measur-

able characteristic of an individual under different treatment states or intervention assignments. For example, when examining the effect of job training, an individual's income with and without training represents two distinct potential outcomes of the same individual. Thus, they usually share certain similarities, and taking such similarities into consideration can lead to improved estimation of treatment effects. However, the aforementioned existing estimation approaches to causal effects do not make use of such information and hence may have potential efficiency loss.

In this paper, we propose to model the conditional distributions of $Y(0)$ and $Y(1)$ given X by two semi-parametric proportional likelihood ratio models (SPLRMs) (Luo and Tsai, 2012) with a common, unknown baseline distribution. The unknown baseline distribution makes this model assumption semiparametric and hence has enough flexibility. Meanwhile, the common baseline distribution makes it convenient to borrow information across treatment groups to improve treatment effects estimation. We estimate all underlying parameters and both the conditional distributions of $Y(0)$ and $Y(1)$ given X by maximum likelihood estimation. With these distribution estimates, conventional causal estimands like ATE, ATT, ATC, and QTEs can all be conveniently estimated. We provide an efficient iteration algorithm, which is a modified version of Luo and Tsai (2012)'s algorithm for the one-sample case, to conveniently calculate our maximum likelihood estimators (MLEs). Theoretically, we establish the asymptotic normality of our MLEs for both the conditional and marginal distributions of the potential outcomes, and various popular causal estimands.

Under our SPLRM framework, it is convenient to construct a likelihood ratio test (LRT) statistic for assessing whether the distributional treatment effect exists. We show that the LRT statistic asymptotically follows a central Chi-squared distribution when there is no distributional treatment effect. This leads to a practically useful and computationally efficient test method for distributional treatment effect. In contrast, conventional Kolmogorov-Smirnov-type tests (e.g., Maier, 2011; Donald and Hsu, 2014) are not the case. Numerical simulations and a real application are provided to show the advantages of the proposed methods.

The remainder of this paper proceeds as follows. In Section 2, after introducing the SPLRM model setup and identification conditions, we present the proposed estimation procedures for the underlying parameters and investigate their asymptotic properties. In Section 3, we present the proposed MLEs for several causal quantities and establish their theoretical properties. We also construct the LRT for assessing the distributional treatment effect. Simulation studies and a real data application are provided in Sections 4 and 5, respectively. All technical proofs and additional simulation results are postponed to the Supplementary Material.

2. Methodology

We adopt the standard assumptions from the causal inference literature to identify causal effects:

Assumption 1. *(i) (Stable Unit Treatment Value) The potential outcomes for any*

individual do not vary with the treatments assigned to other individuals. (ii) (Unconfoundedness) The treatment indicator D is conditionally independent of the potential outcomes $(Y(0), Y(1))$ given X, i.e. $(Y(0), Y(1)) \perp D|X$. (iii) (Overlap) $0 < \text{pr}(D = 1|X) < 1$ *for any* X.

Under Assumption 1(i), the observed outcome for an individual is $Y = D \cdot Y(1) + (1-D) \cdot Y(0)$. Assumption 1(ii) is the commonly-used unconfoundedness assumption (Rosenbaum and Rubin, 1983). Assumption 1(iii) is used to guarantee that there are enough observations in each treatment group.

2.1 Model framework

Suppose $(Y_i, X_i, D_i), i = 1, 2, \ldots, n$, are n independent and identically distributed (i.i.d.) observations from (Y, X, D). For convenience, we re-write the observed data within each treatment group as $\{(Y_{ki}, X_{ki})\}_{i=1}^{n_k}$ for $k = 0, 1$, where n_k denotes the number of units with $D_i = k$ and $n = n_0 + n_1$. Let $F_{(k)}(y)$ denote the marginal distribution of $Y(k)$, and let $F_{(k)}(y \mid x)$ denote its conditional distribution given the covariate $X = x$. Their corresponding density functions are denoted by $f_{(k)}(y)$ and $f_{(k)}(y \mid x)$, respectively.

As we mentioned in the introduction, $Y(0)$ and $Y(1)$ often have certain similarities. To capture such similarities, we assume that their conditional distributions satisfy SPLRMs (Luo and Tsai, 2012; Cheng et al., 2025) with a common baseline

distribution:

$$dF_{(k)}(y|x) = \frac{\exp(y \cdot \beta_k^\top x) dF(y)}{\int \exp(t \cdot \beta_k^\top x) dF(t)}, \quad k \in \{0, 1\}, \tag{2.1}$$

where the vector-valued parameter $\beta_k \in \mathbb{R}^d$ and the baseline "carrier density" $F(y)$ are left unspecified. The SPLRM in (2.1) demonstrates considerable versatility by subsuming many widely-used models as special cases, such as generalized linear models (McCullagh and Nelder, 1989, GLMs), the exponential tilt regression model, the density ratio model (Qin, 1998), and the semiparametric odds ratio model (Chen, 2004). Under model (2.1), the conditional distributions of the potential outcomes share a common baseline distribution (F) but exhibit distinct exponential tiltings. This shared baseline implies two key properties. (1) Both conditional distributions $F_{(0)}(y|x)$ and $F_{(1)}(y|x)$ belong to the same parametric family, differing only in their tilt parameters. (2) The baseline F establishes a direct relationship between $F_{(0)}(y|x)$ and $F_{(1)}(y|x)$, enabling information borrowing from each other and comparative inference through their tilted representations.

2.2 Estimation of model parameters

As the first step of our causal inference procedure, we estimate all the underlying parameters in model (2.1). Under the unconfoundedness assumption, $\text{pr}(Y(k) = y|X = x, D = k) = \text{pr}(Y(k) = y|X = x)$ for any $k \in \{0, 1\}$. This implies that $F_{(k)}(y|x)$ can be identified as $P(Y \leq y|x, D = k)$, denoted by $F_k(y|x)$.

Let $\beta = (\beta_0^\top, \beta_1^\top)^\top$. Based on the observations $\{(Y_{ki}, X_{ki}), i = 1, 2, \ldots n_k, k = 0, 1\}$, the likelihood function is

$$L_n(F, \beta) = \prod_{k=0}^{1} \prod_{i=1}^{n_k} \frac{\exp(Y_{ki} \cdot X_{ki}^\top \beta_k) dF(Y_{ki})}{\int \exp(t \cdot X_{ki}^\top \beta_k) dF(t)}. \tag{2.2}$$

In the principle of empirical likelihood, we model the baseline distribution F by a step function, i.e. $F(y) = \sum_{k=0}^{1} \sum_{i=1}^{n_k} p_{ki} I(Y_{ki} \le y)$, where $p_{ki} \ge 0$ and $\sum_{k=0}^{1} \sum_{i=1}^{n_k} p_{ki} = 1$. Then the log-likelihood of (β, F) becomes

$$\ell_n(F, \beta) = \sum_{k=0}^{1} \left[\sum_{i=1}^{n_k} Y_{ki} \cdot X_{ki}^\top \beta_k + \log(p_{ki}) - \log \left\{ \sum_{s=0}^{1} \sum_{r=1}^{n_s} p_{sr} \cdot \exp(Y_{sr} \cdot X_{ki}^\top \beta_k) \right\} \right] \tag{2.3}$$

To maximize the log-likelihood (2.3), Lee et al. (2025) introduced a B-splines-based approximate maximum likelihood estimation, and Lin et al. (2021) proposed a kernel-based estimation method. However, these methods both require tuning parameters, and they estimate $F_{(k)}(y|x)$ based only on the data corresponding to $D = k$, and therefore may have potential efficiency loss. To bypass these issues, we shall introduce an iterative algorithm to directly maximize the log-likelihood (2.3), and we estimate (F, β) by their maximum empirical likelihood estimator (MELE), say $(\hat{F}, \hat{\beta})$ with $\hat{F}(y) = \sum_{k=0}^{1} \sum_{i=1}^{n_k} \hat{p}_{ki} I(Y_{ki} \le y)$.

