Statistica Sinica Preprint No: SS-2025-0421	
Title	CTE2: Conditional Tail Expectation Treatment Effect
Manuscript ID	SS-2025-0421
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202025.0421
Complete List of Authors	Zi Miao, Liujun Chen and Deyuan Li
Corresponding Authors	Deyuan Li
E-mails	deyuanli@fudan.edu.cn
Notice: Accepted author version.	

Statistica Sinica: Newly accepted Paper

CTE_2: Conditional tail expectation treatment effect

Zi Miao, Fudan University

Liujun Chen, University of Science and Technology of China

Deyuan Li*, Fudan University

Abstract: The subgroup average treatment effect, which evaluates effects for predefined subpopulations, is often more interpretable than the pointwise conditional average treatment effect. In this paper, we focus on the extreme subpopulations and propose the conditional tail expectation treatment effect (CTE_2). By combining inverse probability weighting technique with extreme value theory, we propose an estimator for the CTE_2 at extreme levels. We establish the asymptotic normality of the estimator through a novel multivariate causal tail empirical process framework. The finite-sample performance of the estimator is demonstrated through simulation studies. We further illustrate the utility of the estimator by applying it to estimate the CTE_2 of college education on wages.

Key words and phrases: extreme value statistics; heavy tails; tail dependence; causal inference

*Corresponding author. E-mail: deyuanli@fudan.edu.cn

Statistica Sinica: Newly accepted Paper

1. Introduction

Causal inference plays a crucial role in diverse fields such as biomedical research, policy evaluation and hydrological data analysis. In the binary treatment framework, one primary goal is to evaluate the effect of a treatment (denoted by $D \in \{0, 1\}$) on potential outcomes $Y(D)$. The observed outcome can be written as $Y = DY(1) + (1 - D)Y(0)$. One common quantity of interest is the average treatment effect, defined as $\text{ATE} = \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]$. For heterogeneous effects, conditional ATE, defined by $\text{CATE}(x) = \mathbb{E}[Y(1)|X = x] - \mathbb{E}[Y(0)|X = x]$, allows for the evaluation of treatment effects for subpopulations characterized by the covariate value x (Abrevaya et al., 2015; Jacob, 2021; Fan et al., 2022).

The practical interpretations of CATE become more challenging when dealing with continuous covariates. In applications, researchers are often more interested in treatment effects over specific ranges of a continuous covariate rather than at individual values. The subgroup average treatment effect $\mathbb{E}[Y(1)|X \in \mathcal{A}] - \mathbb{E}[Y(0)|X \in \mathcal{A}]$, where $\mathcal{A}$ represents some subset of the covariate space, is often more interpretable than the pointwise CATE (Yang et al., 2021; Li et al., 2024). It can be applied in various fields like econometrics or biostatistics, where researchers are interested in estimating treatment effects for subpopulations defined by income brackets, such as

Dong et al. (2020), or physiological measurement ranges, such as Robertson et al. (2021) and Li et al. (2024).

The subgroup average treatment effect typically concerns a fixed region $\mathcal{A}$. In this paper, we consider the special case in which the target subgroup $\mathcal{A}$ corresponds to a tail region of the covariate distribution. In this setting, the tail region $\mathcal{A}$ varies as the sample size increases. Specifically, we consider the conditional tail expectation treatment effect defined as

$$\mathrm{CTE}_2(p) = \mathbb{E}[Y(1)|X > Q_X(1-p)] - \mathbb{E}[Y(0)|X > Q_X(1-p)],$$

where p is a small probability level and Q_X is the quantile function of X. $\mathrm{CTE}_2(p)$ can capture average causal effects on sparse subpopulations characterized by the extreme values of X, enabling more targeted analysis in applications such as one treatment or regimen for patients with severe biomarker levels in clinical research, or causal effects of one policy on extremely low- or high-income populations in policy evaluation.

In applications, the probability p can be taken extremely small, indicating that the relevant subpopulations fall within the tail of the distribution. In this case, conventional nonparametric methods usually have poor performance in the estimation of $\mathrm{CTE}_2(p)$ due to the data sparsity in the tails. Extreme value theory (EVT), which focuses on modeling the tail behavior of distributions, provides a theoretical framework for analyzing such

Statistica Sinica: Newly accepted Paper

tail events. We refer readers to the monographs Beirlant et al. (2004) and de Haan and Ferreira (2006) for an introduction of EVT. EVT has proven to be valuable for studying causal tail effects. Zhang (2018) and Deuber et al. (2024) proposed estimators for extremal quantile treatment effect. Huang et al. (2022) considered the estimation of extreme average treatment effect. Wang and Miao (2024) introduced an extreme-based causal effect learning method for the estimation of causal effects.

There is a large body of literature that has analyzed the causal effects of education on wages (e.g., Heckman et al., 2018 and Deuber et al., 2024). However, relatively little attention has been paid to the analysis of average causal effects within the subpopulations defined by family incomes lying in the tails, particularly for individuals from extremely low- or high-income families. High-income family background may serve as a favorable factor for the promotion of wages, as it can provide access to broader opportunities and resources associated with platforms and higher education for these subpopulation. Witteveen and Attewell (2017) found that family background has significant effect on offspring earnings for college graduates. These considerations motivate us to consider the causal effect of college education on wages for subpopulations from high-income families.

In many empirical applications like economics and finance, the out-

comes such as wages and returns can be modeled with heavy-tailed distributions. Therefore, in this paper, we assume that the potential outcomes are heavy-tailed and apply extrapolation technique in extreme value theory to estimate the extremal level of $\mathrm{CTE}_2(p)$ through the estimators of extreme value indices of the potential outcomes and $\theta_{k/n}^{(j)}$ $(j = 0, 1)$ at the intermediate level with $k = k(n) \to \infty, k/n \to 0$ as $n \to \infty$. We propose a three-step approach for estimating $\mathrm{CTE}_2(p)$ at an extreme level p, i.e., $p = p(n) \to 0, np = O(1)$ as $n \to \infty$, based on EVT. Firstly, we construct the inverse probability weighted empirical estimators for $\theta_{k/n}^{(1)}$ and $\theta_{k/n}^{(0)}$, where $\theta_{k/n}^{(j)} = \mathbb{E}[Y(j)|X > Q_X(1 - k/n)]$ $(j = 0, 1)$, and k is a chosen intermediate sequence satisfying $k = k(n) \to \infty, k/n \to 0$. Secondly, we obtain the estimators of $\theta_p^{(1)}, \theta_p^{(0)}$ (denoted by $\widehat{\theta}_p^{(1)}$ and $\widehat{\theta}_p^{(0)}$, respectively) through extrapolation technique (Weissman, 1978) in extreme value statistics. Finally, we estimate $\mathrm{CTE}_2(p)$ by $\widehat{\theta}_p^{(1)} - \widehat{\theta}_p^{(0)}$.

To develop asymptotic theories for our proposed estimator, we introduce the multivariate causal tail empirical process (MCTEP). We establish its weak convergence to a limiting Wiener process (see Section S5 of the Supplementary Material). Since our estimator can be expressed as a functional of the MCTEP, its asymptotic properties follow directly from this weak convergence result. This theoretical framework not only provides the

foundation for our estimator, but also offers a general tool for analyzing multivariate extremes in causal inference.

The rest of the paper is organized as follows. In Section 2, we propose the estimator of $\mathrm{CTE}_2(p)$ at extreme levels and establish the asymptotic theories of the proposed estimator. In addition, we propose a variance estimator and construct the confidence interval (CI) for $\mathrm{CTE}_2(p)$. In Section 3, we carry out simulations to show the finite sample performance of our proposed estimator of $\mathrm{CTE}_2(p)$. A real data application is given in Section 4. All technique details and additional figures are gathered in the Supplementary Material.

2. Methodology

2.1 Preliminaries

In this section, our objective is to estimate $\mathrm{CTE}_2(p) = \theta_p^{(1)} - \theta_p^{(0)}$ at an extreme level p, where $p = p(n) \to 0$ and $np = O(1)$ as $n \to \infty$. The available data are $\{Y_i, D_i, \boldsymbol{Z_i}\}_{i=1}^n$, where $\boldsymbol{Z_i} = (Z_{i1}, \ldots, Z_{ir}) \in \mathbb{R}^r$ and X is one component of $\boldsymbol{Z}$.

