Statistica Sinica Preprint No: SS-2021-0254	
Title	Outlier-Resistant Estimators for Average Treatment Effect in Causal Inference
Manuscript ID	SS-2021-0254
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202021.0254
Complete List of Authors	Kazuharu Harada and Hironori Fujisawa
Corresponding Authors	Kazuharu Harada
E-mails	kharada201612@gmail.com
Notice: Accepted version subject to English editing.	

Statistica Sinica

OUTLIER-RESISTANT ESTIMATORS FOR AVERAGE TREATMENT EFFECT IN CAUSAL INFERENCE

Kazuharu Harada[1] and Hironori Fujisawa[2,1,3]

The Graduate University for Advanced Studies (SOKENDAI), Japan[1]

The Institute of Statistical Mathematics, Japan[2]

Center for Advanced Intelligence Project, RIKEN, Japan[3]

Abstract: The inverse probability (IPW) and doubly robust (DR) estimators are often used to estimate the average causal effect (ATE), but are vulnerable to outliers. The IPW/DR median can be used for outlier-resistant estimation of the ATE, but the outlier resistance of the median is limited and it is not resistant enough for heavy contamination. We propose extensions of the IPW/DR estimators with density power weighting, which can eliminate the influence of outliers almost completely. The outlier resistance of the proposed estimators is evaluated through the unbiasedness of the estimating equations. Unlike the median-based methods, our estimators are resistant to outliers even under heavy contamination. Interestingly, the naive extension of the DR estimator requires bias correction to keep the double robustness even under the most tractable form of contamination. In addition, the proposed estimators are found to be highly resistant to

outliers in more difficult settings where the contamination ratio depends on the covariates. The outlier resistance of our estimators from the viewpoint of the influence function is also favorable. Our theoretical results are verified via Monte Carlo simulations and real data analysis. The proposed methods were found to have more outlier resistance than the median-based methods and estimated the potential mean with a smaller error than the median-based methods.

Key words and phrases: Causal inference, Doubly robust, Missing data, Propensity score, Robust statistics

1. Introduction

Statistical causal inference provides various estimators for causal quantities like the average treatment effect (ATE). To estimate such quantities, the propensity score is widely applied in various ways, such as stratification, matching, inverse probability weighting (IPW), and the doubly robust (DR) estimator (Robins, Rotnitzky, and Zhao, 1994; Rosenbaum and Rubin, 1983; Bang and Robins, 2005). These estimators are designed to control confounding, and they are consistent with the target quantity under some assumptions.

As discussed in the later section, the IPW and DR estimators are vulnerable to outliers since they partially use the sample mean. An outlier in a multivariate setting is classified into three types: a vertical outlier, a good

leverage point, and a bad leverage point (Rousseeuw and van Zomeren, 1990). Figure 1 illustrates the three types of outliers. Canavire-Bacarreza et al. (2021) has investigated the influence of these types of outliers on the estimators of the ATE including IPW through exhaustive Monte-Carlo simulations; they have pointed out that the vertical outliers in the outcome variable lead to a serious bias of the ATE estimation. In this paper, we are interested in reducing the bias caused by the vertical outliers.

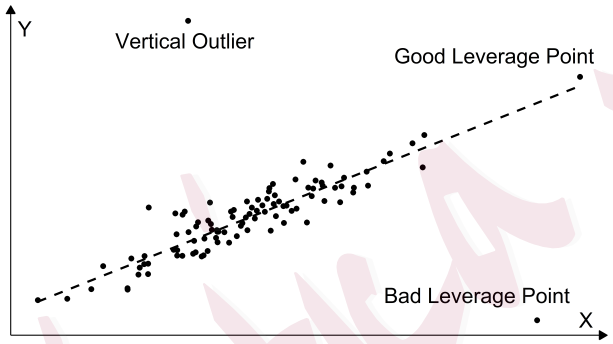

Figure 1: Three types of outliers.

Outlier-resistant statistics have been studied for long; however, most literature does not consider a causal setting (Huber, 2004; Hampel et al., 2011; Maronna et al., 2019). The established methods of outlier-resistant statistics are not directly applicable to causal settings. The median-based estimators are the only examples which are applicable to the estimation of the ATE under outlier contamination (Firpo, 2007; Zhang et al., 2012; Díaz, 2017; Sued, Valdora, and Yohai, 2020). It is well known that the

sample median is more resistant to outliers than the sample mean, but it is still affected; in particular, when the contamination ratio is not small and the outliers lie on one side of the data-generating density, the influence becomes so large as it cannot be ignored (Fujisawa and Eguchi, 2008).

In this paper, we propose extensions of the IPW and DR estimators for the mean of the potential outcome with more outlier resistance than the median-based methods. We discuss the outlier resistance of these estimators from the viewpoint of the unbiasedness of the estimating equation and influence function (IF). In most literature on outlier-resistant statistics, the contamination ratio is assumed to be small and be independent of covariates; however, we discuss the outlier resistance of the proposed estimators under more general assumptions, including the case where the contamination ratio is not small and related to covariates. Interestingly, a straight extension of the DR estimator loses the robustness to model misspecification under contamination. We also propose a bias-corrected version of the extended DR estimator, which holds the double robustness under contamination. Furthermore, the theoretical advantages of our estimators are verified through Monte-Carlo simulations and real data analysis.

The remainder of this paper is organized as follows. In Section 2, we introduce the potential outcome framework for causal inference and the

basic concept of outliers. In Section 3, we propose novel estimators and discuss the outlier resistance from the viewpoint of the unbiasedness of the estimating equations. In Section 4, we evaluate the outlier resistance in terms of the IF. In Section 5, we discuss asymptotic properties. Finally, in Sections 6 and 7, we present the experimental results.

2. Preliminaries

2.1 Potential Outcome and Treatment Effect

Let (Y, T, X) be the observable random variables; X is the outcome, T is the treatment, and X is the confounder. We assume that Y is continuous and T is binary; it is easy to extend T to multiple discrete treatments. We have the observations $(Y_i, T_i, X_i)_{i=1}^n$ drawn from the distribution of (Y, T, X) in an i.i.d. manner. Denote the potential outcome under $T = t$ by $Y^{(t)}$ and let $\mu^{(t)} = \mathbb{E}[Y^{(t)}]$. $Y^{(t)}$ is uniquely defined for every treatment as a random variable, namely, well-defined. Note that i.i.d. sampling and well-definedness of the potential outcome are collectively called the stable unit treatment value assumption (SUTVA; Imbens and Rubin, 2015). The ATE is defined as $\mu^{(1)} - \mu^{(0)}$. The ATE cannot be estimated directly since we cannot observe $Y^{(1)}$ and $Y^{(0)}$ simultaneously; instead, we use the observed variables under the common assumptions (e.g. Imbens and Rubin, 2015):

1. *Conditional Unconfoundedness*: $Y^{(t)} \perp\!\!\!\perp T$ for all $t \in \{0, 1\}$,

2. *Consistency*: $Y = Y^{(t)}$ if $T = t$,

3. *Positivity*: $P(T = 1|X) > c$ for some constant $c > 0$.

The ATE can be estimated from the observed variables under these assumptions, namely, identifiable. Hereafter, we assume the triple assumption holds and focus on the estimation of $\mu^{(1)}$ for simplicity. $\mu^{(0)}$ is estimated in a similar way, and the ATE is estimated by the difference between the estimates of $\mu^{(1)}$ and $\mu^{(0)}$.

We introduce three consistent estimators of the potential mean. The IPW estimator (Rosenbaum and Rubin, 1983) is based on the propensity score (PS). Let $\pi(x; \alpha) \in (0, 1)$ be the PS, which models $P(T = 1|x)$. We assume the PS is correctly specified, in other words, there exists α^* such that $\pi(x; \alpha^*) = P(T = 1|x)$ for every x. The IPW estimator has several forms (Lunceford and Davidian, 2004), but we use the weighted average form: $\hat{\mu}^{(1)}_{IPW} = \left(\sum_{i=1}^n T_i Y_i/\pi(X_i; \hat{\alpha})\right)/\left(\sum_{i=1}^n T_i/\pi(X_i; \hat{\alpha})\right)$, where $\hat{\alpha}$ is an estimate of α obtained in a consistent manner such as the maximum likelihood estimation (MLE). The IPW estimator is consistent with $\mu^{(1)}$ if the PS model is correctly specified. The IPW estimator can be seen as the

root of the following estimating equation:

$$\sum_{i=1}^{n} \frac{T_i}{\pi(X_i;\hat{\alpha})}(Y_i - \mu) = 0. \tag{2.1}$$

Outcome regression (OR) is also popular. To construct the OR estimator, we model $\mathbb{E}[Y|T=1,X]$ by some function $m_1(X;\beta)$. Then, the OR estimator is obtained as $n^{-1}\sum_{i=1}^{n} m_1(X_i;\hat{\beta})$, where $\hat{\beta}$ is a consistent estimate of β. The IPW and OR estimators are asymptotically consistent with $\mu^{(1)}$ when the model used in each estimator is correctly specified; the consistency is not assured if the model is misspecified. The DR estimator (Scharfstein, Rotnitzky, and Robins, 1999; Bang and Robins, 2005) combines the IPW and OR estimators. Since the DR estimator is consistent with $\mu^{(1)}$ if either the PS or OR model is correctly specified, it is said to be "doubly robust." Besides, if both models are correctly specified, the DR estimator is semiparametrically efficient (Robins and Rotnitzky, 1995; Tsiatis, 2006). Although there are many estimators equipped with double robustness, we refer the root of the following estimating equation as the DR estimator $\hat{\mu}^{(1)}_{DR}$, which is a special case of the augmented IPW estimator:

$$\sum_{i=1}^{n}\left[\frac{T_i}{\pi(X_i;\hat{\alpha})}(Y_i-\mu) - \frac{T_i-\pi(X_i;\hat{\alpha})}{\pi(X_i;\hat{\alpha})}\{m_1(X_i;\hat{\beta})-\mu\}\right] = 0. \tag{2.2}$$