Our iterative algorithm is built on the profile log-likelihood function of β. This is equivalent to calculating the maximum point of $\ell(\beta, F)$ with respect to F given β,

say $F_\beta(y) = \sum_{u=0}^{1}\sum_{v=1}^{n_k} p_{uv}(\beta) I(Y_{ki} \le y)$. Define a Lagrange multiplier function

$$H(F, \beta, \lambda) = \ell_n(\beta, F) + \lambda \left(\sum_{k=0}^{1}\sum_{i=1}^{n_k} p_{ki} - 1 \right),$$

where $\lambda \in \mathbb{R}$ is a Lagrange multiplier. Setting the partial derivatives of $H(F, \beta, \lambda)$ with respect to $p_{uv}, u = 0, 1, v = 1, \ldots, n_u$, to zeros leads to

$$0 = \frac{1}{p_{uv}} - \sum_{k=0}^{1}\sum_{i=1}^{n_k} \frac{\exp(Y_{uv} \cdot X_{ki}^\top \beta_k)}{\sum_{s=0}^{1}\sum_{r=1}^{n_s} p_{sr} \exp(Y_{sr} \cdot X_{ki}^\top \beta_k)} + \lambda. \tag{2.4}$$

Multiplying both sides of (2.4) by p_{uv} and taking summation with respect to u and v, we have

$$0 = n - \sum_{u=0}^{1}\sum_{v=1}^{n_u} p_{uv} \sum_{k=0}^{1}\sum_{i=1}^{n_k} \frac{\exp(Y_{uv} \cdot X_{ki}^\top \beta_k)}{\sum_{s=0}^{1}\sum_{r=1}^{n_s} p_{sr} \exp(Y_{sr} \cdot X_{ki}^\top \beta_k)} + \lambda = \lambda.$$

This together with (2.4) implies that given β, the log-likelihood $\ell_n(\beta, F)$ takes its maximum when p_{uv} satisfy

$$p_{uv} = \frac{1}{\sum_{k=0}^{1}\sum_{i=1}^{n_k} \exp(Y_{uv} \cdot X_{ki}^\top \beta_k) / A(X_{ki}^\top \beta_k; F)}, \tag{2.5}$$

where $A(t; F) = \sum_{k=0}^{1}\sum_{r=1}^{n_k} p_{kr} \exp(t \cdot Y_{kr})$.

Although the solution is $p_{uv}(\beta)$, we cannot derive a closed form for $p_{uv}(\beta)$ as $A(t; F)$ also involves p_{uv}. Even so, this motivates us to construct an iterative algo-

rithm for calculating $p_{uv}(\beta)$, F_β, and hence the profile log-likelihood function of β, as did by Luo and Tsai (2012). Given β, we do the following to calculate the profile log-likelihood of β, i.e. $\ell_n(\beta, F_\beta)$.

Step 1. Set $l = 0$ and set initial values $p_{uv}^{(l)} = 1/n$ for $1 \le v \le n_u$ and $u = 0, 1$.

Step 2. Given $F^{(l)}(y) = \sum_{k=0}^{1} \sum_{i=1}^{n_k} p_{ki}^{(l)} I(Y_i \le y)$, calculate

$$p_{uv}^{(l+1)} = \frac{1}{\sum_{k=0}^{1} \sum_{i=1}^{n_k} \exp(Y_{uv} \cdot X_{ki}^\top \beta_k)/A(X_{ki}^\top \beta_k; F^{(l)})}.$$

Step 3. Repeat Step 2 until a pre-specified tolerance is met.

Step 4. Let $p_{uv}(\beta)$ denote the final $p_{uv}^{(l+1)}$ and $F_\beta(y) = \sum_{u=0}^{1} \sum_{v=1}^{n_k} p_{uv}(\beta) I(Y_{ki} \le y)$. Then the profile log-likelihood of β is $\ell_n(\beta, F_\beta)$.

Once the profile log-likelihood $\ell_n(\beta, F_\beta)$ is obtained, we immediately obtain the MELE $\hat{\beta}$ by maximizing $\ell_n(\beta, F_\beta)$ using popular optimization algorithms such as the Nelder-Mead algorithm. Accordingly, the MELE of F is $\hat{F}(y) = F_{\hat{\beta}}(y) = \sum_{u=0}^{1} \sum_{v=1}^{n_k} \hat{p}_{uv} I(Y_{ki} \le y)$ with $\hat{p}_{uv} = p_{uv}(\hat{\beta})$. Compared with the methods of Lee et al. (2025) and Lin et al. (2021), our iterative method has at least two advantages: it does not require any tuning parameter, and it makes full use of the information from all data.

In the implementation of our algorithm, an initial value for β is needed. As the SPLRM in (2.1) implies a heteroscedastic single-index model $\mathbb{E}(Y(k)|X) = m(X^\top \beta_k)$

for a function $m(\cdot)$, we may construct a consistent estimate $\tilde{\beta}_k$ for β_k by existing estimation approaches for single-index models, such as the average derivative estimator of Härdle and Stoker (1989) and the outer-product gradient estimator of Powell et al. (1989). Then we may take $\tilde{\beta} = (\tilde{\beta}_0^\top, \tilde{\beta}_1^\top)^\top$ as an initial value of β.

Assumption 2. *(1) The parameter space $\mathcal{B}$ of β is a compact set and (2) the support set $\mathcal{X}$ of X is bounded; (3) There exist $-\infty < l < u < \infty$ such that $F(l) = 0$ and $F(u) = 1$; (4) $\mathbb{C}\mathrm{ov}(X)$ is positive definite.*

The compactness of the parameter space and boundedness of covariate support in Assumption 2 are standard technical conditions that enable uniform convergence arguments. The compact support assumption on F guarantees that the response variable is properly defined over a closed interval, while the positive definiteness of $\mathbb{C}\mathrm{ov}(X)$ ensures the asymptotic variance of our proposed estimator $\hat{\beta}$ remains well-defined. These conditions in Assumption 2 are relatively mild and were similarly employed by Luo and Tsai (2012) in their Conditions 1 and 2.

Under Model (2.1), let (F^*, β^*) denote the true value of (F, β), where $\beta^* = (\beta_0^{*\top}, \beta_1^{*\top})^\top$. As a variant of Luo and Tsai (2012)'s Theorem 1, the following proposition 1 establishes the asymptotic normality of the MELE $(\hat{F}, \hat{\beta})$ and also shows that $\hat{\beta}$ is asymptotically semiparametric efficient.

Proposition 1. *Suppose that Model (2.1) and Assumption 2 are satisfied. Then $(\hat{F}, \hat{\beta})$ is consistent: $||\hat{F} - F^*||_\infty$ and $||\hat{\beta} - \beta^*||_2$ converge almost surely to 0 as $n \to \infty$. If β^* is an interior of $\mathcal{B}$, then $\sqrt{n}||\hat{F} - F^*||_\infty$ is bounded in probability and*

$\sqrt{n}(\hat{\beta} - \beta^*)$ *converges in distribution to a* $2d$*-dimensional normal distribution with mean* 0 *and efficient variance.*

Under Assumption 1, once reasonable estimators $\hat{F}$, $\hat{\beta}_0$ and $\hat{\beta}_1$ are obtained, a natural estimator of $\hat{F}_{(k)}(y|x)$ $(k = 0, 1)$ can be constructed as

$$\hat{F}_{(k)}(y|x) = \sum_{s=0}^{1} \sum_{r=1}^{n_s} \hat{p}_{sr}(\hat{\beta}) \frac{\exp(x^\top \hat{\beta}_k Y_{sr})}{\hat{A}(x^\top \hat{\beta}_k; \hat{F})} I(Y_{kr} \leq y), \tag{2.6}$$

where $\hat{A}(x^\top \hat{\beta}_k; \hat{F}) = \sum_{s=0}^{1} \sum_{r=1}^{n_s} \hat{p}_{sr}(\hat{\beta}) \exp(x^\top \hat{\beta}_k Y_{sr})$. Accordingly, we propose to estimate $F_{(k)}(y)$ by

$$\hat{F}_{(k)}(y) = \frac{1}{n} \sum_{i=1}^{n} \hat{F}_{(k)}(y|X_i). \tag{2.7}$$

From (2.6) and (2.7), it is clear that the proposed conditional distribution estimator $\hat{F}_{(k)}(y|x)$ and marginal distribution estimator $\hat{F}_{(k)}(y)$ are proper distribution estimators as they are both monotonically non-decreasing with respect to y, and are always bounded between 0 and 1.

Let $\mathbb{P}_n$ be the empirical measure induced by the data-set $\{Z_i : 1 \leq i \leq n\}$ with $Z = (Y_i, X_i, D_i)$, and denote P_0 the probability measure of $Z = (Y, X, D)$. Let $\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - P_0)$. To avoid confusion, we denote by F_X the marginal distribution of X, and write $F_{(k)}(y|x)$ as $F_{Y(k)|X}(y|x)$.