The parameters $\theta_p^{(1)}$ and $\theta_p^{(0)}$ are generally not identifiable from observational data without making additional assumptions. In this paper, we assume that the following conditions hold.

Statistica Sinica: Newly accepted Paper

(a) $(Y(1), Y(0)) \perp\!\!\!\perp D|\boldsymbol{Z}$. There exists $c > 0$, so that $c \leq \min_{\boldsymbol{z}\in\mathcal{Z}} \Pi(\boldsymbol{z}) \leq \max_{\boldsymbol{z}\in\mathcal{Z}} \Pi(\boldsymbol{z}) \leq 1 - c$, where $\mathcal{Z}$ is the support of $\boldsymbol{Z}$ and $\Pi(\boldsymbol{z}) = \Pr\{D = 1|\boldsymbol{Z} = \boldsymbol{z}\}$.

Remark 1. *The first part of Assumption (a) is the unconfoundedness condition (e.g. Rosenbaum and Rubin, 1983), which means that there are no additional unobserved confounders affecting the treatment and potential outcomes conditional on the observed covariate $\boldsymbol{Z}$. The second part is common support condition, which is standard and commonly used in causal inference literatures, see, e.g. Firpo (2007) and Zhang (2018). It ensures that the treatment assignment probabilities are uniformly bounded away from 0 and 1 over $\mathcal{Z}$, guarantees sufficient overlap between treatment and control groups and makes our proposed estimator of $CTE_2(p)$ reliable. Assumption (a) serves as a fundamental causal assumption that facilitates the inverse probability weighting technique applied in the proposed estimation of $\theta_p^{(j)}$ feasible and simplifies the theoretical derivations.*

Lemma 1. *Suppose that Assumption (a) holds. Then,*

$$\theta_p^{(1)} = \frac{1}{p}\mathbb{E}\left[\frac{YD\mathcal{I}\{X > Q_X(1-p)\}}{\Pi(\boldsymbol{Z})}\right],$$
$$\theta_p^{(0)} = \frac{1}{p}\mathbb{E}\left[\frac{Y(1-D)\mathcal{I}\{X > Q_X(1-p)\}}{1-\Pi(\boldsymbol{Z})}\right].$$

Although Lemma 1 ensures the identifiability of $\theta_p^{(1)}$ and $\theta_p^{(0)}$, conven-

Statistica Sinica: Newly accepted Paper

tional nonparametric estimators perform poorly, or may become infeasible, when p is at an extreme level due to data sparsity in the tails. In this paper, we use the extrapolation technique (Weissman, 1978) in extreme value statistics to construct estimators for $\theta_p^{(1)}$ and $\theta_p^{(0)}$. To proceed, we assume some mild conditions on the distributions of $(X, Y(0), Y(1))$.

We assume that both $Y(0)$ and $Y(1)$ follow heavy-tailed distributions. Heavy-tailedness is a common phenomenon in fields such as finance, econometrics, and environmental science. In extreme value analysis, it is typical to assume the existence of constants $\gamma_1 > 0, \gamma_0 > 0$ such that, for any $u > 0$,

$$\lim_{t\to\infty} \frac{U_{Y(j)}(tu)}{U_{Y(j)}(t)} = u^{\gamma_j}, \quad j = 0, 1, \tag{2.1}$$

where $U_{Y(j)}(t) = F^{\leftarrow}_{Y(j)}(1 - 1/t)$ with $^{\leftarrow}$ denoting the left continuous inverse function. This condition is satisfied by many heavy-tailed distributions, including Pareto, Fréchet, and t-distributions.

Additionally, for $(X, Y(1))$ and $(X, Y(0))$, we assume that they are both tail dependent in the right tail throughout the paper. Mathematically, for all $(x, y) \in [0, \infty]^2 \backslash \{(\infty, \infty)\}$, the following limits exist and are positive:

$$R^{(j)}(x, y) = \lim_{t\to\infty} t \Pr\{1 - F_X(X) \le x/t, 1 - F_{Y(j)}(Y(j)) \le y/t\}, \quad j = 0, 1. \tag{2.2}$$

If $R^{(j)}(x, y) \equiv 0$ for all $(x, y) \in [0, \infty]^2 \backslash \{(\infty, \infty)\}$, $(X, Y(j))$ is called upper

Statistica Sinica: Newly accepted Paper

tail independent; otherwise $(X, Y(j))$ is called upper tail dependent. In particular, independence between X and $Y(j)$ implies the upper tail independence. Note that $R^{(j)}(1,1)$ is called the upper tail dependence coefficient of the random pair $(X, Y(j))$ and can be represented as

$$R^{(j)}(1,1) = \lim_{t\to\infty} \Pr\{Y(j) \geq F^{\leftarrow}_{Y(j)}(1-1/t) | X \geq F^{\leftarrow}_{X}(1-1/t)\}, \quad j = 0, 1.$$

This coefficient measures the limiting probability that $Y(j)$ exceeds a high quantile conditional on X exceeding a corresponding high quantile.

In the following, we first consider the case in which $Y(1)$ and $Y(0)$ are nonnegative in Section 2.2. Then, we extend the results to the general case where $Y(1), Y(0) \in \mathbb{R}$ in Section 2.3.

2.2 Nonnegative Y

We assume that $Y(1), Y(0) \in \mathbb{R}^+$. Recall that $k = k(n)$ is an intermediate sequence satisfying $k \to \infty, k/n \to 0$ as $n \to \infty$. We begin by the estimation of $\theta^{(j)}_{k/n}$, $j = 0, 1$. Based on Lemma 1, we propose the following inverse probability weighted estimators:

$$\widehat{\theta}^{(1)}_{k/n} = \frac{1}{k}\sum_{i=1}^{n} \frac{Y_i D_i}{\widehat{\Pi}(\boldsymbol{Z_i})} \mathcal{I}\{X_i > X_{n-k,n}\}, \quad \widehat{\theta}^{(0)}_{k/n} = \frac{1}{k}\sum_{i=1}^{n} \frac{Y_i (1-D_i)}{1-\widehat{\Pi}(\boldsymbol{Z_i})} \mathcal{I}\{X_i > X_{n-k,n}\},$$

where $\widehat{\Pi}(\boldsymbol{z})$ is a suitable non-parametric estimator of the propensity score $\Pi(\boldsymbol{z})$ and $X_{n-k,n}$ denotes the $(n-k)$-th order statistic of $\{X_i\}_{i=1}^n$ with

Statistica Sinica: Newly accepted Paper

$X_{1,n} \le X_{2,n} \le \cdots \le X_{n,n}$. In this paper, we adopt the sieve method for the estimation of $\Pi(\boldsymbol{z})$ (see, e.g. Hirano et al. (2003) and Deuber et al. (2024)). Detailed implementation of the sieve method is provided in Section S1 of the Supplementary Material.

Under conditions (2.1) (heavy-tailed assumption) and (2.2), Proposition 1 of Cai et al. (2015) implies that

$$\lim_{p\to 0} \frac{\theta_p^{(j)}}{U_{Y(j)}(1/p)} = \int_0^\infty R^{(j)}(1, y^{-1/\gamma_j})dy, \quad j = 0, 1.$$

It follows that, as $n \to \infty$,

$$\theta_p^{(j)} \sim \frac{U_{Y(j)}(1/p)}{U_{Y(j)}(n/k)} \theta_{k/n}^{(j)} \sim \left(\frac{k}{np}\right)^{\gamma_j} \theta_{k/n}^{(j)}, \quad j = 0, 1.$$

These asymptotic relationships motivate us to estimate $\mathrm{CTE}_2(p)$ by the following three steps. Firstly, we estimate $\theta_{k/n}^{(1)}$ and $\theta_{k/n}^{(0)}$ by IPW and nonparametric techniques. At this intermediate level k/n, there are still sufficiently many observations (X_i, D_i, Y_i) for $D = 1$ and $D = 0$ such that $X_i > X_{n-k,n}$. Secondly, we can apply the extrapolation technique to obtain the estimator of $\theta_p^{(j)}$ as follows:

$$\widehat{\theta}_p^{(1)} = \left(\frac{k}{np}\right)^{\widehat{\gamma}_1} \widehat{\theta}_{k/n}^{(1)} \quad \text{and} \quad \widehat{\theta}_p^{(0)} = \left(\frac{k}{np}\right)^{\widehat{\gamma}_0} \widehat{\theta}_{k/n}^{(0)},$$

where $\widehat{\gamma}_1$ and $\widehat{\gamma}_0$ are some suitable estimators of γ_1 and γ_0, respectively. Finally, we obtain the estimator of $\mathrm{CTE}_2(p)$ by

$$\widehat{\mathrm{CTE}}_2(p) := \widehat{\theta}_p^{(1)} - \widehat{\theta}_p^{(0)}.$$