2.2 IPW/DR M-estimators

Let $\sum_{i=1}^{n} \psi(Y_i, \theta) = 0$ be an estimating equation, where ψ is a known vector-valued map, and θ is the parameter of interest. An estimator $\hat{\theta}$ solving the estimating equation is called an M-estimator. M-estimator is a large class of estimators involving MLE, IPW, OR, and DR. If the estimating equation is unbiased, say $\mathbb{E}_\theta[\psi(Y, \theta)] = 0$, the M-estimator is consistent with the truth under some regularity conditions (e.g. Chap. 5 of Van der Vaart, 2000).

By replacing $Y_i - \mu$ in (2.2) with an estimating function $\psi(Y_i; \theta)$, the IPW and DR estimators can be expanded to a general M-estimator. If we are interested in the same parameter θ with respect to $Y^{(1)}$, the IPW and DR M-estimators (Tsiatis, 2006) are available:

$$\sum_{i=1}^{n} \frac{T_i}{\pi(X_i; \hat{\alpha})} \psi(Y_i; \theta) = 0, \tag{2.3}$$

$$\sum_{i=1}^{n} \left[\frac{T_i}{\pi(X_i; \hat{\alpha})} \psi(Y_i; \theta) - \frac{T_i - \pi(X_i; \hat{\alpha})}{\pi(X_i; \hat{\alpha})} \mathbb{E}_{\hat{q}}[\psi(Y_i; \theta) | T = 1, X_i] \right] = 0. \tag{2.4}$$

The conditional expectation $\mathbb{E}_{\hat{q}}[\psi(Y_i; \theta) | T = 1, X_i]$ is calculated using the parametric OR model $q(y | T = 1, x; \hat{\beta})$ via direct calculation or Monte-Carlo approximation (Hoshino, 2007). When the original M-estimating equation is unbiased, the IPW/DR estimating equations are unbiased under the proper model specification. The asymptotic properties of the IPW and

DR M-estimators follow from the standard theory of M-estimators.

2.3 Outlier-resistant Estimation

This section provides a brief review of the outlier-resistant estimation of the mean in a one-variable and non-causal setting. Let $\tilde{g}$ be the density function of a random variable $Z \in \mathbb{R}$. Assume that the density is contaminated as $\tilde{g}(z) = (1-\varepsilon)f_{\mu^*}(z) + \varepsilon\delta(z)$, where f_{μ^*} is the density of Z without contamination equipped with the mean μ^*, ε is the contamination ratio, and δ is the density of outliers. Our goal is to estimate μ^* from i.i.d. observations $\{Z_1, ..., Z_n\}$ drawn from $\tilde{g}$. If we model the contamination in this way, the sample mean converges to $(1-\varepsilon)\mu^* + \varepsilon\mathbb{E}_\delta[Z]$; if the mean of outliers is far from μ^*, the sample mean is asymptotically biased. To deal with contamination, many types of M-estimators are applied. The unbiasedness of the estimating equation does not usually hold under contamination because

$$\mathbb{E}_{\tilde{g}}[\psi(Z,\mu^*)] = (1-\varepsilon)\underbrace{\mathbb{E}_{f_{\mu^*}}[\psi(Z,\mu^*)]}_{=0} + \varepsilon\mathbb{E}_\delta[\psi(Z,\mu^*)] \neq 0. \tag{2.5}$$

By designing ψ to eliminate or bound $\mathbb{E}_\delta[\psi(Z,\mu^*)]$, the influence of outliers can be reduced. Let θ^*_ψ denote a root of $\mathbb{E}_{\tilde{g}}[\psi(Z,\theta)] = 0$; then, the latent bias is defined as $\theta^*_\psi - \theta^*$. If δ is Dirac's delta and ε is sufficiently small, the

latent bias is approximated by the IF. The IF-based discussion in Section 4 provides some insights into the outlier resistance of the estimators when the contamination ratio is small. The latent bias and M-estimators are discussed in detail elsewhere (e.g. Huber, 2004; Fujisawa, 2013; Fujisawa and Eguchi, 2008).

2.4 IPW and DR Under Contamination

Next, we move to a causal setting. In this paper, we consider the vertical outliers. In other words, we assume that only the outcome Y may be contaminated, and that the contamination does not affect the causal mechanism among (Y, T, X). A typical example is contamination of laboratory values in medical research with foreign substances. Let $\delta_{Y|TX}$ be the conditional density of outliers given (T, X), and let $\varepsilon_t(x)$ be the contamination ratio. Then, the contaminated conditional density given (T, X) is defined as

$$\tilde{g}_{Y|TX}(y|t,x) = (1-\varepsilon_t(x))g_{Y|TX}(y|t,x) + \varepsilon_t(x)\delta_{Y|TX}(y|t,x), \tag{2.6}$$

where g without tilde denotes the density without contamination; the tilde indicates that the distribution is contaminated. For simplifying the nota-

tions, we often drop the subscripts of density functions as long as there would be no confusion and write $\delta_t(y|x) = \delta_{Y|TX}(y|t,x)$ below. The contamination ratio and their density depend on the treatment T and the confounder X. Since we estimate $\mu^{(t)}$ for each treatment separately, the dependence on T is tractable. In contrast, the dependence on X is critical in our analysis. The X-dependent contamination is referred to as heterogeneous contamination. We also discuss the special case in which ε and δ are not dependent on X, called homogeneous contamination. Note that we do not assume $\varepsilon_t(x)$ to be small enough to be negligible, except in Section 4.

We are interested in the marginal mean of $Y^{(1)}$. Let $f_{Y^{(1)}}(y;\mu^{(1)})$ be the true marginal density of $Y^{(1)}$. It is obtained by integrating X out from $g_{Y|TX}(y|T,X)$ under $T=1$:

$$f_{Y^{(1)}}(y;\mu^{(1)}) = \int g_{Y^{(1)}|X}(y|x)g_X(x)dx = \int g_{Y|TX}(y|1,x)g_X(x)dx. \quad (2.7)$$

The second equality holds from the triple assumption in Section 2.1. We often write $f_{Y^{(1)}}(y;\mu^{(1)})$ as $f_1(y)$ for simplicity.

Under contamination, the IPW estimating equation is severely biased

even if the true PS is obtained as $\pi(X|\alpha^*) = P(T=1|X)$:

$$\mathbb{E}_{\tilde{g}}\left[\frac{T}{\pi(X|\alpha^*)}(Y-\mu^{(1)})\right] = \mathbb{E}_g\left[\varepsilon_1(X)\mathbb{E}_{-g+\delta}\left[(Y-\mu^{(1)})|X\right]\right] \neq 0. \quad (2.8)$$

It is found that the remaining term contains the expectation of Y with respect to δ, which implies the estimating equation is severely affected by outliers. The DR estimating equation is also biased. To estimate $\mu^{(1)}$ accurately, we have to remove the influence of contamination.

3. Outlier-Resistant Extensions of IPW and DR

Before we propose novel estimators, we introduce an assumption on outliers. Intuitively, we assume that the outliers are sufficiently far from the main outcome density. Figure 2 shows real examples of outliers that satisfy this assumption. It is found that these outliers are far from the main body of the density both conditionally and marginally.