Theorem 1 (Asymptotic normalities of $\hat{F}_{(k)}(y|x)$ and $\hat{F}_{(k)}(y)$)**.** *Suppose that Assumptions 1-2 and Model* (2.1) *are satisfied, and* $n/n_k \to \rho_k \in (1, \infty)$ *for* $k = 0$

and 1 *as* $n \to \infty$. *Denote* $Z_{nk}(y,x) = \sqrt{n}(\hat{F}_{(k)}(y|x) - F_{(k)}(y|x))$ *and* $Z_{nk}(y) = \sqrt{n}(\hat{F}_{(k)}(y) - F_{(k)}(y))$. *Then as* $n \to \infty$,

(1) $Z_{nk}(y,x) \xrightarrow{d} Z_k(y,x)$, *where* $Z_k(y,x)$, *defined in equation (S2.8) in the supplementary material, is a zero-mean tight Gaussian process with a.s uniformly continuous paths.*

(2) $Z_{nk}(y) \xrightarrow{d} \int Z_k(y,x)dF_X(x) + \mathbb{G}\{F_{Y(k)|X}\} := Z_k(y)$, *a zero-mean tight Gaussian process with a.s. uniformly continuous sample path.*

Theorem 1 shows that our MELEs for both the conditional distributions $F_{(k)}(y|x)$ and the marginal distributions $F_{(k)}(y)$ are asymptotically normal. This result provides the theoretical foundation for subsequent inference on the causal estimands derived from these distributions.

Remark 1. A key component of our method is the SPLRM framework in equation (2.1), which assumes a proportional likelihood ratio structure with a common baseline distribution F. Since model misspecification can adversely affect downstream statistical inference, it is crucial to assess the adequacy of this assumption in practice. This leads to the following two simultaneous testing questions:

(i) For each group $k \in \{0,1\}$, and under the unconfoundedness assumption, does the conditional distribution satisfy $dF_{(k)}(y \mid x) = \exp\left(y\,\beta_k^\top x\right)\, dF_k(y) / \int \exp\left(t\,\beta_k^\top x\right)\, dF_k(t)$, where F_k denotes the group-specific baseline distribution?

(ii) Do the two baseline distributions F_0 and F_1 coincide?

To address these hypotheses, we construct a score test and a likelihood ratio test, respectively. These tests are then combined via a Bonferroni adjustment to control the family-wise error rate, providing an overall assessment of model validity. Further details are available in Section S7 of the supplementary material.

3. Inferences on causal estimands

Based on the proposed MELEs in the previous section, in this section, we construct estimates for various commonly used causal estimands and then study the problem of whether there exists a distributional causal effect.

3.1 Estimation of commonly-used causal estimands

The first commonly used causal effect is the conditional average treatment effect (CATE),

$$\tau(X) = \mathbb{E}\{Y(1) - Y(0)|X\}.$$

Let $\tau_k(X) = \mathbb{E}\{Y(k)|X\}$ be the conditional mean of the potential outcome $Y(k)$ for $k = 0, 1$. Under Model (2.1), a natural estimator of the conditional mean $\tau_k(X)$ is

$$\hat{\tau}_k(X) = \sum_{s=0}^{1} \sum_{r=1}^{n_s} \frac{\exp(X^\top \hat{\beta}_k Y_{sr})}{\hat{A}(X^\top \hat{\beta}_k; \hat{F})} Y_{sr} \hat{p}_{sr}. \tag{3.8}$$

Accordingly, a natural estimator of the CATE $\tau(X)$ is

$$\hat{\tau}(X) = \hat{\tau}_1(X) - \hat{\tau}_0(X).$$

The most commonly used causal effect is the average treatment effect

$$\tau_{\text{ate}} = \mathbb{E}\{Y(1) - Y(0)\} = \mathbb{E}\{\tau(X)\}.$$

Given the estimator $\hat{\tau}(X)$, we immediately have an ATE estimator

$$\hat{\tau}_{\text{ate}} = \frac{1}{n}\sum_{i=1}^{n} \hat{\tau}(X_i) = \frac{1}{n}\sum_{i=1}^{n} \{\hat{\tau}_1(X_i) - \hat{\tau}_0(X_i)\}. \tag{3.9}$$

In many areas, the average treatment effect on treated (ATT) or that on control (ATC) are of particular interest. For example, in job training evaluations, ATT reveals how the program affected actual participants' earnings, while ATC estimates how non-participants might have benefited — distinct policy questions requiring separate estimation. ATC and ATT are defined as $\tau_{\text{atc}} = \mathbb{E}\{Y(1) - Y(0)|D = 0\}$ and $\tau_{\text{att}} = \mathbb{E}\{Y(1) - Y(0)|D = 1\}$, respectively. Naturally, their estimators can be constructed as

$$\hat{\tau}_{\text{att}} = \frac{1}{n_1}\sum_{i=1}^{n} D_i \left\{\hat{\tau}_1(X_i) - \hat{\tau}_0(X_i)\right\}, \quad \hat{\tau}_{\text{atc}} = \frac{1}{n_0}\sum_{i=1}^{n} (1 - D_i) \left\{\hat{\tau}_1(X_i) - \hat{\tau}_0(X_i)\right\}, \tag{3.10}$$

respectively.

In economics, researchers are usually more interested in the quantile treatment effect (QTE). In all cases, knowledge of all QTEs is more informative than that of the ATE, as the mean can always be estimated by integrating over the quantiles. The QTE at q-level is defined as $\tau_{\text{qte}}(q) = \inf\left\{y : F_{(1)}(y) \geq q\right\} - \inf\left\{y : F_{(0)}(y) \geq q\right\}$. We estimate the QTE at level q by

$$\hat{\tau}_{\text{qte}}(q) = \inf\left\{y : \hat{F}_{(1)}(y) \geq q\right\} - \inf\left\{y : \hat{F}_{(0)}(y) \geq q\right\},$$

where $\hat{F}_{(k)}(y)$ is the estimator in (2.7).

Below, we show that under our model assumption, all these estimators are asymptotically normal. For ease of presentation, we define necessary notations. For any function g, let $A_{kg}(u; F) = \int y^k g(y) \exp(u \cdot y) dF(y)$, and we write $A_{k1}(u; F)$ when $g(y) \equiv 1$ for $k = 0$ or 1. Define

$$K_{11}(y, x) = \exp(\beta_1^{*\top} xy) \cdot \left[\frac{y}{A_{01}(\beta_1^{*\top} x; F^*)} - \frac{A_{11}(\beta_1^{*\top} x; F^*)}{A_{01}^2(\beta_1^{*\top} x; F^*)}\right]$$

is a bounded function of bounded variation,

$$K_{12}(x) = x \cdot \left[\frac{A_{21}(\beta_1^{*\top} x; F^*)}{A_{01}(\beta_1^{*\top} x; F^*)} - \frac{A_{11}^2(\beta_1^{*\top} x; F^*)}{A_{01}^2(\beta_1^{*\top} x; F^*)}\right]$$

is a d-variate bounded vector function of bounded variation. Let $K_{01}(y, x)$ and $K_{02}(x)$

be the counterparts of $K_{11}(y,x)$ and $K_{12}(x)$, respectively, obtained by replacing β_1^* with β_0^*. Denote $K^1(y,x) = K_{11}(y,x) - K_{01}(y,x)$, $K^2(x) = (K_{02}(x)^\top, K_{12}(x)^\top)^\top$ and $K^{12}(y,x) = (K^1(y,x), K^2(x)^\top)^\top$. Let σ be the information operator defined in equation (S1.6) in the supplementary material and $\tilde{\sigma}$ its inverse operator. For any h, $\sigma[h] = (\sigma_1[h], \sigma_2[h])$, where $\sigma_1[h]$ is a bounded function of bounded variation and $\sigma_2[h]$ is a $(2d)$-dimensional vector. Denote $h^* = (h_1^*, h_2^*) = \tilde{\sigma}[h] = (\tilde{\sigma}_1[h](\cdot), \tilde{\sigma}_2[h])$. Theorem 2 demonstrates that given x, the CATE estimator $\hat{\tau}(x)$ follows asymptotically a normal distribution.

Theorem 2. *Suppose that the assumptions in Theorem 1 are satisfied. As $n \to \infty$, for each fixed x, $\sqrt{n}(\hat{\tau}(x) - \tau(x)) \xrightarrow{d} N(0, \sigma_{cate}^2(x))$, where $\sigma_{cate}^2(x) = \langle K^{12}(Y,x), \tilde{\sigma}[K^{12}(Y,x)] \rangle$ and the inner product is with respect to P_0.*

With the log-likelihood (2.3), we can derive a linear approximation of $\hat{F}$ and $\hat{\beta}$, i.e.

$$\begin{aligned}\sqrt{n}\{\int h_1 d(\hat{F} - F^*) + h_2^\top(\hat{\beta} - \beta^*)\} &= \mathbb{G}_n\{\ell_{1,F^*,\beta^*}(Z)[h_1^*] + (h_2^*)^\top \ell_{2,F^*,\beta^*}(Z)\} + o_p(1) \\ &= \mathbb{G}_n\{S_{F^*,\beta^*}[\tilde{\sigma}(h)](Z)\} + o_p(1),\end{aligned}$$

where $S_{F^*,\beta^*}[\tilde{\sigma}(h)](Z) = \ell_{1,F^*,\beta^*}(Z)[h_1^*] + (h_2^*)^\top \ell_{2,F^*,\beta^*}(Z)$ is the score function along the direction h, and $\ell_{1,F^*,\beta^*}(Z)[h_1^*]$ and $(h_2^*)^\top \ell_{2,F^*,\beta^*}(Z)$ are defined in equations (S1.3) and (S1.4) in the supplementary material.