Statistica Sinica: Newly accepted Paper

For the estimation of γ_1 and γ_0, we use the causal Hill estimators introduced in Deuber et al. (2024):

$$\widehat{\gamma}_1 := \frac{1}{k}\sum_{i=1}^{n}\left(\log(Y_i) - \log(\widehat{Q}_{Y(1)}(1-k/n))\right)\frac{D_i}{\widehat{\Pi}(\boldsymbol{Z_i})}\mathcal{I}\{Y_i > \widehat{Q}_{Y(1)}(1-k/n)\},$$

$$\widehat{\gamma}_0 := \frac{1}{k}\sum_{i=1}^{n}\left(\log(Y_i) - \log(\widehat{Q}_{Y(0)}(1-k/n))\right)\frac{1-D_i}{1-\widehat{\Pi}(\boldsymbol{Z_i})}\mathcal{I}\{Y_i > \widehat{Q}_{Y(0)}(1-k/n)\},$$

where

$$\widehat{Q}_{Y(1)}(1-k/n) = \arg\min_{q}\sum_{i=1}^{n}\frac{D_i}{\widehat{\Pi}(\boldsymbol{Z_i})}(Y_i - q)\left(1-k/n-\mathcal{I}\left\{Y_i \le q\right\}\right),$$

$$\widehat{Q}_{Y(0)}(1-k/n) = \arg\min_{q}\sum_{i=1}^{n}\frac{1-D_i}{1-\widehat{\Pi}(\boldsymbol{Z_i})}(Y_i - q)\left(1-k/n-\mathcal{I}\left\{Y_i \le q\right\}\right).$$

To establish the asymptotic theories of $\widehat{\theta}_p^{(1)}$ and $\widehat{\theta}_p^{(0)}$, we make the following assumptions.

(b) For $j = 0, 1$, there exist eventually positive or negative, regularly varying functions A_j with index $\rho_j < 0$ such that, for any $u > 0$,

$$\lim_{t\to\infty}\frac{\frac{U_{Y(j)}(tu)}{U_{Y(j)}(t)} - u^{\gamma_j}}{A_j(t)} = u^{\gamma_j}\frac{u^{\rho_j}-1}{\rho_j}.$$

(c) For $j = 0, 1$, there exist $\beta_j > \gamma_j$ and $\tau_j < 0$ such that, as $t \to \infty$,

$$\sup_{1/2\le x\le 2,\ 0<y<\infty}\frac{|t\Pr\{1-F_X(X) < x/t, 1-F_{Y(j)}(Y(j)) < y/t\} - R^{(j)}(x,y)|}{y^{\beta_j}\wedge 1} = O(t^{\tau_j}).$$

(d) The following limits exist for $(x, y) \in [0,\infty]^2\backslash\{(\infty,\infty)\}$:

$$\widetilde{R}^{(1)}(x,y) = \lim_{t\to\infty} t\mathbb{E}\left[\frac{1}{\Pi(\boldsymbol{Z})}\mathcal{I}\{1-F_X(X) < x/t, 1-F_{Y(1)}(Y(1)) < y/t\}\right],$$

$$\widetilde{R}^{(0)}(x,y) = \lim_{t\to\infty} t\mathbb{E}\left[\frac{1}{1-\Pi(\boldsymbol{Z})}\mathcal{I}\{1-F_X(X) < x/t, 1-F_{Y(0)}(Y(0)) < y/t\}\right].$$

Statistica Sinica: Newly accepted Paper

Moreover, the following limit exists for all $(x, y, z) \in [0, \infty]^3 \backslash \{(\infty, \infty, \infty)\}$:

$$R_{X,Y(0),Y(1)}(x, y, z)$$
$$= \lim_{t\to\infty} t \Pr\{1 - F_X(X) < x/t, 1 - F_{Y(0)}(Y(0)) < y/t, 1 - F_{Y(1)}(Y(1)) < z/t\}.$$

(e) For $j = 0, 1$, $Y(j)$ and $Y(j)|\boldsymbol{Z}$ are continuously distributed with density functions f_j and $f_{j|\boldsymbol{Z}}$, respectively. Moreover, f_j is monotonic in its right tail.

(f) As $n \to \infty$, $\sqrt{k}\{|A_1(n/k)| + |A_0(n/k)|\} \to 0$ and $k = O(n^{\frac{1}{2}\wedge\alpha_1\wedge\alpha_0})$ with $\alpha_j < \min\{-2\tau_j/(-2\tau_j + 1), 2\gamma_j\rho_j/(2\gamma_j\rho_j + \rho_j - 1)\}$.

(g) Assumptions (g1) and (g2) in the Supplementary Material are satisfied.

Assumption (b) quantifies the speed of convergence in relation (2.1). This assumption is commonly assumed in the extreme value analysis to control the biases of estimators, see, e.g. de Haan and Stadtmüller (1996), de Haan and Ferreira (2006). Assumption (c) characterizes the convergence rate in relation (2.2). A similar condition is also imposed in Cai et al. (2015). Assumption (d) provides regularity conditions that guarantee the existence of asymptotic variances in the limiting distributions of $\widehat{\theta}_p^{(1)}$ and $\widehat{\theta}_p^{(0)}$. Assumption (e) is a regularity condition on the potential outcome distributions and is commonly assumed in the literatures (Zhang, 2018;

Statistica Sinica: Newly accepted Paper

Deuber et al., 2024). Assumption (f) specifies the choice of the intermediate sequence k. For Assumption (g), (g1) is a sufficient condition to ensure that the estimation error from replacing Π by $\widehat{\Pi}$ is asymptotically negligible for the proposed $\widehat{\mathrm{CTE}}_2(p)$; (g2) ensures the existence of asymptotic variances for the causal Hill estimators $\widehat{\gamma}_0$ and $\widehat{\gamma}_1$.

We establish the asymptotic normalities of $\widehat{\theta}^{(j)}_{k/n}$ and $\widehat{\theta}^{(j)}_{p}$ in the following two theorems.

Theorem 1. *Suppose Assumptions (a)-(g) hold with* $\gamma_0, \gamma_1 \in (0, 1/2)$. *Then, as* $n \to \infty$,

$$\sqrt{k}\left(\frac{\widehat{\theta}^{(j)}_{k/n}}{\theta^{(j)}_{k/n}} - 1\right) \xrightarrow{d} \Theta_j, \quad j = 0, 1,$$

where Θ_j *is a normal random variable with* $\mathbb{E}\Theta_j = 0$ *and*

$$\mathrm{Var}(\Theta_j) = \gamma_j^2 - 1 - \left\{\int_0^\infty R^{(j)}(1, s^{-1/\gamma_j})ds\right\}^{-2} \int_0^\infty \widetilde{R}^{(j)}(1, s)ds^{-2\gamma_j}.$$

Theorem 2. *Suppose Assumptions (a)-(g) hold with* $\gamma_0, \gamma_1 \in (0, 1/2)$. *Moreover, assume that* $d_n := k/(np) \to \infty$ *and* $\log(d_n) = o(\sqrt{k})$ *as* $n \to \infty$. *Then, we have*

$$\left(\frac{\sqrt{k}}{\log(d_n)}\left(\widehat{\theta}^{(1)}_p/\theta^{(1)}_p - 1\right), \frac{\sqrt{k}}{\log(d_n)}\left(\widehat{\theta}^{(0)}_p/\theta^{(0)}_p - 1\right)\right) \xrightarrow{d} (\Gamma_1, \Gamma_0),$$

Statistica Sinica: Newly accepted Paper

where $(\Gamma_1, \Gamma_0) \sim N(0, B\widetilde{\Sigma}B^T)$, *and*

$$B := \begin{pmatrix} 1 & 0 & -\gamma_1 & 0 \\ 0 & 1 & 0 & -\gamma_0 \end{pmatrix}, \quad \widetilde{\Sigma} := \begin{pmatrix} \widetilde{G}_1 & 0 & \widetilde{J}_1 & 0 \\ 0 & \widetilde{G}_0 & 0 & \widetilde{J}_0 \\ \widetilde{J}_1 & 0 & \widetilde{H}_1 & 0 \\ 0 & \widetilde{J}_0 & 0 & \widetilde{H}_0 \end{pmatrix},$$

and G_j, J_j, H_j *are defined in Assumption (g2) in the Supplementary Material for* $j = 0, 1$.