To formulate this assumption, we introduce density power weighting. The density power weight is used to enhance the outlier resistance in non-causal settings (Windham, 1995; Basu et al., 1998; Jones et al., 2001; Fujisawa and Eguchi, 2008). Let $h(y;\mu)^\gamma$ $(\gamma > 0)$ be a density power weight for $Y^{(1)}$, where $h(y;\mu)$ is a symmetric density function with the location

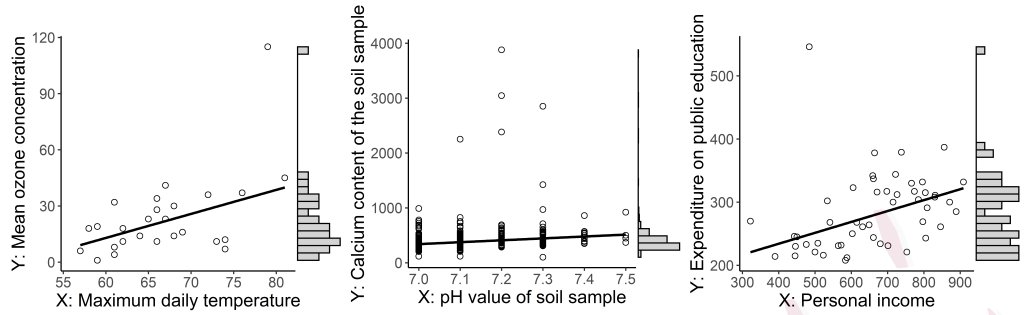

Figure 2: Real examples of outliers which satisfy Assumption 1. All datasets are included in the R package "robustbase" (Maechler et al., 2021): airmay (left), condroz (center), education (right).

parameter μ. The density $h(y;\mu^{(1)})$ is not necessarily equal to the true marginal density $f_1(y)$, but we assume that both h and the true density $f_1(y)$ are symmetric about $\mu^{(1)}$. The assumption of symmetry is common in outlier-resistant estimation, and it is a prerequisite to use the sample median as an estimator of the population mean. Any symmetric density is available for h as long as it satisfies Assumption 1 below. Typically, we assume h is Gaussian. The tuning parameter γ controls the variability of the weight; this leads to the trade-off between outlier resistance and asymptotic efficiency. Assumption 1 formally describes the assumption on outliers.

Assumption 1. Let $h(y;\mu)$ be a weighting density symmetric about μ.

Then, there exists $\gamma > 0$ such that

$$\xi_1(X, \gamma) = \int \delta_1(y|X) h(y; \mu^{(1)})^\gamma (y - \mu^{(1)}) dy \approx 0 \quad a.e. \tag{3.9}$$

Denote an arbitrary bounded function by $\phi(x)$. Assumption 1 implies

$$\nu_1(\phi) := \mathbb{E}[\phi(X)\xi_1(X, \gamma)] = \int \phi(x)\xi_1(x, \gamma) g_X(x) dx \approx 0. \tag{3.10}$$

In particular, let $\phi(x) = 1$; then, the outliers are marginally negligible:

$$\nu_1(1) = \mathbb{E}[\xi_1(X, \gamma)] = \int \delta_1(y) h(y; \mu^{(1)})^\gamma (y - \mu^{(1)}) dy \approx 0. \tag{3.11}$$

Throughout this paper, we assume that γ is sufficiently large so that Assumption 1 holds. Assumption 1 is reduced to a simpler form when $\delta_1(y|X)$ is Dirac's delta at y_0; this is one of the core assumptions in Section 4.

Assumption 1a. Let $h(y; \mu)$ be a weighting density that is symmetric about μ, and assume that the density of outliers is Dirac's delta at y_0 ($\neq \mu^{(1)}$), say $\delta_{y_0}(y)$. Then, there exists $\gamma > 0$ such that

$$\int \delta_{y_0}(y) h(y; \mu^{(1)})^\gamma (y - \mu^{(1)}) dy = h(y_0; \mu^{(1)})^\gamma (y_0 - \mu^{(1)}) \approx 0. \tag{3.12}$$

For example, if h is a Gaussian density with mean $\mu^{(1)}$ and fixed variance, (3.12) tends to 0 as $|y_0| \to \infty$ for fixed $\gamma > 0$ since $h(y_0; \mu^{(1)})^\gamma (y_0 - \mu^{(1)}) \propto \exp\left(-\gamma (y_0 - \mu^{(1)})^2\right)(y_0 - \mu^{(1)})$.

3.1 IPW-type Estimator

First, we introduce an extension of the IPW estimator, called the density power inverse probability weighting (DP-IPW) estimator. The DP-IPW estimator is defined as a root of the following estimating equation:

$$\sum_{i=1}^{n} \frac{T_i}{\pi(X_i; \hat{\alpha})} h(Y_i; \mu)^\gamma (Y_i - \mu) = 0. \tag{3.13}$$

Under no contamination, the DP-IPW estimating equation is unbiased.

Theorem 1. *Assume that the true propensity score $\pi(X; \alpha^*)$ is given. Then, under no contamination, we have*

$$\mathbb{E}_g \left[\frac{T}{\pi(X; \alpha^*)} h(Y; \mu^{(1)})^\gamma (Y - \mu^{(1)}) \right] = 0. \tag{3.14}$$

Only an estimate $\pi(X; \hat{\alpha})$ is available in practice, but the asymptotic consistency of (DP-)IPW still holds if the model $\pi(X; \alpha)$ is correctly specified.

Now we consider the contaminated case. The bias of the DP-IPW estimating equation takes a different form from (2.8).

Theorem 2. *Assume Y is contaminated as* (2.6)*. Under the same assumptions as those in Theorem 1, the expectation of the DP-IPW estimating equation is expressed as*

$$\mathbb{E}_{\tilde{g}}\left[\frac{T}{\pi(X;\alpha^*)}h(Y;\mu^{(1)})^{\gamma}(Y-\mu^{(1)})\right] = B_1 + \nu_1(\varepsilon_1), \tag{3.15}$$

$$\textit{where}\quad B_1 = -\int \varepsilon_1(x)\int h(y;\mu^{(1)})^{\gamma}(y-\mu^{(1)})g(y|x)dy\ g(x)dx.$$

In particular, under homogeneous contamination, B_1 reduces to 0.

The DP-IPW estimating equation is still biased even if $\nu_1(\varepsilon_1)$ is small. Since we assume that $\nu_1(\varepsilon_1)$ is negligible, B_1 is dominant. However, compared to (2.8), the dominant bias of DP-IPW does not contain δ_1. This implies that the bias of DP-IPW is not strongly affected by the absolute value of outliers. Under homogeneous contamination, the dominant term disappears, so the bias is negligible.

3.2 DR-type Estimator

Next, we introduce the density power doubly robust (DP-DR) estimator. The DP-DR estimator is a straight application of the DR M-estimator and

defined as a root of the following estimating equation:

$$\sum_{i=1}^{n}\left[\frac{T_i h(Y_i;\mu)^\gamma}{\pi(X_i;\hat{\alpha})}(Y_i-\mu)-\frac{T_i-\pi(X_i;\hat{\alpha})}{\pi(X_i;\hat{\alpha})}\mathbb{E}_{\hat{q}}\left[h(Y;\mu)^\gamma(Y-\mu)|T=1,X\right]\right]=0. \tag{3.16}$$

As we have discussed in Section 2.1, $\mathbb{E}_{\hat{q}}\left[h(Y;\mu)^\gamma(Y-\mu)|T=1,X\right]$ is obtained by direct calculation or Monte Carlo approximation based on the parametric OR model $\hat{q}:=q(y|T=1,X;\hat{\beta})$. In the Appendix, we present the explicit forms of $\mathbb{E}_{\hat{q}}\left[h(Y;\mu)^\gamma(Y-\mu)|T=1,X\right]$ when h and q are assumed to be Gaussian. The parameter β is usually estimated in an outlier-resistant manner: Huber regression (Huber, 2004, Chap.7), MM estimator (Yohai, 1987), density power regression (Basu et al., 1998; Kanamori and Fujisawa, 2015), and γ-regression (Fujisawa and Eguchi, 2008; Kawashima and Fujisawa, 2017), for example.

The DP-DR estimator is doubly robust under no contamination as with the general DR M-estimator.

Theorem 3. *Assume either the true PS or the true OR model is given. Then, if there is no contamination, the DP-DR estimating equation is unbiased.*

Now, we evaluate the bias of the DP-DR estimating equation under

contamination.

Theorem 4. *Assume that Y is contaminated as* (2.6)*. If the true PS model is given, the expectation of the DP-DR estimating equation is expressed as*

$$-\int \varepsilon_1(x) \int h(y;\mu^{(1)})^{\gamma}(y-\mu^{(1)})g(y|x)dy \; g(x)dx + \nu_1(\varepsilon_1). \tag{3.17}$$

In particular, under homogeneous contamination, (3.17) *reduces to $\nu_1(\varepsilon_1)$. If the true OR model is given, the expectation of the DP-DR estimating equation is expressed as*

$$-\int \varepsilon_1(x)\frac{P(T=1|x)}{\pi(x;\alpha)} \int h(y;\mu^{(1)})^{\gamma}(y-\mu^{(1)})g(y|x)dy \; g(x)dx + \nu_1(\varepsilon_1 P(T=1|\cdot)/\pi(\cdot;\alpha)). \tag{3.18}$$

Under homogeneous contamination, (3.18) *becomes*

$$-\varepsilon_1 \int \frac{P(T=1|x)}{\pi(x;\alpha)} \int h(y;\mu^{(1)})^{\gamma}(y-\mu^{(1)})g(y|x)dy \; g(x)dx + \nu_1(\varepsilon_1 P(T=1|\cdot)/\pi(\cdot;\alpha)). \tag{3.19}$$

Assuming that $\pi(\cdot;\alpha)$ is bounded away from 0 and 1, we find that $P(T=1|\cdot)/\pi(\cdot;\alpha)$ is bounded. Then, from Assumption 1, $\nu_1(\varepsilon_1 P(T=1|\cdot)/\pi(\cdot;\alpha))$ is negligible. As with DP-IPW, the dominant bias is independent of δ, in-

dicating that the influence of outliers is reduced. Unfortunately, DP-DR is still biased in the PS-incorrect and OR-correct case even under homogeneous contamination because the dominant term of (3.19) is not eliminated.