Let $\mathbb{P}_{n1}$ be the empirical measure induced by the data-set $\{X_i : D_i = 1, 1 \le i \le n\}$,

and denote P_{01} as the conditional probability measure of X given $D = 1$. Define $\mathbb{P}_{n0}$ and P_{00} similarly. Accordingly we write $\mathbb{G}_{n0} = \sqrt{n_0}(\mathbb{P}_{n0} - P_{00})$ and $\mathbb{G}_{n1} = \sqrt{n_0}(\mathbb{P}_{n1} - P_{01})$. And let $\mathbb{E}_X$, $\mathbb{E}_X^0$ and $\mathbb{E}_X^1$ denote the expectation of X, conditional expectation of X given $D = 0$ and $D = 1$ respectively. Similar to Theorem 2, we show that the proposed estimators for the average causal effects under different groups, ATE, ATT, and ATC, all have normal limiting distributions.

Corollary 1. *Suppose that the assumptions in Theorem 1 are satisfied. Let X_* follow the same distribution as X. Then as n goes to infinity, $\sqrt{n}(\hat{\tau}_{\text{ate}} - \tau_{\text{ate}}) \xrightarrow{d} N(0, \sigma^2_{\text{ate}})$, $\sqrt{n}(\hat{\tau}_{att} - \tau_{att}) \xrightarrow{d} N(0, \sigma^2_{att})$ and $\sqrt{n}(\hat{\tau}_{atc} - \tau_{atc}) \xrightarrow{d} N(0, \sigma^2_{atc})$, where*

$$
\begin{aligned}
\sigma^2_{\text{ate}} &= \mathbb{E}_{X_*}\langle K^{12}(Y, X_*), \tilde{\sigma}[K^{12}(Y, X_*)]\rangle + P_0\{\mathbb{E}_{X_*}(S_{F^*,\beta^*}[\tilde{\sigma}[K^{12}(Y, X_*)]](Z))\}^{\otimes 2}, \\
\sigma^2_{\text{att}} &= \rho_1 \mathbb{E}^1_{X_*}\langle K^{12}(Y, X_*), \tilde{\sigma}[K^{12}(Y, X_*)]\rangle + P_0\{\mathbb{E}^1_{X_*}(S_{F^*,\beta^*}[\tilde{\sigma}[K^{12}(Y, X_*)]](Z))\}^{\otimes 2}, \\
\sigma^2_{\text{atc}} &= \rho_0 \mathbb{E}^0_{X_*}\langle K^{12}(Y, X_*), \tilde{\sigma}[K^{12}(Y, X_*)]\rangle + P_0\{\mathbb{E}^0_{X_*}(S_{F^*,\beta^*}[\tilde{\sigma}[K^{12}(Y, X_*)]](Z))\}^{\otimes 2};
\end{aligned}
$$

The asymptotic normality of the proposed QTE estimator, as presented in Theorem 3, is more challenging than the other causal effect estimators.

Theorem 3. *Suppose that the assumptions in Theorem 1 are satisfied. Suppose $f_{(0)}(y)$ and $f_{(1)}(y)$ are continuous and bounded away from 0 on $\mathcal{Y}$. As n goes to infinity,*

$$
\sqrt{n}(\hat{\tau}_{qte}(q) - \tau_{qte}(q)) \xrightarrow{d} \frac{Z_0(Q_{Y(0)}(q))}{f_{Y(0)}(Q_{Y(0)}(q))} - \frac{Z_1(Q_{Y(1)}(q))}{f_{Y(1)}(Q_{Y(1)}(q))},
$$

where $Z_k(y), k = 0, 1$ *is the zero-mean tight Gaussian processes in Theorem 1 and* $Q_{Y(k)}(q)$ *is the quantile of potential outcome* $Y(k)$ *at the level* $q \in (0, 1)$, $k = 0, 1$.

3.2 Distributional causal effect test

Each of the aforementioned causal effects, CATE, ATE, ATT, ATC, and QTE, characterizes only certain aspects of the overall causal effects of the treatment. Prior to estimation, it is essential to assess whether a causal effect exists, that is, whether the effect is significantly different from zero. We address this problem by testing whether there exists a distributional causal effect. Under model (2.1), this is equivalent to testing $H_0 : \beta_0 = \beta_1$. We propose to test this hypothesis by using the likelihood ratio test statistic

$$R = 2 \left\{ \sup_{\beta, F} \ell_n(F, \beta) - \sup_{\beta_1, F} \ell_n(F, (\beta_1^\top, \beta_1^\top)). \right\}$$

Despite the existence of the infinite-dimensional nuisance parameter F, we find that the likelihood ratio test statistic follows a central chisquare limiting distribution under $H_0 : \beta_0 = \beta_1$. This is formally stated as the following theorem.

Theorem 4. *Suppose that the assumptions in Theorem 1 are satisfied and the null-hypothesis* $H_0 : \beta_0 = \beta_1$ *is true. Then* n *goes to infinity,* $R \xrightarrow{d} \chi^2_d$, *where* d *is the dimension of* X.

At the significance level α, we recommend to reject $H_0 : \beta_0 = \beta_1$ if $R > \chi^2_{d,1-\alpha}$.

Theorem 4 guarantees that this likelihood ratio test has an asymptotically correct Type I error rate. In practical applications, before estimating various causal effects, we may first apply the likelihood ratio test to examine the existence of distributional causal effects. If the null hypothesis is rejected, then we proceed to make an in-depth analysis of the causal relationship between the treatment and the outcome variable. Otherwise, when the null hypothesis is not rejected, then subsequent causal effect estimation is unnecessary.

4. Simulation Study

In this section, we manipulate simulation experiments to explore the finite-sample performance of the proposed estimation and testing methods.

4.1 Methods under comparison

For comparison, we take into consideration methods based on inverse probability (or propensity score) weighting (Rosenbaum and Rubin, 1983, IPW) and regression adjustment (Hahn et al., 2020; Caron et al., 2022). When applying IPW-based methods, we may estimate the propensity score using a parametric logistic model (GLM), power series (Donald and Hsu, 2014, POW), general random forest (Wager and Athey, 2018, GRF), and gradient boosting machine (Friedman, 2001, GBM). We also incorporate two regression-adjustment strategies inspired by the structure of the T-learner meta-algorithm commonly used in the causal inference literature (Powers

et al., 2018; Künzel et al., 2019). While most meta-algorithm frameworks rely on flexible base learners (gradient booting methods, neural networks, etc) to estimate CATEs, we adopt a simpler design: the regression model within each treatment group is assumed to be either correctly specified (T-PAR) or misspecified (F-PAR). This allows us to assess the sensitivity of regression adjustment to model misspecification. By abstracting away the variability introduced by different base learners, this setup enables a controlled comparison of regression-adjustment strategies under correct and incorrect model specifications.

In addition, we also consider a simple pooled regression approach and the widely used AIPW approach. Specifically, our simulation study evaluates the performance of the following methods:

(1) SPLRM: the proposed estimation method;

(2) T-PAR: a parametric method with $F_k(y|x)$ correctly specified;

(3) F-PAR : a parametric method with $F_k(y|x)$ incorrectly specified;

(4) GLM: the IPW method with propensity score estimated by GLM;

(5) POW: the IPW method with propensity score estimated by POW;

(6) GRF: the IPW method with propensity score estimated by GRF;

(7) GBM: the IPW method with propensity score estimated by GBM;

(8) POOL: a simple pooled outcome regression method that models Y on (X, D);

(9) AIPW-1: the AIPW method with both $F_k(y|x)$ and the propensity score correctly specified;

(10) AIPW-2: the AIPW method with $F_k(y|x)$ incorrectly specified but the propensity score correctly specified;

(11) IDEAL: the direct moment estimation method assuming that all potential outcomes were observable.