Remark 2. *The condition that* $\log(d_n) = o(\sqrt{k})$ *(equivalently,* $\log(np) = o(\sqrt{k})$*) allows that* $np \to c \in [0, \infty)$, *and ensures that the convergence rate of* $p \to 0$ *is not too fast.*

To explore the asymptotic properties of $\widehat{\mathrm{CTE}}_2(p) = \widehat{\theta}_p^{(1)} - \widehat{\theta}_p^{(0)}$, we make a further assumption below.

(h) Ratio Assumption:

$$\lim_{p \to 0} \frac{\theta_p^{(1)}}{\theta_p^{(0)}} = \kappa \in [0, \infty].$$

The following theorem establishes the asymptotic normality of $\widehat{\mathrm{CTE}}_2(p)$.

Theorem 3. *Under the conditions of Theorem 2, along with Assumption (h), we have that as* $n \to \infty$,

$$\beta_n(\widehat{\mathrm{CTE}}_2(p) - \mathrm{CTE}_2(p)) \xrightarrow{d} \min\{1, \kappa\}\Gamma_1 - \min\{1, 1/\kappa\}\Gamma_0,$$

Statistica Sinica: Newly accepted Paper

where $\beta_n = \dfrac{\sqrt{k}}{\log(d_n)\max\{\theta_p^{(1)}, \theta_p^{(0)}\}}$. *Furthermore, if* $\kappa \neq 1$, *we have*

$$\frac{\sqrt{k}}{\log(d_n)}\left(\frac{\widehat{\mathrm{CTE}}_2(p)}{\mathrm{CTE}_2(p)} - 1\right) \xrightarrow{d} \frac{\kappa}{\kappa - 1}\Gamma_1 - \frac{1}{\kappa - 1}\Gamma_0.$$

Therefore, the asymptotic variance of $\widehat{\mathrm{CTE}}_2(p)$ *is* $\sigma^2 := v_\kappa^T B \tilde{\Sigma} B^T v_\kappa$, *where* B, $\tilde{\Sigma}$ *are the same as in Theorem 2, and*

$$v_\kappa := \begin{pmatrix} \min\{1, \kappa\} \\ -\min\{1, 1/\kappa\} \end{pmatrix}.$$

2.3 $Y(1)$ and $Y(0)$ are real

In this subsection, we extend the results to the case where $Y(1)$ and $Y(0)$ are real-valued. For $j = 0, 1$, denote $Y(j)^+ = \max(Y(j), 0)$, $Y(j)^- = \min(Y(j), 0)$, and $\theta_p^{(j+)} = \mathbb{E}\{Y(j)^+ | X > Q_X(1-p)\}$. We list two more assumptions below.

(i) $\mathbb{E}[|Y(1)^-|^{1/\gamma_1}] < \infty$, $\mathbb{E}[|Y(0)^-|^{1/\gamma_0}] < \infty$.

(j) $k = o(p^{\max\{2\tau_1(1-\gamma_1), 2\tau_0(1-\gamma_0)\}})$ as $n \to \infty$.

Assumptions (i) and (j) are two technical conditions imposed to guarantee that, as $n \to \infty$,

$$\theta_p^{(j)}/\theta_p^{(j+)} = 1 + o(1/\sqrt{k}), \quad j = 0, 1.$$

Statistica Sinica: Newly accepted Paper

Therefore, we can estimate $\theta^{(1)}_{k/n}, \theta^{(0)}_{k/n}$ through some modification of the estimators in Section 2.2 as

$$\widehat{\theta}^{(1)}_{k/n} = \frac{1}{k}\sum_{i=1}^{n} \frac{Y_i D_i}{\widehat{\Pi}(\boldsymbol{Z_i})} \mathcal{I}\{Y_i > 0, X_i > X_{n-k,n}\},$$

$$\widehat{\theta}^{(0)}_{k/n} = \frac{1}{k}\sum_{i=1}^{n} \frac{Y_i (1-D_i)}{1-\widehat{\Pi}(\boldsymbol{Z_i})} \mathcal{I}\{Y_i > 0, X_i > X_{n-k,n}\}.$$

The corresponding estimators $\widehat{\theta}^{(1)}_p$ and $\widehat{\theta}^{(0)}_p$ are obtained as

$$\widehat{\theta}^{(1)}_p = \left(\frac{k}{np}\right)^{\widehat{\gamma}_1} \widehat{\theta}^{(1)}_{k/n} \quad \text{and} \quad \widehat{\theta}^{(0)}_p = \left(\frac{k}{np}\right)^{\widehat{\gamma}_0} \widehat{\theta}^{(0)}_{k/n},$$

and the estimator for $\mathrm{CTE}_2(p)$ is given by $\widehat{\mathrm{CTE}}_2(p) = \widehat{\theta}^{(1)}_p - \widehat{\theta}^{(0)}_p$. The estimators used here are identical to in Section 2.2 when the potential outcomes are nonnegative. For simplicity, we retain the same notations and do not introduce additional symbols.

Theorem 4. *Under the conditions of Theorem 2, along with Assumptions (i) and (j), we have that, as* $n \to \infty$,

$$\left(\frac{\sqrt{k}}{\log(d_n)}\left(\widehat{\theta}^{(1)}_p/\theta^{(1)}_p - 1\right), \frac{\sqrt{k}}{\log(d_n)}\left(\widehat{\theta}^{(0)}_p/\theta^{(0)}_p - 1\right)\right) \xrightarrow{d} (\Gamma_1, \Gamma_0).$$

Furthermore, under Assumption (h) with $\kappa \neq 1$, *we have*

$$\frac{\sqrt{k}}{\log(d_n)}\left(\frac{\widehat{\mathrm{CTE}}_2(p)}{\mathrm{CTE}_2(p)} - 1\right) \xrightarrow{d} \frac{\kappa}{\kappa-1}\Gamma_1 - \frac{1}{\kappa-1}\Gamma_0.$$

Statistica Sinica: Newly accepted Paper

2.4 Construction of Confidence Interval for $\mathrm{CTE}_2(p)$

In this subsection, we construct the CI of $\mathrm{CTE}_2(p)$. Based on Theorems 3 and 4, the asymptotic variance of $\widehat{\mathrm{CTE}}_2(p)$ is $\sigma^2 = v_\kappa^T B \widetilde{\Sigma} B^T v_\kappa$. We propose the following estimator of σ^2:

$$\widehat{\sigma}^2 := \widehat{v}_\kappa^T \widehat{B} \widehat{\Sigma} \widehat{B}^T \widehat{v}_\kappa,$$

where

$$\widehat{B} := \begin{pmatrix} 1 & 0 & -\widehat{\gamma}_1 & 0 \\ 0 & 1 & 0 & -\widehat{\gamma}_0 \end{pmatrix}, \quad \widehat{v}_\kappa := \begin{pmatrix} \min\{1, \widehat{\kappa}\} \\ -\min\{1, 1/\widehat{\kappa}\} \end{pmatrix}$$

with $\widehat{\kappa} = \widehat{\theta}_p^{(1)} / \widehat{\theta}_p^{(0)}$, and $\widehat{\Sigma}$ is defined in Section S4 of the Supplementary Material. The consistency of $\widehat{\sigma}^2$ is obtained by the following Theorem 5.

Theorem 5. *Under the conditions of Theorems 3 and 4, we have that, as* $n \to \infty$,

$$\widehat{\sigma}^2 \xrightarrow{P} \sigma^2.$$

Then, we can construct the following approximated $(1-\alpha)-$CI for $\mathrm{CTE}_2(p)$:

$$\left[\widehat{\mathrm{CTE}}_2(p) \pm z_{\alpha/2} \frac{\widehat{\sigma}}{\widehat{\beta}_n} \right],$$

where $\widehat{\beta}_n = \dfrac{\sqrt{k}}{\log(d_n) \max\{\widehat{\theta}_p^{(1)}, \widehat{\theta}_p^{(0)}\}}$, and $z_{\alpha/2}$ is the $(1-\alpha/2)-$quantile of the standard normal distribution.

