Journal of the Royal Statistical Society Series B: Statistical Methodology, 2023, **00**, 1–24
https://doi.org/10.1093/jrsssb/qkad072

Journal of the Statistics Society Series B Statistical Methodology

Original Article

Semi-parametric estimation of treatment effects in randomised experiments

Susan Athey[1,2], Peter J. Bickel[3], Aiyou Chen[5,4], Guido W. Imbens[1,2,6] and Michael Pollmann[7]

[1]Graduate School of Business, Stanford University, CA, USA
[2]National Bureau of Economic Research, MA, USA
[3]Department of Statistics, University of California, Berkeley, CA, USA
[4]Google LLC, Mountain View, CA, USA
[5]Waymo, Mountain View, CA, USA
[6]Department of Economics, Stanford University, CA, USA
[7]Department of Economics, Duke University, NC, USA

Address for correspondence: Guido W. Imbens, Graduate School of Business, Stanford University, CA, USA. Email: imbens@stanford.edu

Abstract

We develop new semi-parametric methods for estimating treatment effects. We focus on settings where the outcome distributions may be thick tailed, where treatment effects may be small, where sample sizes are large, and where assignment is completely random. This setting is of particular interest in recent online experimentation. We propose using parametric models for the treatment effects, leading to semi-parametric models for the outcome distributions. We derive the semi-parametric efficiency bound for the treatment effects for this setting, and propose efficient estimators. In the leading case with constant quantile treatment effects, one of the proposed efficient estimators has an interesting interpretation as a weighted average of quantile treatment effects, with the weights proportional to minus the second derivative of the log of the density of the potential outcomes. Our analysis also suggests an extension of Huber's model and trimmed mean to include asymmetry.

Keywords: average treatment effects, potential outcomes, quantile treatment effects, semi-parametric efficiency bound

1 Introduction

Historically, randomised experiments were often carried out in medical and agricultural settings. In these settings, sample sizes were often modest, typically on the order of hundreds or (more rarely) thousands of units. Outcomes commonly studied included mortality or crop yield, and were characterised by relatively well-behaved distributions with thin tails. Standard analyses in those settings typically involved estimating the average effect of the treatment using the difference in average outcomes by treatment group, followed by constructing confidence intervals using Normal distribution-based approximations. These methods originated in the 1920s, e.g. Neyman (1923/1990) and Fisher (1937), but they continue to be the standard in modern applications. See Wu and Hamada (2011) for a recent discussion.

More recently many experiments are conducted online (see Kohavi et al., 2020 for an overview), leading to substantially different settings. Gupta et al. (2019, p. 20) claim that 'Together these organizations [Airbnb, Amazon, Booking.com, Facebook, Google, LinkedIn, Lyft, Microsoft, Netflix, Twitter, Uber, and Yandex] tested more than one hundred thousand experimental treatments last year'. The settings for these online experiments are substantially

Received: September 6, 2021. **Revised:** April 10, 2023. **Accepted:** April 28, 2023

different from those in biomedical and agricultural settings. First, the experiments are often on a vastly different scale, with the number of units on the order of millions to tens of millions. Second, the outcomes of interest, variables such as time spent by a consumer, sales per consumer or payments per service provider, are characterised by distributions with extremely thick tails. Third, the treatment effects are often extremely small relative to the standard deviation of the outcomes, even if their magnitude remains substantively important. For example, Lewis and Rao (2015) analyse challenges with statistical power in experiments designed to measure the effect of digital advertising on consumer expenditures. They discuss a hypothetical experiment where the average expenditure per potential customer is \$7 with a standard deviation of \$75, and where an average treatment effect of \$0.35 (0.005 of a standard deviation) would be substantial in the sense of being highly profitable for the company given the cost of advertising. In the Lewis and Rao example, an experiment with power 0.8 for a treatment effect of \$0.35, and a significance level for the two-sided test of means of 0.05, would require a sample size of 1.4 million customers. As a result, confidence intervals for the average treatment effect are likely to include zero even if the true effects were substantively important and samples are large. Even if a confidence interval for the average treatment effect includes zero, there may be evidence about the presence of causal effects of the treatment. Using Fisher exact p-value calculations (Fisher, 1937) with well-chosen statistics (e.g. the Hodges–Lehman difference in average ranks, Rosenbaum, 1993), one may well be able to establish conclusively that treatment effects are present. However, the magnitude of the treatment effect, rather than its presence, is typically important for decision makers.

This sets the stage for the problem we address in this paper. In the absence of additional information, there exists no estimator for the average treatment effect that is more efficient than the difference in means. To obtain more precise estimates, we either need to change the focus away from the average treatment effect, or we need to make additional assumptions. One approach to changing the question, at least slightly, is to transform the outcome (e.g. taking logarithms or winsorising) followed by a standard analysis estimating the average effect of the treatment on the transformed outcome. In this paper, like Taddy et al. (2016) and Tripuraneni et al. (2021), we choose a different approach, namely making additional assumptions on the joint distribution of the outcomes and treatment indicator.