We implement the GRF and GBM methods for propensity score estimation with the `R` package `grf` and `GBM`, respectively. Under the unconfoundedness assumption, the propensity score can be consistently estimated by flexible nonparametric or machine-learning methods. In practice, however, correctly specifying the outcome distribution $F_k(y \mid x)$ is more challenging. Within the doubly robust framework, we investigate the impact of outcome model misspecification by comparing two AIPW estimators: one based on a correctly specified model for $F_k(y \mid x)$ and the other on a misspecified one. In both cases, the propensity score is estimated using GRF. The IDEAL method is not practically useful but is taken as a benchmark for comparison.

4.2 Data generation settings

We consider a d-variate covariate X, the components of which are independent and identically distributed as $N(1,1)$. Given $X = (x_1, x_2, \ldots, x_d)^\top$, we generate D from a Bernoulli distribution with success probability $e(x)$ and two choices of $e(x)$ are considered: $\text{logit}\{e(x)\} = -3x_1$ (linear case, **ps-L** for short), and $\text{logit}\{e(x)\} = -3x_1x_2^2x_3$

(nonlinear case, **ps-N**). Under these settings, the proportions of the treated units are about 20% and 33%, respectively. Given an observation $X_i = (X_{i1}, X_{i2}, \ldots, X_{id})^\top$, we generate the potential outcomes using the following two examples.

Example 1. $Y_i(0) = \beta_0^\top X_i + \varepsilon_i(0)$ and $Y_i(1) = \beta_1^\top X_i + \varepsilon_i(1)$, where $\varepsilon_i(0)$ and $\varepsilon_i(1)$ are i.i.d. from $N(0,1)$. Two choices of β_0 and β_1 are considered: $\beta_{0j} = \beta_{1j} = 0.4$ (**HOM** for short) and $\beta_{0j} = 0.4$, $\beta_{1j} = 0.5$ (**HET** for short) for $j = 1, \ldots, d$.

Example 2. $Y_i(0)|X \overset{i.i.d.}{\sim} U(0,5)$ and $Y_i(1)|X \overset{i.i.d.}{\sim} Gamma(a(X), 1)$, where $a(X_i) = b(X_{i1})b(X_{i2})$ is the shape parameter and $b(x) = 3/(1 + \exp(0.7(x - 0.5)))$.

We set d to be 5 or 10, set the sample size to $n = 500$, and conduct 1000 simulations in each simulation setting. Based on each simulated dataset, we calculate estimates of ATE, ATT, ATC, and QTE at the 10%, 50%, and 90% quantile levels.

4.3 Modelling details underlying the estimators under comparison

When data were generated from Example 1, the SPLRM model is correctly specified for the proposed method. For the parametric method (F-PAR), the working parametric model for $Y|(X = x, D = k)$ is set to $\tilde{X}^\top \eta_k + \varepsilon$, where $\varepsilon \sim N(0,1)$, $\tilde{X} = (1, X_1, X_1^2)$ is 3-dimensional vector and X_1 is the first component of X. For the parametric pooled regression (POOL) estimator, we fit a fully interacted linear regression model $Y = \alpha_1^\top X + \alpha_2 D + D\,\alpha_3^\top X + \varepsilon, \varepsilon \sim N(0, \sigma^2)$, which accommodates treatment effect heterogeneity through treatment–covariate interactions.

Example 2 was designed to explore the robustness of the proposed method as

the SPLRM is misspecified. When data were generated from Example 2, the parametric model in the F-PAR method is a Gamma distribution with scale parameter $\exp(\theta_{0k})$, and shape parameter $\exp(\tilde{X}^\top\theta_{1k})$, where $\tilde{X}^\top = (1, X)$ and $(\theta_{0k}, \theta_{1k})$ is unknown. When implementing the POOL estimator, we impose a Gamma distribution assumption on $Y|(X, D)$. After model selection, the shape and scale parameters are specified as $\exp(s_1 + s_2X_1D + s_3X_1X_2)$ and $\exp(t_1 + t_2X_1D)$, respectively, under the **ps-L** setting, and are $\exp(s_1 + s_2D + s_3X_1)$ and $\exp(t_1 + t_2D + t_3X_1 + t_4X_2)$, respectively, under the **ps-N** setting.

For the AIPW-2 estimator, the outcome distribution $F_k(y|x)$ is always modeled using the same working parametric specification as in F-PAR. Tables 1–2 report 100 times the simulated biases (Bias), variances (Variance), and mean squared errors (MSE) of the competing estimators when data were generated from Examples 1 and 2 with $d = 5$.

4.4 Simulation results

In Example 1, the outcome models for SPLRM and POOL are correctly specified, in addition to T-PAR. As shown in Table 1, these three estimators yield biases close to zero and comparable to the IDEAL benchmark. And SPLRM exhibits smaller bias than T-PAR and POOL under the **HET** setting in the estimation of τ_{ate}, τ_{att}, and τ_{atc}. For the **ps-L** setting, both AIPW estimators also produce negligible biases in both the **HOM** and **HET** settings. However, their biases are generally larger than

those of SPLRM, T-PAR, and POOL. In terms of variance and MSE, our SPLRM has the best performance among all estimators except IDEAL in the estimation of all causal estimands. This advantage reflects the benefit of jointly modeling treated and control data and borrowing information from them. Although POOL also combines information from the two groups of data, its variance and MSE remain larger than those of SPLRM. The IPW and AIPW estimators have even larger variance and MSE. For the **ps-N** setting in Example 1, the general simulation results (see Table S1 in Section S8 in the supplementary material) are similar to those (Table 1) for the **ps-L** setting. The major differences lies in the performance of AIPW-2: it has negligible bias and large variance in the **ps-L** setting but large bias and negligible variance in the **ps-N** setting. When the outcome model is misspecified, the finite-sample bias of AIPW depends critically on the propensity score weighting structure. This may explain the larger bias of AIPW-2 in the **ps-N** setting.

Example 2 is designed such that the SPLRM estimator is misspecified. As the models underlying T-PAR and AIPW-1 are correctly specified, Table 2 shows that these two estimators yield biases close to zero and comparable to that of IDEAL for all causal estimands under both **ps-L** and **ps-N**. As suffering from model misspecification, the other estimators, including SPLRM, the IPW estimators, F-PAR, POOL and AIPW-2, all exhibit non-negligible bias, although among these estimators, SPLRM and AIPW-2 generally have the smallest bias. In terms of variance and MSE, except the two ideal estimators IDEAL and T-PAR, our SPLRM is in general the

best among all the rest competitors. This can be attributed to SPLRM's ability to leverage both treated and control data.

In the meantime, although the IPW method is widely used in causal inference for causal effect estimation, its estimates are highly sensitive to small-value probabilities. When the weight assigned to a particular observation is close to zero, the corresponding IPW estimate becomes extremely unstable. As shown in Table 1 – Table 2 , the MSEs of all the IPW methods are generally no less than those of the proposed SPLRM method, regardless of the propensity score estimation method used, although their MSEs remain within an acceptable range in some cases. They have similar performance in terms of Bias. Overall, none of the IPW methods consistently outperforms the other IPW methods. In our simulations, both parametric and machine learning approaches were used for propensity score estimation. Nevertheless, the inherent instability of the IPW method persists and cannot be fully mitigated through improved propensity score estimation.

We also conduct simulations with a larger dimension $d = 10$; the results are reported in Section S8 in supplementary material. We find that the relative performance of all estimators remains largely unchanged when the covariate dimension increases from 5 to 10.

Table 3 presents simulated rejection rates of the proposed likelihood ratio test (LRT) in Section 3.2 at the 5% significance level. In the case of **HOM** in Example 1, the null hypothesis $H_0 : \beta_0 = \beta_1$ holds, or there is no distributional treatment effect.

The corresponding rejection rates in Table 3 are type I errors. Roughly speaking, the type I error of our LRT is under control, although it is slightly inflated in the case of high-dimensional covariates ($d = 10$). In all the other scenarios, the null hypothesis is violated, and the reported numbers are all powers. The LRT has powers of at least 72.9%, indicating that it can successfully detect the existence of the distributional treatment effect with a large probability, which is desirable.

Overall, our simulation results demonstrate that the proposed test yields reliable conclusions and can serve as a valuable tool in causal analysis. In practice, for a given causal problem, we recommend first applying the LRT to assess the existence of distributional causal effects. If the test indicates significant distributional effects, researchers can then proceed to estimate common causal quantities to further investigate the direction and magnitude of various causal effects.

5. Real Data Analysis

To demonstrate the practical applicability of the proposed method, we analyze data from the National Supported Work (NSW) Demonstration, a federally funded labor training program, by estimating its impact on post-intervention earnings (LaLonde, 1986). The analysis utilizes the `LLvsPSID` dataset from the `cem` package in `R`, which combines 297 treated participants from the original NSW experiment with 2,490 non-experimental controls drawn from the Panel Study of Income Dynamics (PSID). Detailed descriptions of this dataset are provided in the `cem` package documentation

Table 1: Simulated Biases, Variances, and MSEs under data generated from Example 1 with low-dimensional ($d = 5$) covariates of the **ps-L** type. All numbers reported have been multiplied by 100.