3. Simulations

In this section, we conduct a simulation study to evaluate the finite-sample performance of the proposed estimator. We compare it with the following estimator based on the naive inverse probability weighting technique:

$$\widehat{\mathrm{CTE}}_{2,\mathrm{naive}}(p) = \widehat{\theta}_p^{(1),\mathrm{naive}} - \widehat{\theta}_p^{(0),\mathrm{naive}},$$

where

$$\widehat{\theta}_p^{(1),\mathrm{naive}} = \frac{1}{\lfloor np \rfloor} \sum_{i=1}^{n} \frac{D_i}{\widehat{\Pi}(\boldsymbol{Z_i})} Y_i \mathcal{I}\{X_i > X_{n-\lfloor np \rfloor,n}\},$$

$$\widehat{\theta}_p^{(0),\mathrm{naive}} = \frac{1}{\lfloor np \rfloor} \sum_{i=1}^{n} \frac{1-D_i}{1-\widehat{\Pi}(\boldsymbol{Z_i})} Y_i \mathcal{I}\{X_i > X_{n-\lfloor np \rfloor,n}\},$$

and $\lfloor a \rfloor$ represents the largest integer not exceeding a.

Throughout our simulation study, for $i = 1, 2, 3, 4$, we use the notation Model (i-t) and (i-g) to denote Model (i) in which (X, W_1, W_0) generated from the t-copula and Gumbel copula, respectively.

3.1 Simulations with $\kappa \in (0, \infty)$

We first consider the setting in which $\theta_p^{(1)}$ and $\theta_p^{(0)}$ are asymptotically equivalent; that is, $\kappa \in (0, \infty)$. The data is generated as follows.

We denote t_ν as univariate t-distribution with degree of freedom ν. We begin by generating (X, W_1, W_0) from a three-dimensional t-copula with degrees of freedom $\nu = 3$ and correlation parameter $\rho = 0.7$, which is

Statistica Sinica: Newly accepted Paper

defined by

$$C_{\nu,\Sigma}(u, v, w) = T_{\nu,\Sigma}(F_{t_\nu}^{-1}(u), F_{t_\nu}^{-1}(v), F_{t_\nu}^{-1}(w)), \quad u, v, w \in [0, 1],$$

where $T_{\nu,\Sigma}$ denotes the CDF of a three-dimensional t-distribution with degrees of freedom ν and scale matrix

$$\Sigma = \begin{pmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{pmatrix},$$

and $F_{t_\nu}^{-1}$ is the quantile function of t_ν. We also generate (X, W_1, W_0) from the three-dimensional Gumbel copula, which is defined by

$$C_\theta(u, v, w) = \exp[-\{(-\log u)^\theta + (-\log v)^\theta + (-\log w)^\theta\}^{1/\theta}], \quad u, v, w \in [0, 1], \quad \theta > 1.$$

Next, let $U \sim \text{Unif}[0, 1]$ be a uniform random variable independent of (X, W_1, W_0). In this subsection, we consider univariate covariate $\boldsymbol{Z} = X$ as Zhang (2018) and Deuber et al. (2024). The extension to multivariate covariates is provided in Section S7 of the Supplementary Material.

We assign treatment according to $D = I\{U \leq \Pi(\boldsymbol{Z})\}$, where the propensity score is defined as $\Pi(\boldsymbol{z}) = \frac{e^{\boldsymbol{z}}}{1+e^{\boldsymbol{z}}} - 0.1$. We consider the following models for the potential outcomes:

(1) $Y(1) = 6F_{t_5}^{-1}((1 + W_1)/2)$, $Y(0) = 2F_{t_5}^{-1}((1 + W_0)/2)$.

Statistica Sinica: Newly accepted Paper

(2) $Y(1) = 8/(-\log(W_1))^{1/6}$, $Y(0) = 4/(-\log(W_0))^{1/6}$.

In Model (1), both $Y(1)/6$ and $Y(0)/2$ follow an absulote t_5-distribution, and $\kappa = 3$. In Model (2), both $Y(1)/8$ and $Y(0)/4$ follow a Fréchet(6) distribution, and $\kappa = 2$. Here, Fréchet(α) denotes the Fréchet distribution with shape parameter α, whose CDF is given by $F(x) = \exp(-x^{-\alpha})$.

We evaluate the performance of the proposed estimator using $M = 1000$ independent datasets with sample sizes $n \in \{1000, 2000\}$. The target parameter is $\mathrm{CTE}_2(p)$, where $p \in \{0.005, 0.001, 0.0005\}$. The true values of $\mathrm{CTE}_2(p)$ are approximated via a large-scale Monte Carlo simulation. Specifically, we generate 20 independent samples of size $N = 10^8$, and for each sample we compute the empirical approximation:

$$\frac{1}{Np}\sum_{i=1}^{N} Y_i(1)\mathcal{I}\{X_i > 1-p\} - \frac{1}{Np}\sum_{i=1}^{N} Y_i(0)\mathcal{I}\{X_i > 1-p\}.$$

We then take the median of the 20 resulting estimates as the approximated true value of $\mathrm{CTE}_2(p)$.

Figure 1 presents boxplots of the ratios between the estimated and true values for both our proposed and naive estimators for Models (1-t) and (2-t). The value of k is chosen as $k = 50$ for $n = 1000$ and $k = 100$ for $n = 2000$. The boxplots demonstrate that the proposed estimator outperforms the naive estimator for $p = 0.001$ and 0.0005 in terms of the

Statistica Sinica: Newly accepted Paper

width of boxes, primarily due to its more efficient utilization of informative tail observations. The naive estimator is not applicable for $np < 1$ and suffers from large variances when np is small. In comparison, our proposed estimator remains accurate and stable as p changes from 0.005 to 0.0005. The performance of the proposed estimator under Model (1-t) is inferior to Model (2-t), which can be attributed to the tail property of the absolute t-distribution and the t-copula. For $j = 0, 1$, the ρ_j in Model (1) is $-2/\nu_j$ with $\nu_j = 5$, whereas in Model (2), $\rho_j = -1$. It is well established that extreme value estimators tend to exhibit better finite-sample performance for smaller values of ρ_j (de Haan and Ferreira, 2006). In addition, for the t-copula, $\tau_j = -1/3$ in Assumption (c) is larger than that under the Gumbel copula ($\tau_j = -1$), implying that the convergence rate of $t \Pr\{1 - F_X(X) < x/t, 1 - F_{Y_{(j)}}(Y(j)) < y/t\}$ to $R^{(j)}(x, y)$ for t-copula is slower. These may adversely affect the performance of the proposed estimator.

To assess the sensitivity of the proposed estimator to the choice of k, we compute the relative mean square error (rMSE) as

$$\text{rMSE} = \frac{1}{M} \sum_{m=1}^{M} \left(\frac{\widehat{\text{CTE}}_{2,m}(p)}{\text{CTE}_2(p)} - 1 \right)^2,$$

where $\widehat{\text{CTE}}_{2,m}(p)$ denotes the estimate of $\text{CTE}_2(p)$ based on the m-th simulated sample. For Models (1-t) and (2-t), the rMSE curves across a range of k values are shown in Figure 2. The results indicate that the rMSE remains

Statistica Sinica: Newly accepted Paper

relatively stable over a wide range of k values, demonstrating robustness of the estimator to the choice of k. Figures S1 and S2 in the Supplementary Material show the simulation results in terms of boxplots and rMSEs under Models (1-g) and (2-g).

Figures S3 and S4 in the Supplementary Material show the empirical coverage probabilities of CIs for $\widehat{\mathrm{CTE}}_2(p)$ across different values of k for Models (1-t) and (2-t), (1-g) and (2-g). The presented empirical coverage results indicate that in Models (1-g) and (2-g), our proposed estimator performs well over a wide range of k, particularly for smaller values of p. In terms of empirical coverage probabilities of CIs, Model (2-g) outperforms Model (1-g). This difference can be attributed to the differences of second-order indices ρ_1 and ρ_0 in Models (1) and (2), as larger values of ρ_j may deteriorate the performance of the proposed estimator.