In the OR-correct case, the reason why DP-DR is biased under homogeneous contamination is as follows. Under Assumption 1, the expectation of the DP-DR estimating function becomes

$$\begin{aligned}
&\mathbb{E}_g\left[\frac{P(T=1|X)}{\pi(X;\alpha)}\left\{\mathbb{E}_{\tilde{g}}[\psi(Y^{(1)};\mu^{(1)})|X]-\mathbb{E}_g[\psi(Y^{(1)};\mu^{(1)})|X]\right\}\right]\\
\approx\ &\mathbb{E}_g\left[\frac{P(T=1|X)}{\pi(X;\alpha)}\left\{(1-\varepsilon_1)\mathbb{E}_g[\psi(Y^{(1)};\mu^{(1)})|X]-\mathbb{E}_g[\psi(Y^{(1)};\mu^{(1)})|X]\right\}\right],
\end{aligned}$$

where we denote the density power estimating function by ψ. In the last formula, it is found that the terms in the curly brackets do not cancel because the first term is reduced by $1-\varepsilon_1$. Based on this consideration, we propose a bias-corrected version of DP-DR, called the εDP-DR estimator. εDP-DR is designed to cancel the dominant bias under homogeneous contamination. The εDP-DR estimator is a root of the following estimating equation:

$$\sum_{i=1}^{n}\left[\frac{T_i h(Y_i;\mu)^\gamma}{\pi(X_i;\hat{\alpha})}(Y_i-\mu)-\frac{T_i-\pi(X_i;\hat{\alpha})}{\pi(X_i;\hat{\alpha})}(1-\hat{\varepsilon}_1)\mathbb{E}_{\hat{q}}\left[h(Y;\mu)^\gamma(Y-\mu)|T=1,X\right]\right]=0, \tag{3.20}$$

where $\hat{\varepsilon}_1$ is a consistent estimator of the expected contamination ratio $\overline{\varepsilon}_1 = \int \varepsilon_1(x)g(x)dx$. $\hat{\varepsilon}_1$ can be obtained simultaneously with the parametric OR model by the unnormalized modeling with the density power score (Kanamori and Fujisawa, 2015), for example. While DP-DR is a special case of the DR M-estimator, εDP-DR goes beyond this framework by the bias correction. Under no contamination, the εDP-DR estimating equation is asymptotically identical to the DP-DR estimating equation. The εDP-DR estimating equation is also biased under heterogeneous contamination; however, the bias takes a different form.

Corollary 1. *If the true PS model is given, the expectation of the εDP-DR estimating equation is equal to* (3.17)*. If the true OR model is given, the expectation of the εDP-DR estimating equation is expressed as*

$$\mathbb{E}_g\left[(\overline{\varepsilon}_1 - \varepsilon_1(X))\frac{P(T=1|X)}{\pi(X;\alpha)}\mathbb{E}_g[h(Y^{(1)};\mu^{(1)})^\gamma(Y^{(1)}-\mu^{(1)})|X]\right] + \nu_1(\varepsilon_1 P(T=1|\cdot)/\pi(\cdot;\alpha)). \tag{3.21}$$

The first term disappears under homogeneous contamination.

Proof. The derivation is the same as that of Theorem 4. If $\varepsilon_1(X)$ is constant ε_1, the first term disappears because $\overline{\varepsilon}_1 = \varepsilon_1 \int g(x)dx = \varepsilon_1$. □

Similar to (3.19), the second term of (3.21) is approximately zero if we

assume that $\pi(\cdot;\alpha)$ is bounded away from 0 and 1.

Remark One may find that "$\varepsilon(X)$"DP-DR would work better than εDP-DR under heterogeneous contamination. In fact, the bias (3.21) will disappear if we replace $\bar{\varepsilon}$ with $\varepsilon(X)$. However, it is necessary to model $\varepsilon(X)$ correctly for consistent estimation of "$\varepsilon(X)$"DP-DR. To the best of our knowledge, no easy method is available for this purpose.

3.3 Summary

We have proposed three types of outlier resistant semiparametric estimators: DP-IPW, DP-DR, and εDP-DR. Table 1 shows the bias of the estimating equations under the conditions discussed above. Under heterogeneous contamination, all estimators are biased, but the biases are hardly influenced by the absolute value of outliers. Furthermore, as discussed in Section 4, outliers have negligible influence if the contamination ratio is sufficiently small. εDP-DR improves DP-DR in the OR-correct case under homogeneous contamination, but we continue to discuss DP-DR for three reasons: the contamination ratio is sometimes hard to estimate, the bias (3.19) is not serious if $\pi(X;\alpha)$ is close to $P(T=1|X)$, and the simulation results presented in Section 6 indicate that DP-DR remains better than the existing methods even in the OR-correct case.

Contamination	model	DP-IPW	DP-DR	εDP-DR
No contamination	PS-correct	0	0	0
	OR-correct	-	0	0
homogeneous: ε	PS-correct	≈ 0	≈ 0	≈ 0
	OR-correct	-	$\approx \varepsilon\mathbb{E}[\phi(X)]$	≈ 0
heterogeneous: $\varepsilon(X)$	PS-correct	$\approx \mathbb{E}[\varepsilon(X)\phi(X)]$	$\approx \mathbb{E}[\varepsilon(X)\phi(X)]$	$\approx \mathbb{E}[\varepsilon(X)\phi(X)]$
	OR-correct	-	$\approx \mathbb{E}[\varepsilon(X)\phi(X)]$	$\approx \mathbb{E}[(\bar{\varepsilon} - \varepsilon(X))\phi(X)]$

Table 1: Summary of the biases of the proposed estimating equations. The function $\phi(X)$ differs cell-by-cell. PS-correct means that the PS model is correctly specified and the OR model may not be; OR-correct means the opposite.

4. Influence-function-based Analysis of Outlier Resistance

As discussed in the previous section, the proposed estimators suffer less from outliers compared with ordinary estimators from the viewpoint of the unbiasedness of the estimating equation. In this section, we demonstrate that they are outlier-resistant from the viewpoint of the IF.

Here, we briefly review the IF for the univariate M-estimator and expand it to evaluate our estimators. Let G be the distribution of $Z \in \mathbb{R}$, and let $T(G)$ be a functional of G, which is the parameter of interest. The IF of $T(G)$ is defined as

$$IF(z_0; G) := \lim_{\varepsilon \to 0} \frac{T((1-\varepsilon)G + \varepsilon\Delta_{z_0}) - T(G)}{\varepsilon} = \frac{\partial}{\partial \varepsilon}\{T((1-\varepsilon)G + \varepsilon\Delta_{z_0}) - T(G)\}\Big|_{\varepsilon=0}, \tag{4.22}$$

where Δ_{z_0} is a degenerate distribution at z_0. We also see that the latent bias $T((1-\varepsilon)G+\varepsilon\Delta_{z_0})-T(G)$ can be approximated by $\varepsilon IF(z_0;G)$. Therefore, the behavior of the IF approximates that of the latent bias. In the population, the M-estimator $T_M(G)$ satisfies $\int \psi(z,T_M(G))dG(z)=0$. Then, the IF for $T_M(G)$ is obtained by differentiating $\int \psi(z,T_M((1-\varepsilon)G+\varepsilon\Delta_{z_0})d\{(1-\varepsilon)G+\varepsilon\Delta_{z_0}\}(z)=0$ with respect to ε. This yields

$$IF(z_0;G)=-\mathbb{E}\left[\frac{\partial}{\partial\eta}\psi(Z,\eta)\Big|_{\eta=T_M(G)}\right]^{-1}\psi(z_0,T_M(G)). \tag{4.23}$$

The function ψ is said to have a redescending property if $\psi(z_0,T_M(G))$ approaches zero as the outlier $|z_0|$ increases. Therefore, when ψ has the redescending property and z_0 is an outlier, the latent bias is sufficiently small. This is favorable for outlier resistance.