	τ_{ate}	τ_{att}	τ_{atc}	$\tau_{0.1}$	$\tau_{0.5}$	$\tau_{0.9}$	τ_{ate}	τ_{att}	τ_{atc}	$\tau_{0.1}$	$\tau_{0.5}$	$\tau_{0.9}$
	HOM:Bias						**HET:Bias**					
SPLRM	0.65	-0.86	1.02	-2.05	0.69	3.51	0.4	-0.73	0.68	-1.17	0.97	3.06
GLM	-11.54	-3.69	-12.36	0.62	-6.54	-31.75	-14.44	-3.68	-15.63	-0.06	-8.33	-38.32
POW	-16.52	-5.48	-15.63	-3.46	-12.06	-36.69	-20.36	-5.47	-19.68	-4.33	-14.65	-45.14
GRF	-20.65	-17	-19.47	-16.37	-17.68	-29.63	-24.8	-16.97	-24.43	-18.84	-22.05	-35.25
GBM	-23.26	-13.17	-25.82	-17.42	-21.72	-31.19	-28.37	-13.36	-32.02	-20.94	-25.62	-38.39
IDEAL	-0.11	-0.92	0.1	-0.04	-0.28	0.3	-0.08	-0.91	0.14	0.08	-0.18	-0.13
T-PAR	0.69	-0.78	1.04	-0.61	0.6	2.24	0.72	-0.76	1.08	-0.3	0.62	1.86
F-PAR	-1.79	0.36	-2.45	-5.77	-3.99	38.21	-2.07	0.38	-2.81	-17	-4.74	11.74
POOL	0.62	-0.93	0.99	-2.06	0.54	3.55	0.65	-0.92	1.04	-1.58	0.56	3.01
AIPW-1	2.02	-0.3	2.57	4.26	2.77	-3.11	1.54	-0.3	2.01	4.32	3.79	-2.8
AIPW-2	2.66	0.52	2.98	-4.01	1.88	4.96	1.76	0.53	1.86	-3.83	2.7	5.15
	HOM:Variance						**HET:Variance**					
SPLRM	4.77	1.59	6.88	2.12	4.96	9.58	4.67	1.63	6.72	1.9	5.23	11.2
GLM	21.79	11.42	29.93	33.37	46.82	41.28	26.74	11.57	36.56	38.62	57.46	51.25
POW	15.37	9.48	22.27	24.17	35.55	36.77	18.19	9.59	26.16	27.39	42.71	44.64
GRF	11.89	4.16	16.82	22.66	21.78	39	14.8	4.48	20.64	26.61	27.35	49.12
GBM	7.63	6.48	10.29	20.6	12.92	25.62	8.63	6.74	11.61	23.51	15.85	31.5
IDEAL	0.41	2.11	0.5	1.76	0.77	1.68	0.42	2.12	0.51	1.85	0.81	1.77
T-PAR	5.29	2.08	7.38	4.23	5.29	9.06	5.31	2.13	7.41	4.16	5.31	9.64
F-PAR	34.76	6.08	51.75	281.45	26.68	1049.58	41.77	6.42	62.25	143.24	32.02	201.57
POOL	5.28	1.65	7.4	3.73	5.28	8.84	5.3	1.7	7.43	3.68	5.3	9.45
AIPW-1	13.76	2.4	20.3	23.03	20.42	31.67	12.75	2.44	18.84	25.12	25.27	38.33
AIPW-2	46.18	6.39	69.2	52.19	44.67	89.36	48.36	6.69	72.78	61.75	57.26	111.8
	HOM:MSE						**HET:MSE**					
SPLRM	4.78	1.59	6.89	2.16	4.97	9.7	4.67	1.63	6.73	1.92	5.24	11.29
GLM	23.12	11.55	31.45	33.37	47.24	51.36	28.82	11.71	39	38.62	58.16	65.93
POW	18.1	9.78	24.71	24.29	37	50.23	22.34	9.89	30.03	27.58	44.86	65.02
GRF	16.15	7.05	20.61	25.34	24.91	47.78	20.95	7.36	26.61	30.16	32.22	61.54
GBM	13.04	8.22	16.95	23.63	17.63	35.34	16.68	8.52	21.87	27.89	22.41	46.24
IDEAL	0.41	2.12	0.5	1.76	0.77	1.68	0.42	2.13	0.51	1.85	0.81	1.77
T-PAR	5.29	2.09	7.39	4.23	5.29	9.11	5.32	2.14	7.43	4.16	5.32	9.67
F-PAR	34.79	6.08	51.81	281.78	26.84	1064.18	41.81	6.42	62.33	146.13	32.24	202.95
POOL	5.28	1.66	7.41	3.77	5.28	8.97	5.3	1.71	7.44	3.71	5.3	9.54
AIPW-1	13.81	2.4	20.37	23.21	20.5	31.77	12.77	2.45	18.88	25.31	25.41	38.41
AIPW-2	46.25	6.39	69.29	52.35	44.7	89.6	48.4	6.69	72.81	61.9	57.33	112.07

Table 2: Simulated Biases, Variances, and MSEs under data generated from Example 2 with low-dimensional ($d = 5$) covariates. All numbers reported have been multiplied by 100.

	τ_{ate}	τ_{att}	τ_{atc}	$\tau_{0.1}$	$\tau_{0.5}$	$\tau_{0.9}$	τ_{ate}	τ_{att}	τ_{atc}	$\tau_{0.1}$	$\tau_{0.5}$	$\tau_{0.9}$
	ps-L:Bias						**ps-N:Bias**					
SPLRM	4.19	-13.32	8.41	-0.57	8.89	22.68	-4.1	-10.9	-0.77	-3.75	2.5	10.16
GLM	17.1	-0.82	19.27	21.82	21.35	7.56	11.39	-0.64	16.38	3.8	13.89	15.76
POW	22.87	-0.37	23.7	23.88	26.76	18.31	9.21	-1.67	17.47	9.55	12.84	3.25
GRF	26.15	0.27	30.68	21.2	28.59	26.79	28.55	-1.2	42.6	13.81	33.42	36.86
GBM	30.02	-0.26	36.58	19.25	33.81	39.08	24.33	-1.15	36.12	12.68	29.69	28.56
IDEAL	0.03	0.81	-0.2	-0.85	-0.63	0.09	-0.41	-1.29	0.01	-0.64	-0.3	-0.96
T-PAR	0.87	0.58	0.91	0.35	1.03	1.22	0.11	-0.45	0.41	-0.5	0.32	0.56
F-PAR	24.87	7.42	29.09	1.12	64.67	-4.77	9.66	5.36	11.77	-5.99	47.84	-25.5
POOL	49.25	29.38	54.01	12.89	87.79	30.86	34.83	28.9	37.75	4.97	73.77	11.98
AIPW-1	0.45	0.29	0.46	6.17	2.15	-8.74	-0.21	-1.06	0.22	-0.28	0.97	-0.31
AIPW-2	11.3	1.06	13.78	12.99	15.73	6.65	6.13	-1.3	9.77	1.61	10.09	8.04
	ps-L:Variance						**ps-N:Variance**					
SPLRM	4.88	4.44	6.56	0.72	7.91	8.47	1.89	2.19	2.13	0.62	3.74	4.37
GLM	21.8	15.15	27.87	7.09	46.12	89	2.62	4.17	3.29	1.42	5.58	14.14
POW	17.22	14.94	23.54	6.96	38.5	75.95	17.12	17.75	17.56	6.92	39.21	58.59
GRF	11.97	6.83	16.58	5.72	22.06	67.38	2.83	3.32	3.66	1.32	5.33	15.73
GBM	9.78	9.4	12.58	4.64	16.98	65.72	3.23	4.46	4.2	1.58	6.14	19.72
IDEAL	0.94	5.86	1.03	0.54	1.91	4.12	0.89	3.13	1.28	0.51	1.85	4.01
T-PAR	1.73	3.51	1.74	0.36	1.78	7.24	1.12	1.85	1.02	0.14	0.88	5.74
F-PAR	12.41	8.95	16.14	1.99	11.41	44.31	4.79	5.79	5.2	0.81	3.8	22.77
POOL	3.84	0.75	6.78	1.8	3.54	19.07	0.95	0.53	1.39	0.42	0.89	4.85
AIPW-1	11.42	5.83	16.33	4.68	19.54	63.39	2.05	2.92	2.28	0.89	4.09	12.1
AIPW-2	14.76	7.24	21.29	5.75	20.27	72.84	2.73	3.44	3.38	1.06	4.89	13.29
	ps-L:MSE						**ps-N:MSE**					
SPLRM	5.05	6.21	7.27	0.72	8.7	13.61	2.05	3.38	2.14	0.76	3.8	5.4
GLM	24.73	15.16	31.59	11.85	50.68	89.57	3.92	4.18	5.98	1.56	7.51	16.63
POW	22.45	14.94	29.16	12.66	45.66	79.3	17.97	17.77	20.61	7.83	40.86	58.69
GRF	18.81	6.83	25.99	10.21	30.23	74.56	10.98	3.33	21.81	3.23	16.5	29.32
GBM	18.8	9.4	25.96	8.35	28.41	80.99	9.14	4.47	17.24	3.19	14.96	27.88
IDEAL	0.94	5.87	1.03	0.55	1.91	4.12	0.89	3.15	1.28	0.51	1.85	4.02
T-PAR	1.74	3.51	1.75	0.36	1.79	7.25	1.12	1.85	1.02	0.14	0.88	5.74
F-PAR	18.59	9.5	24.6	2	53.23	44.54	5.72	6.08	6.59	1.17	26.69	29.27
POOL	28.1	9.39	35.95	3.46	80.6	28.6	13.08	8.88	15.64	0.67	55.31	6.28
AIPW-1	11.42	5.83	16.33	5.06	19.58	64.15	2.05	2.93	2.28	0.89	4.1	12.1
AIPW-2	16.03	7.25	23.19	7.43	22.74	73.28	3.11	3.45	4.34	1.08	5.91	13.93