The key conceptual contribution is that we postulate a semi-parametric model for the outcome distributions by treatment group. The leading example of this semi-parametric model corresponds to restricting the quantile treatment effects to be identical across quantiles, thus assuming that the two conditional outcome distributions differ only by a shift. We do not directly use parametric models for the outcome distributions by treatment group, because specifying such a model that well approximates the full outcome distribution is more challenging than postulating a model for the treatment effects. Unlike outcomes, treatment effects tend to be small and often have little variation. For this semi-parametric set-up (e.g. Bickel & Doksum, 2015; Bickel et al., 1993), we derive the influence function, the semi-parametric efficiency bound, and we propose semi-parametrically efficient estimators.

It turns out that the parametrisation of the treatment effect can be very informative, potentially making the asymptotic variance for the corresponding semi-parametric estimators substantially smaller than the asymptotic variance for the difference in means estimator. For example, if the potential outcomes have Cauchy distributions, the variance bound for the average treatment effect is infinite because the moments of the Cauchy distribution do not exist. However, under the constant additive treatment effect assumption (implying that the quantile treatment effects are identical), the semi-parametric variance bound for the treatment effect is finite.

In addition, even if this model for the treatment effect is misspecified, the estimand corresponding to proposed estimators continue to have a causal interpretation, as a weighted average of quantile treatment effects, making it an easy-to-implement and attractive choice in practice.

The remainder of the paper is organised as follows. First, in Section 2, we consider the leading case where we assume the two potential outcome distributions differ only by a shift, so that the quantile treatment effects are all identical. This is implied by, but does not require, the assumption that the treatment effect is additive and constant. In Section 3, we consider the case where we have more flexible parametric models linking the two conditional outcome distributions. In Section 4, we provide some simulation evidence regarding the finite sample properties of the proposed methods in controlled settings and provide real data illustrations. Section 5 concludes. A software implementation for R is available at https://github.com/michaelpollmann/parTreat.

2 Constant quantile treatment effects

In this section, we focus on a special case with constant quantile treatment effects. After setting up the problem formally, we discuss robust estimation in the one-sample case to motivate a class of weighted quantile treatment effect estimators. We then discuss the formal semi-parametric problem and show adaptivity of the proposed estimators. Finally, we consider partial adaptivity and robustness.

This case is closely related to the classical two-sample problem, as discussed in Hodges and Lehmann (1963), and to problems considered in the literature on robust descriptive statistics as in Bickel and Lehmann (1975a, 1975b), K. Doksum (1974), K. A. Doksum and Sievers (1976), and in particular Jaeckel (1971a, 1971b). Section 2 can be interpreted as an extension of Jaeckel's work, in a causal inference framework, to the two-sample context in the setting of semi-parametric theory. In the process of doing so, we generalise Huber's model (Huber, 1964) and the estimator based on trimmed means to include asymmetry, and present a simplified version of the results of Chernoff et al. (1967) (see also Bickel (1967), Govindarajulu et al. (1967) and Stigler (1974) on linear combinations of order statistics). In particular, we exhibit efficient M (maximum-likelihood type) and L (linear combination of order statistics) estimates for outcome distributions that are known up to a shift. We then analyse fully adaptive estimates of both types, as discussed in Bickel et al. (1993), and partially adaptive estimates, in particular flexible trimmed means (Jaeckel, 1971a). In this setting, the problem is closely related to the literature on robust estimation of locations (e.g. Bickel & Lehmann, 1975a, 1975b, 1976, 2012; Hampel et al., 2011; Huber, 2011).

2.1 Set-up

We consider a set-up with a randomised experiment with n observations drawn randomly from a large population. With probability $p \in (0, 1)$ a unit is assigned to the treatment group. Let n_1 and $n_0 = n - n_1$ denote the number of units assigned to the treatment and control group. Following Neyman (1923/1990), Rubin (1974), and Imbens and Rubin (2015), let $Y_i(0)$ and $Y_i(1)$ denote the two potential outcomes for unit i, and let the treatment be denoted by $Z_i \in \{0, 1\}$. We assume that the treatment assignment for one unit does not affect the outcomes for any other unit. For all units in the sample we observe the pair (Z_i, Y_i), where $Y_i \equiv Y_i(Z_i)$. The cumulative distribution functions for the two potential outcomes are $F_0(y)$ and $F_1(y)$ with inverses $F_0^{-1}(u)$ and $F_1^{-1}(u)$, and means and variances μ_0, μ_1, σ_0^2, and σ_1^2. Note that by the random assignment assumption the distribution of the potential outcome $Y_i(z)$ is identical to the conditional distribution of the realised outcome Y_i conditional on $Z_i = z$: $F_z(y) \equiv \Pr(Y_i(z) \leq y) = \Pr(Y_i \leq y \mid Z_i = z)$.