Since ε_1 is dependent on X, we cannot apply the IF directly to our estimators. To overcome this issue, we consider the IF with fixed covariates $\{X_i\}_{i=1}^n$; this approach is similar to the fixed carrier model in Hampel et al. (2011, Chap.6). Consider the following estimating equation:

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{\tilde{g}}\left[\psi(Y,T,X_i;\mu)\middle|\,X_i\right]=0. \tag{4.24}$$

If the fixed sample $\{X_i\}_{i=1}^n$ consists of i.i.d. observations, then the left-hand

side of (4.24) converges to $\mathbb{E}_{\tilde{g}}[\psi(Y,T,X;\mu)]$ as $n \to \infty$. Let $\tilde{\mu}_n^{(1)}$ denote a root of (4.24), and let $\tilde{\mu}^{(1)}$ be a root of $\mathbb{E}_{\tilde{g}}[\psi(Y,T,X;\mu)]$. Then, $\tilde{\mu}_n^{(1)}$ also converges to $\tilde{\mu}^{(1)}$. Therefore, $\tilde{\mu}_n^{(1)}$ shows roughly the same behavior as that of the target estimator $\tilde{\mu}^{(1)}$. The contaminated density $\tilde{g}$ is defined as (2.6), and $\delta_1(y|X_i)$ is assumed to be Dirac's delta at y_0. The IF of $T_n(\tilde{G})$ at X_i is obtained by differentiating (4.24) with respect to $\varepsilon_1(X_i)$ at $\varepsilon_1(X_i) = 0$.

Because of the space, we discuss only the εDP-DR. Assume that $\bar{\varepsilon}_1 = \frac{1}{n}\sum_{i=1}^n \varepsilon_1(X_i)$, then the IF of εDP-DR is

$$
\begin{aligned}
&-\mathbb{E}_g\left[\left.\frac{\partial\psi}{\partial\mu}\right|_{\mu=\mu_n^{(1)}}\middle| X_i\right]^{-1}\left[\frac{P(T=1|X_i)}{\pi(X_i;\alpha)}h(y_0-\mu_n^{(1)})^\gamma(y_0-\mu_n^{(1)})\right.\\
&\left.-\frac{n-1}{n}\frac{P(T=1|X_i)-\pi(X_i;\alpha)}{\pi(X_i;\alpha)}\mathbb{E}_{\hat{q}}\left[h(Y;\mu_n^{(1)})^\gamma(Y-\mu_n^{(1)})|T=1,X\right]\right].
\end{aligned}
\tag{4.25}
$$

In the PS-correct case, the second term in square brackets is equal to zero, and then the IF tends to zero as $|y_0| \to \infty$. In the OR-correct case, the second term does not disappear. Considering the limit of $|y_0| \to \infty$, the IF converges to

$$
\frac{n-1}{n}\mathbb{E}_g\left[\left.\frac{\partial\psi}{\partial\mu}\right|_{\mu=\mu_n^{(1)}}\middle| X_i\right]^{-1}\left[\frac{P(T=1|X_i)-\pi(X_i;\alpha)}{\pi(X_i;\alpha)}\mathbb{E}_{\hat{q}}[h(Y;\mu_n^{(1)})^\gamma(Y-\mu_n^{(1)})|T=1,X_i]\right].
\tag{4.26}
$$

Thus, the εDP-DR estimator has the redescending property only in the PS-correct case. In the OR-correct case, the influence cannot be eliminated, but the IF tends to a constant when $|y_0|$ tends to infinity, implying that the influence of the outlier is not serious. DP-DR has an IF similar to that of εDP-DR, and DP-IPW has an IF similar to that of εDP-DR whose PS is correct. The derivations of all IFs are presented in the Appendix.

Under homogeneous contamination, the ordinary IF is applicable, and we can see that the proposed estimators have the redescending property in the PS-correct case. Besides, εDP-DR has the redescending property even in the OR-correct case; this result is consistent with Corollary 1. The IF-based analysis under homogeneous contamination is presented in the Appendix.

5. Asymptotic Properties

We discuss the asymptotic properties of the εDP-DR estimator. For the other proposed estimators, we obtain similar results with small changes. The asymptotic properties can be obtained in a manner similar to that described in Hoshino (2007). Assume that the PS and OR models are regular and are estimated consistently if the models are correctly specified. Furthermore, the contamination ratio ε_1 is known. Note that when the

contamination ratio is consistently estimated simultaneously with the OR model by Kanamori and Fujisawa (2015), we can replace β with $(\varepsilon_1, \beta^T)^T$ in the following discussion.

We write (3.20) as $\frac{1}{n}\sum_{i=1}^n \psi_i(\mu; \hat{\alpha}, \hat{\beta})$, and let $\frac{1}{n}\sum_{i=1}^n s_i^{PS}(\alpha) = 0$ and $\frac{1}{n}\sum_{i=1}^n s_i^{OR}(\beta) = 0$ be the estimating equations for the PS and OR models, respectively. Let $\lambda = (\mu, \alpha^T, \beta^T)^T$ be the parameter vector, and let the full estimating equation be defined as

$$\sum_{i=1}^n S_i(\lambda) = \sum_{i=1}^n \begin{pmatrix} \psi_i(\mu; \alpha, \beta) \\ s_i^{PS}(\alpha) \\ s_i^{OR}(\beta) \end{pmatrix} = \mathbf{0}. \tag{5.27}$$

Let $\lambda^* = (\mu^*, \alpha^{*T}, \beta^{*T})^T$ be a root of (5.27) in population. Note that, in this section, $*$ does not necessarily mean that the model is correctly specified. With the results presented in Van der Vaart (2000, Chap.5), the following theorem holds under some regularity conditions.

Theorem 5. *Under the regularity conditions presented in the Appendix,*

the following asymptotic properties hold:

$$\hat{\lambda} \xrightarrow{p} \lambda^*, \tag{5.28}$$

$$\sqrt{n}(\hat{\lambda} - \lambda^*) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \mathbf{V}^{\tilde{g}}(\lambda^*)\right), \tag{5.29}$$

where $\mathbf{V}^{\tilde{g}}(\lambda^*) = \mathbf{J}^{\tilde{g}}(\lambda^*)^{-1}\mathbf{K}^{\tilde{g}}(\lambda^*)\{\mathbf{J}^{\tilde{g}}(\lambda^*)^T\}^{-1}$, $\mathbf{J}^{\tilde{g}}(\lambda^*) = \mathbb{E}_{\tilde{g}}\left[\partial S_i(\lambda^*)/\partial\lambda^T\right]$, *and* $\mathbf{K}^{\tilde{g}}(\lambda^*) = \mathbb{E}_{\tilde{g}}\left[S_i(\lambda^*)S_i(\lambda^*)^T\right]$.

By using this and applying the results presented in Section 3.2, we find that the limit μ^* is in the neighborhood of $\mu^{(1)}$.

Theorem 6. *Let* $\lambda^{**} = (\mu^{(1)}, \alpha^{*T}, \beta^{*T})^T$ *and assume that* $\mathbf{J}_{11}^{\tilde{g}}(\lambda)$ *is nonzero within the interval* $[\lambda^*, \lambda^{**}]$. *Under Assumption 1 and homogeneous contamination, if either the PS or the OR model is correct, it then holds that*

$$\mu^* = \mu^{(1)} + \mathcal{O}(\nu_1(\phi)), \tag{5.30}$$

where $\phi(\cdot) = \varepsilon_1$ *(constant) in the PS-correct case and* $\phi(\cdot) = \varepsilon_1 P(T = 1|\cdot)/\pi(\cdot;\alpha)$ *in the OR-correct case.*

The proof of Theorem 6 and further discussions on the asymptotic variance are available in the Appendix.

6. Monte-Carlo Simulation

We conduct Monte-Carlo simulations to evaluate the performance of the proposed estimators. Our methods are compared with the naive IPW and DR estimators and some existing outlier-resistant methods (Firpo, 2007; Zhang et al., 2012; Díaz, 2017; Sued, Valdora, and Yohai, 2020). Since these methods focus on the median of the potential outcome, they are resistant to outliers at a certain level; but the median-based methods are not so resistant to heavy contamination. To the best of our knowledge, no method other than the proposed method has more outlier resistance than the median. Firpo's IPW estimator (Firpo, 2007) is defined as

$$\hat{\mu}_{\text{Firpo}} = \arg\min_{\mu} \sum_{i=1}^{n} \frac{T_i}{\pi(X_i;\hat{\alpha})}(Y_i - \mu)(0.5 - \mathbb{I}(Y_i \leq \mu)), \tag{6.31}$$

where the function $\mathbb{I}$ is an indicator function. Zhang's IPW median (Zhang et al., 2012) is based on the IPW-empirical distribution. Firpo's IPW and Zhang's IPW are almost equivalent except for a slight difference in their computation. Zhang's and Sued's DR methods (Zhang et al., 2012; Sued, Valdora, and Yohai, 2020) estimate the empirical distribution in a doubly robust way. They incorporate an IPW-type estimator into the first term. The remaining term of Zhang's DR is based on the Gaussian cumulative dis-

tribution function of Y given X. In contrast, Sued's DR constructs the remaining term in a nonparametric manner. Diaz's DR median (Díaz, 2017) is a different approach; it employs the targeted maximum likelihood estimator (TMLE) (Van Der Laan and Rubin, 2006). We implemented our methods, Zhang's IPW/DR, and Sued's DR in R. For Firpo's IPW and TMLE, we used the *causalquantile* package (https://github.com/idiazst/causalquantile; Updated on 31 Aug 2017).