Table 3: Simulated rejection rates of the proposed likelihood ratio tests when data were generated from Examples 1 and Example 2.

	$d = 5$	$d = 10$
Example 1: **ps-L,HOM**	4.9%	7.1%
Example 1: **ps-N,HOM**	5.3%	7.1%
Example 1: **ps-L,HET**	72.9%	100%
Example 1: **ps-N,HET**	95.9%	100%
Example 2: **ps-L**	100%	99.4%
Example 2: **ps-N**	100%	100%

and the references therein.

We define the response variable as earnings in 1978 (`re78`), representing income one year after the training program concluded. The analysis focuses on individuals younger than 30 years, motivated by the following consideration. The full `LLvsPSID` dataset includes individuals aged 17 to 55, with a median age of 30 and a mean age of 33. Among individuals under 30, 249 out of 1,275 (19.5%) received treatment, whereas only 48 out of 1,512 (3.2%) individuals aged 30 or older participated in the program. This severe imbalance indicates limited overlap for older individuals, potentially violating the overlap assumption required for valid causal inference.

We first assess the presence of distributional treatment effect using the likelihood ratio test introduced in Section 3.2. The extremely small p-value 1.36×10^{-10} provides conclusive statistical evidence that the training program significantly affects earnings among participants aged below 30 years.

For individuals younger than 30, the maximum outcome in the treatment group is 6.03×10^4, whereas that in the control group reaches 11.82×10^4 (about twice of 6.03×10^4), indicating limited overlap in the support. To obtain reliable estimation

under SPLRM, we first replace the treated outcome Y_1 with $2Y_1$, and fit the proposed SPLRM model using the rescaled treated outcome. After obtaining the estimated potential outcome distributions, we transform it back to the original scale of Y_1.

We apply the same set of comparison methods as in the simulation study, except for the IDEAL method, which is infeasible in practice. For the regression-adjustment approach, we select an appropriate base learner via a model selection procedure, and refer to the resulting method as REG. REG fits separate outcome models for the treated and control groups, whereas POOL imposes a common parametric model across groups. The propensity score in AIPW is estimated using GRF, and the outcome distribution model coincides with that selected for REG.

Table 4 presents various estimates and 95% bootstrap confidence intervals for ATE, ATT, and ATC. Both the proposed SPLRM and REG produce significant negative estimates, with comparable magnitudes. Their signs are also aligned with those from AIPW and most IPW-based methods, indicating consistent evidence of negative treatment effects. Moreover, the magnitude of ATC is systematically larger (in absolute value) than that of ATT across these methods. In contrast, POOL yields a non-significant positive ATE estimate, a significant negative ATT estimate, and a significant positive ATC estimate, exhibiting a markedly different pattern. This divergence likely reflects misspecification arising from the common parametric constraint imposed by POOL, which may fail to accommodate distributional differences between the two groups.

The findings in Table 4 may be interpreted from two perspectives. First, the outcome variable measures real earnings in 1978, whereas the National Supported Work (NSW) program was implemented during 1975–1976. The outcome thus reflects labor market performance approximately two to three years after program initiation, corresponding to a short- to medium-run evaluation horizon. A negative impact over this period is plausible, as training participation may delay labor market entry and generate temporary opportunity costs. Second, Figure S1 in the supplementary material displays the age distributions of treated and control individuals younger than 30. The treated group is younger on average (median age 22) and largely consists of recent high school graduates or college interns, who may have had weaker baseline employment prospects and relatively low opportunity costs. Consequently, even if their income temporarily declines due to participating in training, the drop won't be significant. In contrast, the control group includes individuals who are more likely to have completed college or accumulated some work experience. For these individuals, training could involve larger short-term losses by interrupting ongoing employment or delaying existing earnings opportunities. This asymmetry provides a potential explanation for why the estimated ATC is more negative than the corresponding ATT.

In Table 4, except for the IPW estimators based on POW and GLM, which yield positive ATE and ATT estimates, all other IPW variants as well as AIPW produce negative estimates of the average causal effects. Nevertheless, substantial numer-

ical differences exist among these various IPW methods. For example, for ATC, the largest IPW estimate (based on GLM) is $-13,421.23$, whereas AIPW yields $-22,799.21$. These pronounced differences highlight the sensitivity of IPW estimators to propensity score estimation, particularly when estimated scores approach zero and induce highly variable inverse probability weights. As noted by Liu and Fan (2023), a concentration of near-zero estimated propensity scores for the treated group leads to substantial instability in standard IPW estimators. Similar patterns are observed in Figure S2 in the supplementary material for individuals no elder than 30, where all four propensity score estimation methods produce a non-negligible fraction of treated units with estimated propensity scores clustered near zero, with POW exhibiting the highest concentration.

Table 4: Estimates and 95% bootstrap confidence intervals ([Lower, Upper]) for ATE, ATT, and ATC on the earnings of participants aged below 30 years based on the LLvsPSID data.

		SPLRM	GLM	POW	GRF	GBM	REG	POOL	AIPW
ATE	EST	-8838.96	-9883.65	835.64	-15272.11	-11431.55	-10464.15	1395.36	-18913.61
	Lower	-10445.15	-15408.24	-22608.50	-22629.92	-15255.96	-13268.64	-900.90	-34133.78
	Upper	-7232.77	-4359.07	24279.79	-7914.30	-7607.14	-7659.66	3691.61	-3693.43
ATT	EST	-3223.36	90.79	2052.94	-3672.18	-1109.07	-4081.19	-7554.10	-2903.04
	Lower	-4881.94	-2118.95	-20550.37	-5581.97	-3359.77	-5906.38	-9307.71	-4624.48
	Upper	-1564.78	2300.52	24656.25	-1762.40	1141.64	-2256.01	-5800.48	-1181.60
ATC	EST	-10201.81	-13421.23	-15336.74	-17408.87	-14183.99	-12013.23	3567.30	-22799.21
	Lower	-12300.87	-19042.31	-21796.85	-26347.69	-19106.64	-15347.88	674.94	-41731.00
	Upper	-8102.74	-7800.15	-8876.64	-8470.05	-9261.33	-8678.59	6459.67	-3867.42

We also calculate the estimates of QTEs at quantile levels $q = 0.1, 0.25, 0.5, 0.75, 0.9$, based on the methods under comparison. The results are presented in Table 5. For

QTE10, all methods except POOL yield non-significant estimates. At higher quantiles, all methods except POW and POOL produce significantly negative effects, with GLM and REG closest to those of SPLRM. A common pattern observed in Table 5 is that, excluding POW and POOL, the magnitude of the quantile treatment effect decreases as the quantile level increases. For interval estimation, SPLRM consistently attains shorter bootstrap confidence intervals than IPW-based methods and REG. The wider intervals of IPW and AIPW are consistent with the high variability induced by near-zero propensity scores (See Section S9 of the supplementary material for more discussion), which inflate inverse probability weights and increase standard errors. In contrast, REG fits outcome models separately within each group, so parameter estimation is based on smaller subsamples, leading to larger standard errors and hence wider intervals. Overall, SPLRM yields comparatively stable quantile effect estimates across quantile levels.