We are interested in the average treatment effect in the population,

$$\tau \equiv \mathbb{E}[Y_i(1) - Y_i(0)]. \tag{2.1}$$

The natural estimator for this average treatment effect is the difference in sample averages

$$\hat{\tau} = \overline{Y}_1 - \overline{Y}_0, \text{ where } \overline{Y}_1 = \frac{1}{n_1}\sum_{i=1}^{n} Z_i Y_i, \quad \overline{Y}_0 = \frac{1}{n_0}\sum_{i=1}^{n}(1 - Z_i)Y_i, \tag{2.2}$$

are the averages of the observed outcomes by treatment group. Under standard conditions $\frac{n_1}{n} \xrightarrow{P} p$, $\frac{n_0}{n} \xrightarrow{P} (1 - p)$, and

$$\sqrt{n}(\hat{\tau} - \tau) \xrightarrow{d} \mathcal{N}\left(0, \frac{\sigma_0^2}{1-p} + \frac{\sigma_1^2}{p}\right). \tag{2.3}$$

The concern is that this conventional estimator $\hat{\tau}$ may be imprecise. In particular in settings where the outcome distribution is thick tailed, sometimes extremely so, confidence intervals may be wide. We address this issue in this paper by imposing some restrictions on the two potential outcome distributions. Following the semi-parametric literature (Bickel & Lehmann, 1975a, 1975b, 1976, 2012), we exploit these restrictions to develop new estimators.

2.2 Weighted average quantile treatment effects

It is useful to start with quantile treatment effects (Lehmann & D'Abrera, 1975), which play an important role in our set-up. For quantile $u \in (0, 1)$, define

$$\Delta(u) \equiv F_1^{-1}(u) - F_0^{-1}(u), \quad 0 \le u \le 1. \tag{2.4}$$

These quantile treatment effects are closely related to what K. Doksum (1974) and K. A. Doksum and Sievers (1976) label the *response* function: $R(y) \equiv F_1^{-1}(F_0(y)) - y = \Delta(F_0(y))$. The natural estimate for the quantile treatment effect is the empirical plug-in, $\hat{\Delta}(u) \equiv \hat{F}_1^{-1}(u) - \hat{F}_0^{-1}(u)$, where $\hat{F}_1^{-1}(u)$ equals $Y^{(1)}_{([n_1 u])}$, defined as the $[n_1 u]$th order statistic of $Y_i \mid Z_i = 1, i = 1, \ldots, n_1$, where $n_1 = \sum_{i=1}^n Z_i$ and similarly for $\hat{F}_0^{-1}(u)$.

A natural class of parameters summarising the difference between the $Y_i(1)$ and $Y_i(0)$ distributions consists of weighted averages of the quantile treatment effects:

$$\tau(F_0, F_1;\, W) \equiv \int_0^1 \Delta(u)\, \mathrm{d}W(u),$$

where the weights integrate to one, $W(0) = 0$, $W(1) = 1$. Different choices for the weight function correspond to different estimands. The constant weight case, $W'(u) \equiv 1$, corresponds to the population average treatment effect $\tau = \mathbb{E}[Y_i(1) - Y_i(0)]$. The median corresponds to the case where $W(\cdot)$ puts all its mass at 1/2. We thus allow $W(\cdot)$ to permit point masses.

For a given weight function $W(\cdot)$, we can estimate the parameter $\tau(F_0, F_1;\, W)$ using a *weighted average quantile* (waq) estimator:

$$\begin{aligned} \hat{\tau}_W \equiv\ \tau(\hat{F}_0, \hat{F}_1;\, W) &= \int_0^1 \left(\hat{F}_1^{-1}(u) - \hat{F}_0^{-1}(u)\right) \mathrm{d}W(u) \\ &= \frac{1}{n_1}\sum_{i=1}^{n_1} w_i^{(1)} Y^{(1)}_{(i)} - \frac{1}{n_0}\sum_{i=1}^{n_0} w_i^{(0)} Y^{(0)}_{(i)}, \end{aligned} \tag{2.5}$$

where

$$w_i^{(z)} \equiv\ W\left(\frac{i}{n_z}\right) - W\left(\frac{i-1}{n_z}\right)$$

and $Y^{(z)}_{(i)}$ again are the order statistics in treatment group z.