6.1 Numerical Algorithm for the Proposed Methods

Since the proposed estimating equations cannot be solved explicitly, we develop an iterative algorithm. Various algorithms are available, but we propose a standard algorithm for M-estimators (Huber, 2004; Hampel et al., 2011). Detailed algorithm is available in the Appendix. Hereafter, we suppose h and q are Gaussian. We also provide explicit updating formulae in this case. Note that some additional parameters of h should be estimated in a roughly unbiased and outlier-resistant way.

6.2 Simulation Model

We simulated random observations based on a simple causal setting. The confounders (X_1, X_2) were independently drawn from a Gaussian or uni-

form distribution with mean zero and unit variance. The treatment T was assigned along with the conditional probability $P(T = 1|X_1, X_2)$ that was defined as a sigmoid function of $0.8X_1 + 0.2X_2$. The potential outcomes $(Y^{(1)}, Y^{(0)})$ were generated according to a linear function of (X_1, X_2) with Gaussian error: $Y^{(1)} = \mu^{(1)} + 1.2X_1 + 0.3X_2 + e$ and $Y^{(0)} = \mu^{(0)} + 1.2X_1 + 0.3X_2 + e$. $\mu^{(1)}$ and $\mu^{(0)}$ were set to 3 and 0, respectively. The standard deviation (SD) of e was set to $\sqrt{0.72}$; then, $\mathrm{SD}[Y^{(1)}] = \mathrm{SD}[Y^{(0)}] = 1.5$. When the confounders were not Gaussian, the potential outcomes were not Gaussian. The observed outcome Y was defined as $Y = TY^{(1)} + (1 - T)Y^{(1)}$ under no contamination. Outliers were drawn from $\mathcal{N}(\mu^{(t)} + 10\sigma^{(t)}, 1)$, with $\sigma^{(t)} = \mathrm{SD}[Y^{(t)}] = 1.5$. For the homogeneous contamination settings, the contamination ratio was set to be a constant ε_t. For the heterogeneous contamination settings, the contamination ratio was set to be $1.5\varepsilon_t$ if $X_1 + X_2 \leq 0$ and $0.5\varepsilon_t$ if $X_1 + X_2 > 0$. The average contamination ratio is set to $\varepsilon_t \in \{0, 0.05, 0.1, 0.2\}$. Then, the observations of Y were randomly replaced with outliers according to the contamination ratio. The sample size was fixed to $n = 100$ throughout the Monte Carlo simulations. Furthermore, we generated datasets in which the outcome follows a symmetric and heavy-tailed distribution. We drew the error term of $Y^{(t)}$ from the standard Cauchy distribution instead of inserting outliers.

6.3 Results

First, we performed a comparative study. The potential mean $\mu^{(1)}$ was estimated using the proposed and comparative methods. In this experiment, we used all settings illustrated in the previous section. The propensity score was estimated by logistic regression. The parametric OR was conducted in two ways: Gaussian MLE with non-outliers or unnormalized Gaussian modeling (the tuning parameter was set to 0.5) (Kanamori and Fujisawa, 2015). For the DR estimators, we investigated three patterns of model misspecification: PS-correct/OR-correct, PS-correct/OR-incorrect, and PS-incorrect/OR-correct. For the model-correct case, we included an intercept and (X_1, X_2) as covariates. For the model-incorrect case, we included only an intercept and X_2. We performed 10,000 simulations for every setting and method. Tables 2 and 3 show the results of the comparative study when the covariates were Gaussian and the OR for the DR-type estimators was the Gaussian MLE with non-outliers. The estimation error was measured by the root mean square error (RMSE). The mean and SD of all estimates, the mean computation time, and the results for the other settings are provided in the Appendix. In Table 2, the naive IPW estimator had a significantly larger RMSE under contamination. Both the median-based methods and DP-IPW dramatically reduced the RMSE.

As the contamination ratio increased, the RMSE increased. The RMSE tended to be larger for heterogeneous contamination than for homogeneous contamination. When the optimal γ was chosen, the proposed method outperformed the comparative methods and had the smallest RMSE for all settings. Looking at Table 3, the results for the DR-type estimators were similar to those for the IPW estimators. The proposed method with a proper γ outperformed the comparative methods and had the smallest RMSE in all settings. DP-DR and εDP-DR performed similarly, although εDP-DR was slightly superior in many settings. Among the median-based methods, TMLE performed better, but it took much more time than the other methods, including the proposed methods, and occasionally ($< 1\%$) failed to converge.

		Homogeneous			**Heterogeneous**		
ε	0.00	0.05	0.10	0.20	0.05	0.10	0.20
Naive	0.222	0.957	1.683	3.153	0.993	1.752	3.253
median (Firpo)	0.257	0.294	0.367	0.649	0.306	0.409	0.769
median (Zhang-IPW)	0.257	0.294	0.367	0.649	0.306	0.409	0.769
DP-IPW ($\gamma = 0.1$)	0.218	0.276	0.531	2.263	0.293	0.609	2.377
DP-IPW ($\gamma = 0.5$)	0.227	0.249	0.272	0.639	0.245	0.287	0.726
DP-IPW ($\gamma = 1.0$)	0.261	0.271	0.275	0.413	0.262	0.281	0.498

Table 2: RMSE of the IPW-type estimators. X was drawn from Gaussian distributions.

Table 4 shows the RMSE of each method on the data with Cauchy error. As well as the above experiments, the proposed method performed

		Homogeneous			**Heterogeneous**		
ε	0.00	0.05	0.10	0.20	0.05	0.10	0.20
(PS-correct/OR-correct)							
Naive	0.184	0.957	1.684	3.154	0.997	1.758	3.265
median (Zhang-DR)	0.239	0.317	0.391	0.733	0.330	0.452	0.905
median (Sued)	0.238	0.316	0.388	0.693	0.329	0.450	0.869
median (TMLE)	0.237	0.280	0.359	0.603	0.295	0.402	0.701
DP-DR ($\gamma = 0.1$)	0.183	0.302	0.564	2.262	0.318	0.649	2.394
DP-DR ($\gamma = 0.5$)	0.202	0.285	0.326	0.697	0.274	0.349	0.834
DP-DR ($\gamma = 1.0$)	0.240	0.288	0.307	0.524	0.287	0.336	0.669
εDP-DR ($\gamma = 0.1$)	0.183	0.296	0.554	2.255	0.314	0.636	2.385
εDP-DR ($\gamma = 0.5$)	0.202	0.264	0.302	0.669	0.271	0.323	0.793
εDP-DR ($\gamma = 1.0$)	0.240	0.287	0.299	0.513	0.286	0.335	0.648
(correct/incorrect)							
Naive	0.237	0.963	1.686	3.156	1.001	1.758	3.262
median (Zhang-DR)	0.275	0.342	0.408	0.741	0.350	0.465	0.912
median (Sued)	0.275	0.342	0.407	0.699	0.350	0.464	0.872
median (TMLE)	0.242	0.284	0.363	0.622	0.297	0.404	0.719
DP-DR ($\gamma = 0.1$)	0.237	0.314	0.561	2.267	0.330	0.644	2.393
DP-DR ($\gamma = 0.5$)	0.247	0.319	0.349	0.714	0.319	0.361	0.839
DP-DR ($\gamma = 1.0$)	0.280	0.334	0.347	0.581	0.329	0.372	0.709
εDP-DR ($\gamma = 0.1$)	0.237	0.311	0.557	2.264	0.328	0.640	2.388
εDP-DR ($\gamma = 0.5$)	0.247	0.317	0.344	0.694	0.313	0.356	0.817
εDP-DR ($\gamma = 1.0$)	0.280	0.333	0.338	0.551	0.327	0.369	0.708
(incorrect/correct)							
Naive	0.181	0.879	1.591	3.026	0.826	1.490	2.813
median (Zhang-DR)	0.237	0.263	0.316	0.503	0.269	0.337	0.548
median (Sued)	0.236	0.272	0.346	0.599	0.277	0.364	0.627
median (TMLE)	0.234	0.260	0.309	0.478	0.265	0.328	0.522
DP-DR ($\gamma = 0.1$)	0.182	0.192	0.345	2.057	0.191	0.299	1.681
DP-DR ($\gamma = 0.5$)	0.199	0.206	0.218	0.366	0.203	0.209	0.283
DP-DR ($\gamma = 1.0$)	0.230	0.232	0.239	0.273	0.230	0.233	0.242
εDP-DR ($\gamma = 0.1$)	0.182	0.193	0.381	2.207	0.194	0.335	1.839
εDP-DR ($\gamma = 0.5$)	0.199	0.203	0.208	0.376	0.203	0.212	0.318
εDP-DR ($\gamma = 1.0$)	0.230	0.230	0.231	0.243	0.231	0.237	0.260

Table 3: RMSE of the DR-type estimators. X was drawn from Gaussian distributions.

better than the comparative methods. In this setting, we only used the unnormalized Gaussian modeling for OR for the DR-type estimators. Only in the PS-correct/OR-incorrect case, the median (TMLE) performed slightly better than the proposed method.