6. Discussion

This paper develops a semi-parametric SPLRM model for causal inference, offering a unified framework for estimating the conditional and marginal distributions of potential outcomes. By leveraging a shared baseline distribution and an exponential link function, the proposed method fully utilizes the data from both treatment and control groups, resulting in improved efficiency and robustness compared to existing approaches.

Table 5: Estimates and 95% bootstrap confidence intervals ([Lower, Upper]) for QTEs on the earnings of participants aged below 30 years based on the LLvsPSID data.

		SPLRM	GLM	POW	GRF	GBM	REG	POOL	AIPW
QTE10	EST	0.00	0.00	-3657.37	-218.70	0.00	-687.22	-2580.50	-129.40
	Lower	-193.08	-1246.58	-22021.15	-1751.93	-1204.28	-2073.00	-4969.69	-3905.98
	Upper	193.08	1246.58	14706.41	1314.52	1204.28	698.56	-191.31	3647.19
QTE25	EST	-4231.70	-5615.36	-3657.37	-8866.36	-7388.63	-7593.18	-5537.45	-8570.87
	Lower	-5144.56	-10300.73	-22485.05	-12792.64	-10354.46	-8946.04	-8398.18	-16323.17
	Upper	-3318.85	-930.00	15170.30	-4940.08	-4422.81	-6240.32	-2676.71	-818.57
QTE50	EST	-9324.33	-11391.45	3131.09	-17732.72	-14018.82	-12089.98	-1132.93	-17730.71
	Lower	-10030.71	-20808.15	-20357.75	-27205.88	-19376.50	-13899.83	-3624.34	-28044.24
	Upper	-8617.95	-1974.76	26619.92	-8259.56	-8661.15	-10280.13	1358.49	-7417.18
QTE75	EST	-13299.54	-13464.85	5224.29	-23230.40	-15281.18	-15481.82	4928.45	-25858.45
	Lower	-14283.83	-22337.56	-28666.35	-36698.68	-19193.52	-18336.23	1834.24	-43024.86
	Upper	-12315.24	-4592.13	39114.93	-9762.13	-11368.85	-12627.40	8022.65	-8692.04
QTE90	EST	-17584.95	-19951.96	7999.13	-21592.34	-19919.27	-16610.25	10859.45	-32657.46
	Lower	-20831.67	-33525.54	-38719.09	-42164.93	-35190.16	-22912.85	5133.66	-58587.88
	Upper	-14338.22	-6378.39	54717.36	-1019.75	-4648.39	-10307.65	16585.24	-6727.04

In the supplementary material (Section S10), we provide simulation results on the estimation of CATE. We compare the proposed SPLRM method with several popular machine learning-based CATE estimators, including GRF, X-learner, and BART. Our simulation results show that compared to the existing methods, SPLRM usually has smaller MSEs and accurate coverage with shorter confidence intervals, even under data imbalance and model misspecification. These findings further demonstrate the practical advantages of our method in capturing heterogeneous treatment effects.

Acknowledgments

This work is supported in part by the National Natural Science Foundation of China (12571283) and the Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China (JYB2025XDXM904). Yukun Liu is the corresponding author.

Supplementary Material

The supplementary material contains the proofs of Theorems 1–4 and Corollary 1, additional simulation results, and supplementary results for the real data analysis.

References

Ai, C., O. Linton, and Z. Zhang (2022). Estimation and inference for the counterfactual distribution and quantile functions in continuous treatment models. *Journal of Econometrics 228*(1), 39–61.

Bang, H. and J. M. Robins (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics 61*(4), 962–973.

Caron, A., G. Baio, and I. Manolopoulou (2022). Estimating individual treatment effects using non-parametric regression models: A review. *Journal of the Royal Statistical Society Series A: Statistics in Society 185*(3), 1115–1149.

Cattaneo, M. D. (2010). Efficient semiparametric estimation of multi-valued treatment effects under ignorability. *Journal of Econometrics 155*(2), 138–154.

Chan, K. C. G., S. C. P. Yam, and Z. Zhang (2016). Globally efficient non-parametric inference of average treatment effects by empirical balancing calibration weighting. *Journal of the Royal Statistical Society Series B: Statistical Methodology 78*(3), 673–700.

Chen, H. Y. (2004). Nonparametric and semiparametric models for missing covariates in parametric regression. *Journal of the American Statistical Association 99*(468), 1176–1189.

Cheng, M., Y. Liu, and J. Qin (2025). A novel semiparametric approach to nonignorable missing data by catching covariate marginal information. *Scandinavian Journal of Statistics 52*(2), 691–709.

Chernozhukov, V., I. Fernández-Val, and B. Melly (2013). Inference on counterfactual distributions. *Econometrica 81*(6), 2205–2268.

Donald, S. G. and Y.-C. Hsu (2014). Estimation and inference for distribution functions and quantile functions in treatment effect models. *Journal of Econometrics 178*, 383–397.

Firpo, S. (2007). Efficient semiparametric estimation of quantile treatment effects. *Econometrica 75*(1), 259–276.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistic-*

s 29(5), 1189–1232.

Frölich, M. and B. Melly (2013). Unconditional quantile treatment effects under endogeneity. *Journal of Business & Economic Statistics 31*(3), 346–357.

Hahn, P. R., J. S. Murray, and C. M. Carvalho (2020). Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects (with discussion). *Bayesian Analysis 15*(3), 965–1056.

Härdle, W. and T. M. Stoker (1989). Investigating smooth multiple regression by the method of average derivatives. *Journal of the American statistical Association 84*(408), 986–995.

Hirano, K., G. W. Imbens, and G. Ridder (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica 71*(4), 1161–1189.

Imbens, G. W. and D. B. Rubin (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Künzel, S. R., J. S. Sekhon, P. J. Bickel, and B. Yu (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences of the United States of America 116*(10), 4156–4165.

LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. *The American Economic Review 76*(4), 604–620.

Lee, S.-h., Y. Ma, and E. Ronchetti (2025). Semiparametric approach to estimation of marginal mean effects and marginal quantile effects. *Journal of Econometrics 249*, 105455.

Lin, L., L. Liu, X. Cui, and K. Wang (2021). A generalized semiparametric regression and its efficient estimation. *Scandinavian Journal of Statistics 48*(1), 1–24.

Liu, Y. and Y. Fan (2023). Biased-sample empirical likelihood weighting for missing data problems: an alternative to inverse probability weighting. *Journal of the Royal Statistical Society Series B: Statistical Methodology 85*(1), 67–83.

Lunceford, J. K. and M. Davidian (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in Medicine 23*(19), 2937–2960.

Luo, X. and W. Y. Tsai (2012). A proportional likelihood ratio model. *Biometrika 99*(1), 211–222.

Maier, M. (2011). Tests for distributional treatment effects under unconfoundedness. *Economics Letters 110*(1), 49–51.

McCullagh, P. and J. Nelder (1989). *Generalized Linear Models*. Chapman and Hall.

Neyman, J. (1990). On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science 5*(4), 465–472.

Powell, J. L., J. H. Stock, and T. M. Stoker (1989). Semiparametric estimation of index coefficients. *Econometrica: Journal of the Econometric Society 57*(6), 1403–1430.

Powers, S., J. Qian, K. Jung, A. Schuler, N. H. Shah, T. Hastie, and R. Tibshirani (2018). Some methods for heterogeneous treatment effect estimation in high dimensions. *Statistics in Medicine 37*(11), 1767–1787.

Qin, J. (1998). Inferences for case-control and semiparametric two-sample density ratio models. *Biometrika 85*(3), 619–630.

Rosenbaum, P. R. and D. B. Rubin (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika 70*(1), 41–55.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies.

Journal of Educational Psychology 66(5), 688–701.

Wager, S. and S. Athey (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association 113*(523), 1228–1242.

Yu, J., H. Zhou, and J. Cai (2022). Semiparametric estimation of average treatment effect with sieve method. *arXiv:2207.03281*.

Manli Cheng

Department of Statistics and Actuarial Sciences, University of Waterloo, Waterloo, ON N2L 3G1, Canada

KLATASDS-MOE, School of Statistics and Academy of Statistics and Interdisciplinary Sciences, East China Normal University, 200062, Shanghai, China

E-mail: cmanli176@163.com

Yukun Liu

KLATASDS-MOE, School of Statistics and Academy of Statistics and Interdisciplinary Sciences, East China Normal University, 200062, Shanghai, China

E-mail: ykliu@sfs.ecnu.edu.cn (The corresponding author)