The empirical coverage performance in Models (1-t) and (2-t) is inferior to that in Models (1-g) and (2-g), with empirical coverage probabilities generally falling below the nominal 95% level. This discrepancy may be explained by the difference between the t-copula and the Gumbel copula, as the slower convergence of $t\Pr\{1 - F_X(X) \le x/t, 1 - F_{Y(j)}(Y(j)) \le y/t\}$ to $R^{(j)}(x, y)$ may adversely affect the finite-sample performance of the proposed estimator, thereby leading to the lower observed coverage

Statistica Sinica: Newly accepted Paper

probabilities of the corresponding CIs.

3.2 Simulations with $\kappa = \infty$

In this subsection, we consider the cases where $\kappa = 0$ or ∞. By symmetry, it suffices to consider the case $\kappa = \infty$ in the following simulations.

We generate (X, W_1, W_0, Z, U, D) in the same manner as in Section 3.1. For the potential outcomes, we consider the following models:

(3) $Y(1) = 5F_{t_3}^{-1}((1+W_1)/2), Y(0) = 5F_{t_4}^{-1}((1+W_0)/2)$.

(4) $Y(1) = 5/(-\log(W_1))^{1/3}, Y(0) = 5/(-\log(W_0))^{1/4}$.

In Model (3), $Y(1)/5$ follows an absolute t_3-distribution, and $Y(0)/5$ follows an absolute t_4-distribution. In Model (4), $Y(1)/5$ follows a Fréchet(3) distribution, and $Y(0)/5$ follows a Fréchet(4) distribution. The simulation results for Models (3) and (4) are shown in Figures S5-S10 in the Supplementary Material, which imply that our proposed estimator performs consistently well in this setting.

4. Applications

In this section, we study the education effect on hourly wage at age 30 for high income population. The dataset we apply comes from National

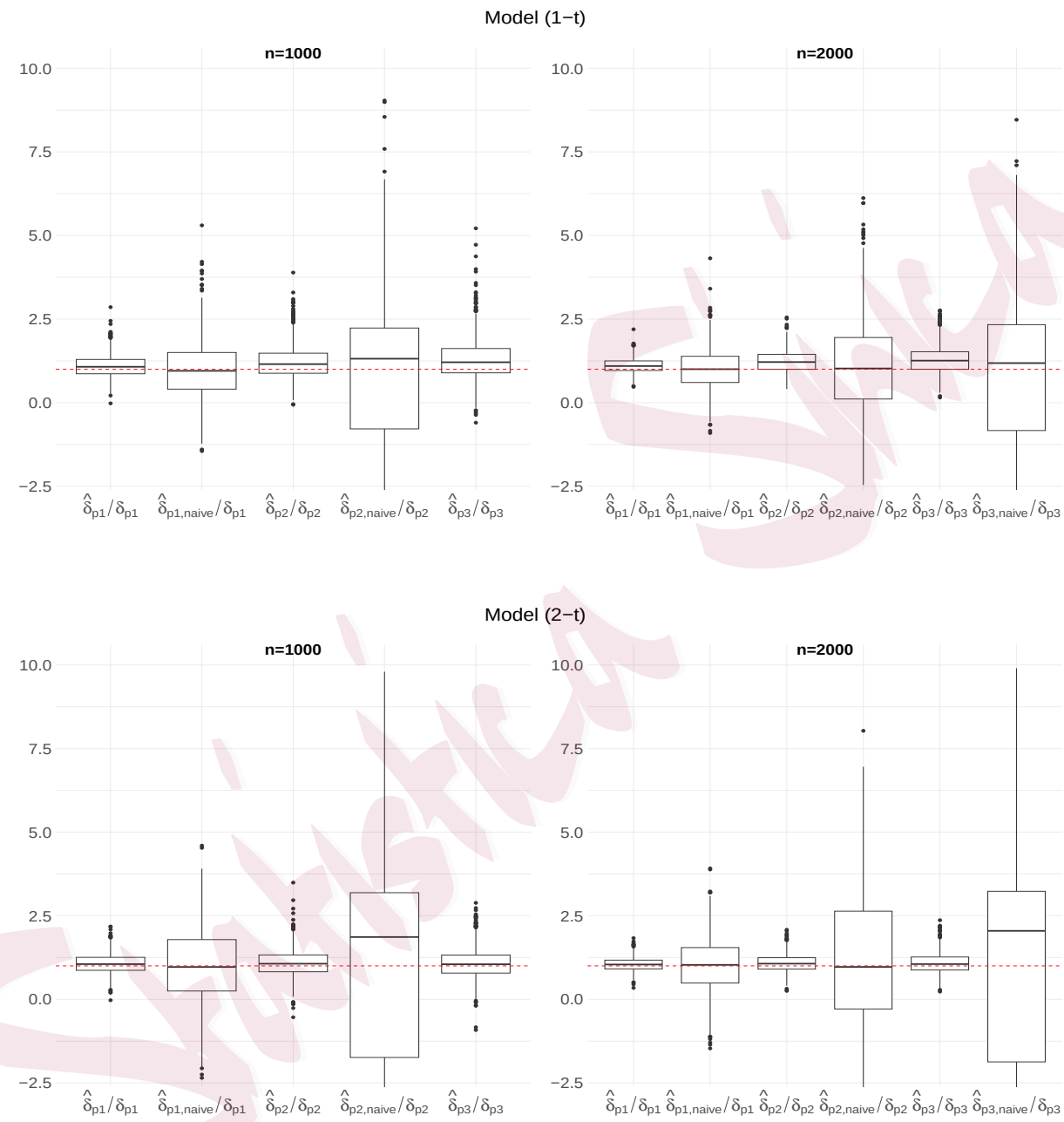

Figure 1: Boxplots for simulations under Models (1-t) and (2-t). Here, δ_p, $\widehat{\delta}_p$ and $\widehat{\delta}_{p,\text{naive}}$ denote $\text{CTE}_2(p)$, $\widehat{\text{CTE}}_2(p)$ and $\widehat{\text{CTE}}_{2,\text{naive}}(p)$, respectively. The values of p considered are $(p1, p2, p3) = (0.005, 0.001, 0.0005)$.