	IPW
Naive	274.024
median (Firpo)	0.414
median (Zhang-IPW)	0.414
DP-IPW ($\gamma = 0.1$)	0.443
DP-IPW ($\gamma = 0.5$)	0.367
DP-IPW ($\gamma = 1.0$)	0.380

		DR	
PS/OR	correct/correct	correct/incorrect	incorrect/correct
Naive	275.447	275.446	263.629
median (Zhang-DR)	0.415	0.456	0.390
median (Sued)	0.408	0.436	0.373
median (TMLE)	0.392	0.394	0.389
DP-DR ($\gamma = 0.1$)	0.501	0.514	0.390
DP-DR ($\gamma = 0.5$)	0.363	0.404	0.358
DP-DR ($\gamma = 1.0$)	0.372	0.418	0.364
εDP-DR ($\gamma = 0.1$)	0.487	0.503	0.377
εDP-DR ($\gamma = 0.5$)	0.361	0.399	0.328
εDP-DR ($\gamma = 1.0$)	0.370	0.412	0.334

Table 4: RMSE of the comparative study using the heavy-tailed data. The covariate X is drawn from Gaussian distribution. The OR model for the DR-type estimators were obtained by the unnormalized Gaussian modeling.

Next, we conduct a γ-sensitivity study. $\mu^{(1)}$ was estimated by the proposed methods with different γs. X had a Gaussian distribution, and the contamination ratio varied in $\{0, 0.05, 0.1, 0.2\}$ under homogeneous contamination. For the DR-type estimators, the outcome regression was performed by the Gaussian MLE using nonoutliers. We simulated 10,000 datasets for

every setting and method. Table 5 shows the results of the γ-sensitivity study. As in the comparative study, when the ratio of outliers increased, the bias increased. Larger γ resulted in increased variance. When the contamination ratio was small, it was sufficient to use a small γ such as $\gamma = 0.1$ or 0.2 to remove the adverse effect of outliers. Even in highly contaminated cases, $\gamma > 1.0$ was not needed. Comparing the estimates of DP-DR and εDP-DR in the PS-incorrect/OR-correct case, it can be found that the DP-DR estimates were biased especially when ε was large, and contrarily, the εDP-DR estimates were almost equal to the true value 3. This result shows that the bias correction by $1 - \hat{\varepsilon}$ worked well in our experiments.

As in many other outlier-resistant statistical methods, parameter tuning is challenging. We suggest a possible policy on this issue based on the solution paths of the proposed estimators, which is provided in the Appendix. Looking at the paths, the influence of outliers decreased as γ increased, and the paths became stable around the true value after reaching a certain γ. Thus, we suggest using the smallest γ for which the estimate is stable.

	PS/OR	ε	$\gamma = 0.0$	0.1	0.2	0.5	1.0	1.5	2.0
DP-IPW	T/-	0.00	3.004 (0.22)	2.998 (0.22)	2.994 (0.22)	2.986 (0.23)	2.980 (0.26)	2.974 (0.30)	2.970 (0.34)
		0.05	3.749 (0.59)	3.030 (0.27)	2.999 (0.26)	2.987 (0.25)	2.978 (0.27)	2.970 (0.30)	2.963 (0.33)
		0.10	4.493 (0.78)	3.142 (0.51)	3.015 (0.32)	2.989 (0.27)	2.977 (0.27)	2.969 (0.30)	2.963 (0.33)
		0.20	5.983 (1.02)	4.492 (1.70)	3.536 (1.39)	3.052 (0.64)	2.990 (0.41)	2.978 (0.39)	2.971 (0.40)
DP-DR	T/T	0.00	2.999 (0.18)	2.998 (0.18)	2.997 (0.19)	2.996 (0.20)	2.992 (0.24)	2.989 (0.28)	2.985 (0.31)
		0.05	3.745 (0.60)	3.029 (0.30)	3.002 (0.27)	2.997 (0.29)	2.991 (0.29)	2.985 (0.31)	2.980 (0.34)
		0.10	4.489 (0.79)	3.140 (0.55)	3.017 (0.36)	3.000 (0.33)	2.992 (0.31)	2.986 (0.32)	2.981 (0.33)
		0.20	5.979 (1.04)	4.465 (1.72)	3.532 (1.41)	3.060 (0.69)	3.009 (0.52)	2.999 (0.51)	2.994 (0.51)
	T/F	0.00	3.004 (0.24)	2.998 (0.24)	2.994 (0.24)	2.986 (0.25)	2.979 (0.28)	2.974 (0.32)	2.968 (0.36)
		0.05	3.750 (0.60)	3.033 (0.31)	3.001 (0.29)	2.989 (0.32)	2.978 (0.33)	2.970 (0.36)	2.963 (0.39)
		0.10	4.494 (0.78)	3.150 (0.54)	3.020 (0.37)	2.992 (0.35)	2.979 (0.35)	2.970 (0.37)	2.963 (0.39)
		0.20	5.984 (1.03)	4.490 (1.71)	3.546 (1.41)	3.059 (0.71)	3.001 (0.58)	2.985 (0.55)	2.975 (0.54)
	F/T	0.00	2.999 (0.18)	2.999 (0.18)	2.999 (0.18)	3.001 (0.20)	3.005 (0.23)	3.010 (0.26)	3.014 (0.29)
		0.05	3.725 (0.50)	2.997 (0.19)	2.976 (0.19)	2.975 (0.20)	2.978 (0.23)	2.982 (0.26)	2.986 (0.29)
		0.10	4.451 (0.65)	3.051 (0.34)	2.956 (0.21)	2.950 (0.21)	2.953 (0.23)	2.956 (0.26)	2.960 (0.28)
		0.20	5.902 (0.86)	4.326 (1.57)	3.301 (1.15)	2.907 (0.35)	2.895 (0.25)	2.897 (0.26)	2.900 (0.28)
εDP-DR	T/T	0.00	2.999 (0.18)	2.998 (0.18)	2.997 (0.19)	2.996 (0.20)	2.992 (0.24)	2.989 (0.28)	2.985 (0.31)
		0.05	3.745 (0.60)	3.028 (0.29)	3.002 (0.27)	2.997 (0.26)	2.991 (0.29)	2.985 (0.31)	2.980 (0.34)
		0.10	4.489 (0.78)	3.138 (0.54)	3.017 (0.35)	2.999 (0.30)	2.991 (0.30)	2.985 (0.32)	2.980 (0.33)
		0.20	5.978 (1.03)	4.464 (1.72)	3.531 (1.40)	3.058 (0.67)	3.007 (0.51)	2.998 (0.50)	2.993 (0.51)
	T/F	0.00	3.004 (0.24)	2.998 (0.24)	2.994 (0.24)	2.986 (0.25)	2.979 (0.28)	2.974 (0.32)	2.968 (0.36)
		0.05	3.750 (0.60)	3.033 (0.31)	3.001 (0.29)	2.989 (0.32)	2.978 (0.33)	2.970 (0.36)	2.963 (0.39)
		0.10	4.493 (0.78)	3.149 (0.54)	3.020 (0.36)	2.992 (0.34)	2.978 (0.34)	2.970 (0.37)	2.963 (0.39)
		0.20	5.983 (1.02)	4.489 (1.71)	3.543 (1.40)	3.057 (0.69)	2.998 (0.55)	2.984 (0.54)	2.976 (0.54)
	F/T	0.00	2.999 (0.18)	2.999 (0.18)	2.999 (0.18)	3.001 (0.20)	3.005 (0.23)	3.010 (0.26)	3.014 (0.29)
		0.05	3.746 (0.50)	3.020 (0.19)	2.998 (0.19)	2.998 (0.20)	3.001 (0.23)	3.005 (0.26)	3.009 (0.29)
		0.10	4.493 (0.66)	3.108 (0.37)	3.004 (0.20)	2.998 (0.21)	3.001 (0.23)	3.004 (0.26)	3.007 (0.28)
		0.20	5.986 (0.87)	4.541 (1.58)	3.486 (1.24)	3.020 (0.38)	3.003 (0.24)	3.005 (0.25)	3.008 (0.27)

Table 5: Results of γ-sensitivity study. Each figure displays the mean (sd) of 10,000 simulations for each setting. In the second column, "T" and "F" denote the correct and the incorrect modeling, respectively.

7. Real Data Analysis

In this section, we demonstrate the estimation of the ATE on a real dataset. We use the data of the National Health and Nutrition Examination Survey Data I Epidemiologic Follow-up Study (NHEFS). The NHEFS is a national longitudinal study that was performed by U.S. public agencies. We use the processed dataset available online (Hernán and Robins, 2020, https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/). The NHEFS dataset contains 1,566 observations of smokers who were enrolled in the study in 1971–75. By the follow-up visit in 1982, 403 (25.7%) participants had quit smoking. The study goal was to evaluate the treatment effect of smoking cessation ($T = 1$) on weight gain (Y). Other than the treatment and outcome, several baseline variables were collected, including sex, age, race, education level, intensity and duration of smoking, physical activity in daily life, recreational exercise, and baseline weight. We used all of them to control for confounding in a similar manner to that of Hernán and Robins (2020). We included linear and quadratic terms for all continuous covariates (age, intensity and duration of smoking, and baseline weight) and dummy terms for the discrete covariates. The propensity score was estimated by logistic regression, and outcome regression was performed by unnormalized Gaussian modeling (the tuning parameter was set to 0.2).

The original dataset does not contain obvious outliers; then, we randomly replaced 10% observations with outliers drawn from $\mathcal{N}(100, 5^2)$. Then, we estimated $\mu^{(1)}$, $\mu^{(0)}$ and the ATE by the same methods in the Monte Carlo simulations. This process was repeated 10,000 times, and we summarized the results in Table 6. For reference, we estimated every target quantity using the naive IPW/DR using the original data.

For the IPW-type estimators, the median-based methods gave larger estimates of $\mu^{(1)}$ and $\mu^{(0)}$ than those in the case of IPW (no outliers). In particular, $\mu^{(0)}$ was estimated to be much larger. As a result, when using the median-based methods, the ATE was estimated to be smaller than that in the case of IPW (no outliers). By contrast, DP-IPW overestimated $\mu^{(1)}$ with $\gamma = 0.05$ and underestimated $\mu^{(1)}$ with $\gamma \geq 0.10$. It overestimated $\mu^{(0)}$ compared to the case of IPW (no outliers), and this tendency was strengthened by increasing γ. However, because the extent of overestimation of $\mu^{(0)}$ was smaller than that in the case of median-based methods, the estimate of the ATE by DP-IPW was closer to that obtained using IPW (no outliers) than by using the median-based methods. The DR-type estimators showed similar results. The median-based methods overestimated $\mu^{(1)}$ and $\mu^{(0)}$. DP-DR and εDP-DR underestimated $\mu^{(1)}$ and overestimated $\mu^{(0)}$. The ATE was estimated better by DP-DR and εDP-DR than by the median-

based methods. DP-DR and εDP-DR had the same tendency of estimation bias and γ; a larger γ value increased the bias.

	Target Quantities		
	$\mu^{(1)}$	$\mu^{(0)}$	ATE
IPW (no outliers)	5.221 (-)	1.780 (-)	3.441 (-)
IPW	14.718 (1.57)	11.607 (0.87)	3.111 (1.78)
median (Firpo)	5.439 (0.21)	2.753 (0.10)	2.686 (0.24)
median (Zhang-IPW)	5.439 (0.21)	2.753 (0.10)	2.686 (0.24)
DP-IPW ($\gamma = 0.05$)	5.597 (0.30)	1.851 (0.07)	3.746 (0.31)
DP-IPW ($\gamma = 0.10$)	5.157 (0.15)	1.819 (0.07)	3.338 (0.17)
DP-IPW ($\gamma = 0.20$)	5.089 (0.15)	1.875 (0.06)	3.215 (0.16)
DP-IPW ($\gamma = 0.50$)	4.949 (0.15)	2.007 (0.06)	2.941 (0.16)
DR (no outliers)	5.136 (-)	1.772 (-)	3.364 (-)
DR	14.574 (1.57)	11.589 (0.90)	2.985 (1.81)
median (Zhang-DR)	5.352 (0.20)	2.743 (0.10)	2.609 (0.22)
median (Sued)	5.353 (0.20)	2.744 (0.10)	2.609 (0.23)
median (TMLE)	5.363 (0.21)	2.739 (0.10)	2.624 (0.23)
DP-DR ($\gamma = 0.05$)	5.478 (0.27)	1.842 (0.07)	3.636 (0.28)
DP-DR ($\gamma = 0.10$)	5.057 (0.16)	1.810 (0.07)	3.248 (0.17)
DP-DR ($\gamma = 0.20$)	4.983 (0.16)	1.865 (0.06)	3.119 (0.17)
DP-DR ($\gamma = 0.50$)	4.834 (0.16)	1.997 (0.06)	2.837 (0.17)
εDP-DR ($\gamma = 0.05$)	5.574 (0.29)	1.851 (0.07)	3.723 (0.30)
εDP-DR ($\gamma = 0.10$)	5.148 (0.15)	1.819 (0.07)	3.330 (0.17)
εDP-DR ($\gamma = 0.20$)	5.080 (0.15)	1.874 (0.06)	3.206 (0.17)
εDP-DR ($\gamma = 0.50$)	4.937 (0.15)	2.007 (0.06)	2.930 (0.16)

Table 6: Results of the NHEFS data analysis. Each figure shows mean (sd) of 10,000 estimates.

Supplementary Materials

Appendices and Tables are available online.

Acknowledgements

This work was partially supported by JSPS KAKENHI Grant Number 17K00065.

References

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

Basu, A., I. R. Harris, N. L. Hjort, and M. Jones (1998). Robust and efficient estimation by minimising a density power divergence. *Biometrika 85*(3), 549–559.

Canavire-Bacarreza, G., L. Castro Peñarrieta, and D. Ugarte Ontiveros (2021, April). Outliers in Semi-Parametric estimation of treatment effects. *Econometrics 9*(2), 19.

Díaz, I. (2017). Efficient estimation of quantiles in missing data models. *Journal of Statistical Planning and Inference 190*, 39–51.

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

Fujisawa, H. (2013). Normalized estimating equation for robust parameter estimation. *Electronic Journal of Statistics 7*, 1587–1606.

Fujisawa, H. and S. Eguchi (2008). Robust parameter estimation with a small bias against heavy contamination. *Journal of Multivariate Analysis 99*(9), 2053–2081.

Hampel, F. R., E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel (2011). *Robust statistics: the approach based on influence functions*, Volume 196. John Wiley & Sons.

Hernán, M. A. and J. M. Robins (2020). *Causal inference: what if*. Boca Raton: Chapman & Hall/CRC.

Hoshino, T. (2007). Doubly robust-type estimation for covariate adjustment in latent variable modeling. *Psychometrika 72*(4), 535–549.

Huber, P. J. (2004). *Robust statistics*, Volume 523. John Wiley & Sons.

Imbens, G. W. and D. B. Rubin (2015, April). *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press.

Jones, M., N. L. Hjort, I. R. Harris, and A. Basu (2001). A comparison of related density-based minimum divergence estimators. *Biometrika 88*(3), 865–873.

Kanamori, T. and H. Fujisawa (2015). Robust estimation under heavy contamination using unnormalized models. *Biometrika 102*(3), 559–572.

Kawashima, T. and H. Fujisawa (2017). Robust and sparse regression via γ-divergence. *Entropy 19*(11), 608.

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.

Maechler, M., P. Rousseeuw, C. Croux, V. Todorov, A. Ruckstuhl, M. Salibian-Barrera, T. Ver-

beke, M. Koller, E. L. Conceicao, and M. A. di Palma (2021). *robustbase: Basic Robust Statistics*. R package version 0.93.9.

Maronna, R. A., R. D. Martin, V. J. Yohai, and M. Salibián-Barrera (2019). *Robust statistics: theory and methods (with R)*. John Wiley & Sons.

Robins, J. M. and A. Rotnitzky (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association 90*(429), 122–129.

Robins, J. M., A. Rotnitzky, and L. P. Zhao (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association 89*(427), 846–866.

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.

Rousseeuw, P. J. and B. C. van Zomeren (1990, September). Unmasking multivariate outliers and leverage points. *J. Am. Stat. Assoc. 85*(411), 633–639.

Scharfstein, D. O., A. Rotnitzky, and J. M. Robins (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association 94*(448), 1096–1120.

Sued, M., M. Valdora, and V. Yohai (2020). Robust doubly protected estimators for quantiles with missing data. *TEST 63*(3), 819–843.

Tsiatis, A. (2006, June). *Semiparametric Theory and Missing Data*. Springer New York.

Van Der Laan, M. J. and D. Rubin (2006). Targeted maximum likelihood learning. *The international journal of biostatistics 2*(1).

Van der Vaart, A. W. (2000). *Asymptotic statistics*, Volume 3. Cambridge university press.

Windham, M. P. (1995). Robustifying model fitting. *Journal of the Royal Statistical Society. Series B (Methodological)*, 599–609.

Yohai, V. J. (1987). High breakdown-point and high efficiency robust estimates for regression. *The Annals of Statistics*, 642–656.

Zhang, Z., Z. Chen, J. F. Troendle, and J. Zhang (2012). Causal inference on quantiles with an obstetric application. *Biometrics 68*(3), 697–706.

Kazuharu Harada

Department of Statistical Science, School of Multidiciplinary Sciences, The Graduate University for Advanced Studies (SOKENDAI), 10-3 Midori-cho, Tachikawa, Tokyo 190-8562, Japan.

E-mail: kharada@ism.ac.jp

Hironori Fujisawa

Department of Statistical Inference and Mathematics, The Institute of Statistical Mathematics, 10-3 Midori-cho, Tachikawa, Tokyo 190-8562, Japan.

E-mail: fujisawa@ism.ac.jp