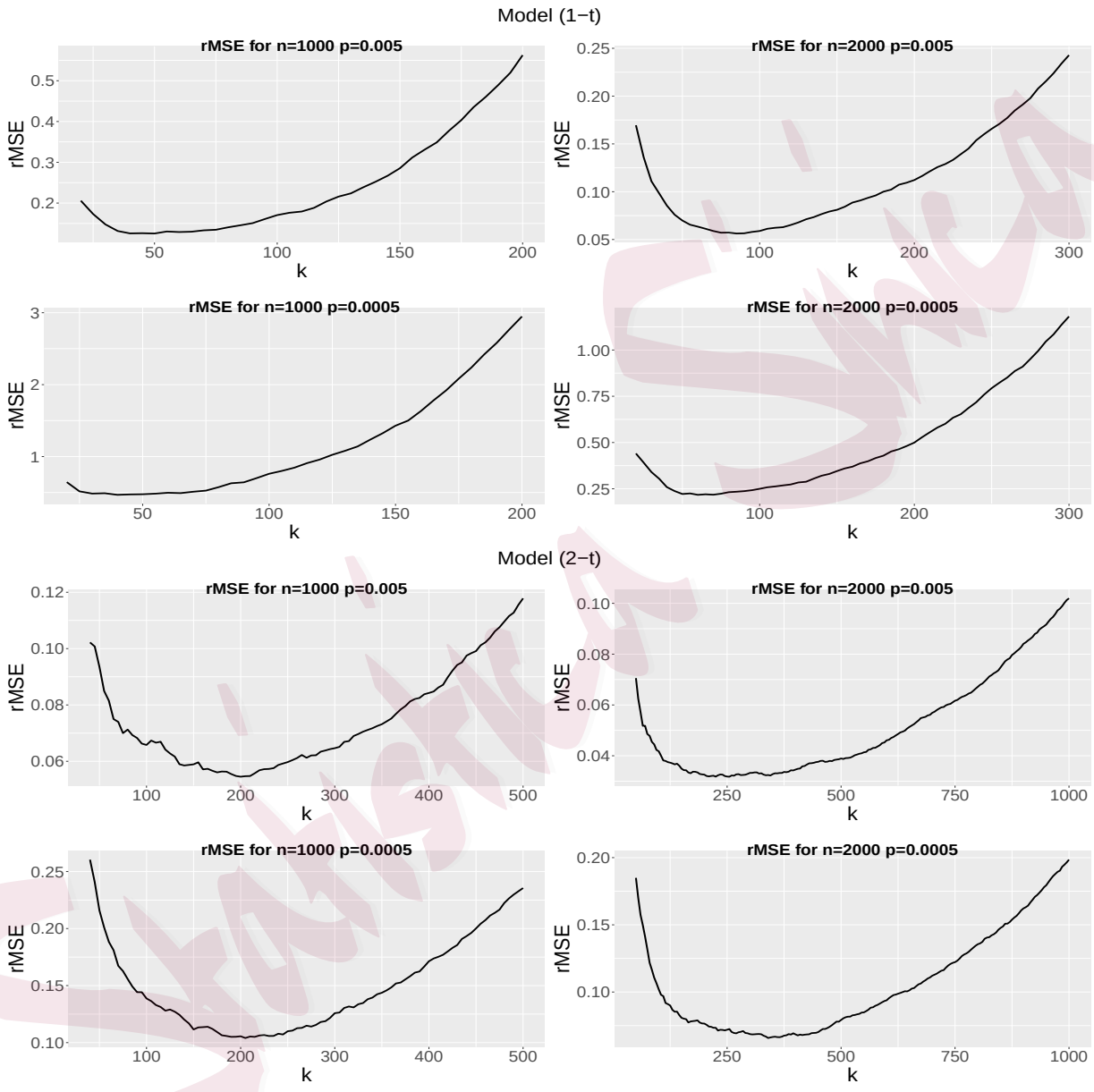

Figure 2: rMSE curves from simulation results of Models (1-t) and (2-t).

Statistica Sinica: Newly accepted Paper

Longitudinal Survey of Youth (NLSY79) and it consists of a sample of young American people who were first interviewed in 1979 at ages 14-21. This dataset has been previously analyzed by Heckman et al. (2006), Heckman et al. (2018) and Deuber et al. (2024), and it is publicly available at https://www.journals.uchicago.edu/doi/suppl/10.1086/698760.

We restrict our analysis to male respondents who completed their education before the age of 30 and did not serve in the military. Additionally, we focus on individuals who graduated from high school.

We let Y be the hourly wage (in U.S. dollar) at age 30, and X be the family income in 1979. The treatment indicator D equals 1 if an individual received college education, and 0 otherwise. The covariates $\boldsymbol{Z}$ in our analysis, following Deuber et al. (2024), include race, region of residence in 1979, urban status in 1979, age in 1979, broken home status, number of siblings, education (highest grade completed) of father and mother, scores from the Armed Services Vocational Aptitude Battery (ASVAB) test, and GPAs from 9th grade core subjects (language, math, science and social science). This results in a total of 19 covariates, making $\boldsymbol{Z}$ a 19-dimensional vector. After omitting observations with missing values, the sample size is $n = 805$ and the number of $D = 1$ is 432. The covariate X contains repeated values, resulting in ties in the data. To mitigate this issue, we add

a small perturbation $\epsilon \sim N(0, \sigma^2)$ to X, where $\sigma = 10^{-3}$.

For the estimation of the propensity score, we follow the suggestions in Deuber et al. (2024) to apply logistic regression with the linear form of all covariates. The goal is to estimate $\mathrm{CTE}_2(p)$ for $p \in \{0.005, 0.002, 0.001\}$, which belongs to the tail region but not too small.

We begin by plotting the estimated extreme value indices $\widehat{\gamma}_j$ across different values of k, in the left panel of Figure 3. Most of the estimated values fall within the interval $(0, 1/2)$, which suggests that our assumption $\gamma_0, \gamma_1 \in (0, 1/2)$ is reasonable. In the right panel of Figure 3, we plot the calculated tail dependence coefficients over different values of k, which are obtained as

$$\widehat{R}^{(1)}(1,1) = \frac{1}{k}\sum_{i=1}^{n} \frac{D_i}{\widehat{\Pi}(\boldsymbol{Z_i})} \mathcal{I}\{Y_i > \widehat{Q}_{Y(1)}(1-k/n), X_i > X_{n-k,n}\},$$

$$\widehat{R}^{(0)}(1,1) = \frac{1}{k}\sum_{i=1}^{n} \frac{1-D_i}{1-\widehat{\Pi}(\boldsymbol{Z_i})} \mathcal{I}\{Y_i > \widehat{Q}_{Y(0)}(1-k/n), X_i > X_{n-k,n}\}.$$

We observe that the estimated values $\widehat{R}^{(1)}(1,1)$ and $\widehat{R}^{(0)}(1,1)$ are larger than 0.3 and 0.15 respectively, across a large range of k, indicating notable upper tail dependence for $(X, Y(1))$ and $(X, Y(0))$.

Figure 4 shows $\widehat{\mathrm{CTE}}(p)$ with corresponding 95% CIs across different k. The results demonstrate that college education has a statistically significant beneficial effect on wages among high-income populations. Based on the plot of the estimated extreme value indices $\widehat{\gamma}_j$ for $j = 0, 1$, we select $k = 150$

for subsequent analysis. Using this choice of k, we compute the estimators of $\theta_p^{(1)}$, $\theta_p^{(0)}$, and consequently $\mathrm{CTE}_2(p)$. The results are summarized in Table 1. We observe that the estimates $\widehat{\mathrm{CTE}}_2(p)$ are positive across different values of k, indicating that for the subpopulation from high-income families, college education has substantial positive effects on wages.

Table 1: Approximated values for $\theta_p^{(1)}$, $\theta_p^{(0)}$ and $\mathrm{CTE}_2(p)$ with $p = 0.005, 0.002, 0.001$ at $k = 150$.

p	$\theta_p^{(1)}$	$\theta_p^{(0)}$	$\mathrm{CTE}_2(p)$
0.005	81.140	19.597	61.543
0.002	113.701	22.948	90.753
0.001	146.762	25.859	120.902

5. Discussion

In this paper, we introduce the conditional tail expectation treatment effect (CTE_2) to evaluate the average treatment effect on extreme subgroups. Our focus has been on the univariate case, where extreme subgroups are defined by the condition $X > Q_X(1-p)$.

This definition extends naturally to multivariate covariates $\boldsymbol{X} = (X_1, \ldots, X_d)^T$.

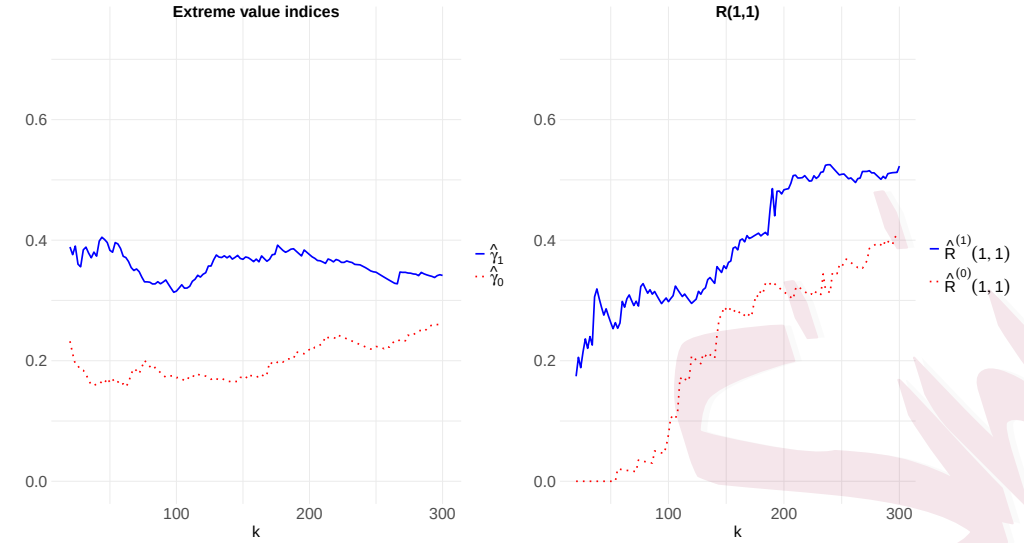

Figure 3: Estimation of γ_j and $R^{(j)}(1,1)$ $(j = 0, 1)$. The blue solid line, and the red dotted line represent the cases of $j = 1$ and $j = 0$, respectively.

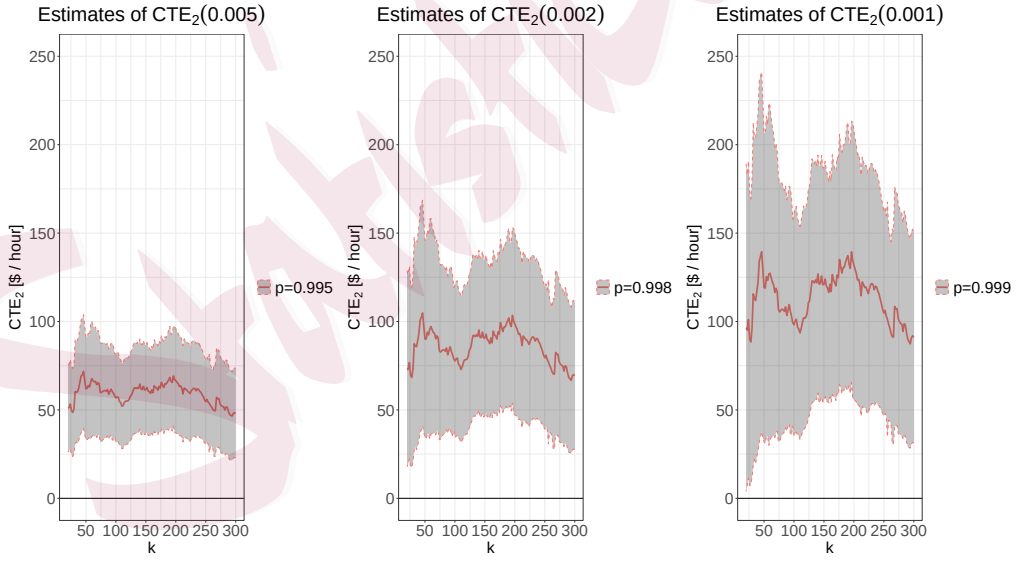

Figure 4: Estimates and shadows of 95% CIs for $\mathrm{CTE}_2(p)$ at $p = 0.005, 0.002, 0.001$.

We can define the following extended multivariate CTE_2:

$$\text{CTE}_2(\mathcal{A}_n) = \mathbb{E}[Y(1)|\boldsymbol{X} \in \mathcal{A}_n] - \mathbb{E}[Y(0)|\boldsymbol{X} \in \mathcal{A}_n],$$

where $\mathcal{A}_n \in \mathbb{R}^d$ is a tail region satisfying $\lim_{n\to\infty} P(\boldsymbol{X} \in \mathcal{A}_n) = 0$. For example, $\mathcal{A}_n$ can be chosen to be $\cap_{j=1}^d (U_{X_j}(1/p), \infty)$ or $\cup_{j=1}^d (U_{X_j}(1/p), \infty)$, where $p = p(n) \to 0$. This generalization enables $\text{CTE}_2(\mathcal{A}_n)$ to be applied across a broad range of contexts. It offers a flexible framework for examining how treatment effects concentrate in extreme subgroups, which are often of central importance in policy evaluation, medical research, and financial risk management.

Despite its conceptual appeal, estimating $\text{CTE}_2(\mathcal{A}_n)$ in the multivariate setting presents significant challenges. The primary difficulty lies in modeling the multivariate tail dependence structures (de Haan and Ferreira, 2006) of the joint distributions of $(\boldsymbol{X}, Y(1))$ and $(\boldsymbol{X}, Y(0))$. We leave this direction for future research.

Supplementary Materials

The supplementary material includes detailed proofs of Theorems 1–5 from the main paper, along with auxiliary explanations, lemmas, and propositions. Additionally, we present supplementary simulation results that complement the findings in the main paper.

Statistica Sinica: Newly accepted Paper

Acknowledgement

The authors are sincerely grateful to the Editor, the Associate Editor and two anonymous referees for their thoughtful reading, insightful comments and helpful suggestions, which have substantially improved the clarity and quality of this paper. Liujun Chen's research was partially supported by the National Natural Science Foundation of China grants 12301387 and 12471279, and the Fundamental Research Funds for the Central Universities grant WK2040250122. Deyuan Li's research was partially supported by the National Natural Science Foundation of China grant 12471279.

References

Abrevaya, J., Hsu, Y. C. and Lieli, R. P. (2015). Estimating conditional average treatment effects. *Journal of Business & Economic Statistics*, **33**(4): 485–505.

Beirlant, J., Goegebeur, Y., Segers, J. and Teugels, J. L. (2004). *Statistics of Extremes: Theory and Applications*. John Wiley & Sons.

Cai, J. J., Einmahl, J. H. J., Haan, L. and Zhou, C. (2015). Estimation of the marginal expected shortfall: the mean when a related variable is extreme. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **77**(2): 417–442.

de Haan, L. and Ferreira, A. (2006). *Extreme Value Theory: an Introduction*. Springer.

de Haan, L. and Stadtmüller, U. (1996). Generalized regular variation of second order. *Journal*

Statistica Sinica: Newly accepted Paper

of the Australian Mathematical Society, **61**(3): 381–395.

Deuber, D., Li, J., Engelke, S. and Maathuis, M. H. (2024). Estimation and inference of extremal quantile treatment effects for heavy-tailed distributions. *Journal of the American Statistical Association*, **119**(547): 2206–2216.

Dong, J., Zhang, J. L., Zeng, S. and Li, F. (2020). Subgroup balancing propensity score. *Statistical Methods in Medical Research*, **29**(3): 659–676.

Fan, Q., Hsu, Y.-C., Lieli, R. P. and Zhang, Y. (2022). Estimation of conditional average treatment effects with high-dimensional data. *Journal of Business & Economic Statistics*, **40**(1): 313–327.

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

Heckman, J. J., Humphries, J. E. and Veramendi, G. (2018). Returns to education: The causal effects of education on earnings, health, and smoking. *Journal of Political Economy*, **126**(S1): S197–S246.

Heckman, J., Stixrud, J. and Urzua, S. (2006). The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior. *Journal of Labor Economics*, **24**(3): 411–482.

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

Statistica Sinica: Newly accepted Paper

Huang, W., Li, S. and Peng, L. (2022). Extreme continuous treatment effects: measures, estimation and inference. *arXiv preprint arXiv:2209.00246*.

Jacob, D. (2021). Cate meets ml: conditional average treatment effect and machine learning. *Digital Finance*, **3**: 99–148.

Li, Y., Kuo, Y.-F. and Li, L. (2024). Propensity score analysis with guaranteed subgroup balance. *arXiv preprint arXiv:2404.11713*.

Robertson, S. E., Leith, A., Schmid, C. H. and Dahabreh, I. J. (2021). Assessing heterogeneity of treatment effects in observational studies. *American Journal of Epidemiology*, **190**(6): 1088–1100.

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

Wang, R. and Miao, W. (2024). Extreme-based causal effect learning with endogenous exposures and a light-tailed error. *arXiv preprint arXiv:2408.06211*.

Weissman, I. (1978). Estimation of parameters and large quantiles based on the k largest observations. *Journal of the American Statistical Association*, **73**(364): 812–815.

Witteveen, D. and Attewell, P. (2017). Family background and earnings inequality among college graduates. *Social Forces*, **95**(4): 1539–1576.

Yang, S., Lorenzi, E., Papadogeorgou, G., Wojdyla, D. M., Li, F. and Thomas, L. E. (2021). Propensity score weighting for causal subgroup analysis. *Statistics in Medicine*, **40**(19):

Statistica Sinica: Newly accepted Paper

4294–4309.

Zhang, Y. (2018). Extremal quantile treatment effect. *The Annals of Statistics*, **46**(6B): 3707–3740.

School of Management, Fudan University

E-mail: 24110690006@m.fudan.edu.cn

Faculty of Business for Science and Technology, School of Management, University of Science and Technology of China

E-mail: ljchen22@ustc.edu.cn

ORCID: https://orcid.org/0009-0001-5302-3527

School of Management, Fudan University

E-mail: deyuanli@fudan.edu.cn

ORCID: https://orcid.org/0000-0003-2155-2552