2.3 Efficient estimation of waq treatment effects using influence functions

To understand the properties of the waq estimator $\hat{\tau}_W$, we begin by considering the non-parametric model for a single sample, with cumulative distribution $F(\cdot)$ where the interest is in the weighted quantile $\int_{-\infty}^{\infty} F^{-1}(u)\, \mathrm{d}W(u)$ for a given weight funtion $W(\cdot)$. For this one-sample case, Jaeckel (1971a, 1971b), building on Chernoff et al. (1967), shows that under simple conditions on W and F, for a sample size of n,

$$\int_0^1 \left(\hat{F}^{-1}(u) - F^{-1}(u)\right) \mathrm{d}W(u) = \int_{-\infty}^{\infty} \psi(x, F, W) d(\hat{F}(x) - F(x)) + o_P(n^{-1/2}), \tag{2.6}$$

where the *influence function* ψ is related to the weight function W by

$$\psi(x, F, W) = -\int_x^{\infty} \frac{1}{f(y)}\, \mathrm{d}W(F(y)) + \int_{-\infty}^{\infty} \frac{F(y)}{f(y)}\, \mathrm{d}W(F(y)). \tag{2.7}$$

The last term ensures that $\int_{-\infty}^{\infty} \psi(x, F, W)\, \mathrm{d}F(x) = 0$.

Note that if the derivatives $\psi\prime(\cdot)$ and $W'(\cdot) = w(\cdot)$ exist, by equation (2.7),

$$\psi(x, F, W) = -\int_x^\infty w(F(y))\,\mathrm{d}y + \int_{-\infty}^\infty F(y)w(F(y))\,\mathrm{d}y$$

so that $\psi\prime(x, F, W) = w(F(x))$. Note that for the median, equation (2.7) yields, $\psi(x, F, W) = \frac{sign(x-F^{-1}(\frac{1}{2}))}{2f(F^{-1}(\frac{1}{2}))}$. Our formula (2.7) is slightly more general than Jaeckel's, and in the online supplementary appendix, we establish sufficient conditions on the cumulative distribution function $F(\cdot)$ and the weight function $W(\cdot)$ for our version of his result to hold.

Expression (2.6) in turn implies that

$$\int_0^1 (\hat{F}^{-1}(u) - F^{-1}(u))\,\mathrm{d}W(u) \xrightarrow{d} \mathcal{N}(0, \sigma^2(F, W)),$$

where the variance equals the expectation of the square of the influence function:

$$\sigma^2(F, \omega) = \int_{-\infty}^\infty \psi(x, F, W)^2\,\mathrm{d}F(x).$$

The results in Jaeckel (1971a, 1971b) for the one-sample case extend in the following way to the two-sample setting that is our primary focus. If $\tau(F_0, F_1, W)$ is estimated by $\hat{\tau}_W$ in equation (2.5), then, under regularity conditions given in the online supplementary appendix, Theorem A.1,

$$\hat{\tau}_W = \tau(F_0, F_1, W) + \frac{1}{n}\sum_{i=1}^n \left(Z_i \frac{\psi(Y_i, F_1, W)}{p} - (1 - Z_i)\frac{\psi(Y_i, F_0, W)}{1-p} \right) + o_P(n^{-1/2}), \tag{2.8}$$

where $\psi(x, F, W)$ is given by equation (2.7).

2.4 Constant quantile treatment effects

Now let us return to the primary focus of this section, the estimation of the average treatment effect under the constant quantile treatment effect assumption. Our key assumption in this section is that the quantile treatment effects are all equal:

$$\tau(u) = \tau, \quad \forall u \tag{2.9}$$

and thus, for any weight functions $W(\cdot)$,

$$\tau(F_0, F_1, W) = \tau. \tag{2.10}$$

Later, in Section 3, we generalise this to allow for a more general parametric function linking the quantile treatment effects. One way to motivate the constant quantile treatment effect assumption is to assume that the unit-level treatment effects are all constant, $Y_i(1) - Y_i(0) = \tau$ for all units $i = 1, \ldots, n$. This implies, but is not implied by, the assumption that all the quantile treatment effects are identical. The assumption of constant unit-level treatment effects is very strong, implying rank-invariance, which is in fact stronger than what we need.

In this section, for expository reasons we further assume that we know the control outcome distribution $F_0(\cdot)$ up to a shift. That is, $F_0(x) = F(x - \eta)$, where $F(\cdot)$ (with derivative f) is known and η unknown. Because of the constant quantile treatment effect assumption, the treated potential outcome distribution is also known up to a shift, $F_1(x) = F(x - \eta - \tau)$. Assuming that $F_0(\cdot)$ is known up to a shift is unrealistic in practice, and we remove this assumption below in Section 2.5, but it allows us to focus in this section on some key insights.

For this fully parametric model (with unknown parameters η and τ), if the Fisher information $I(f) = \int (\frac{f'}{f})^2(x)f(x)\,\mathrm{d}x = \int (-\frac{f'}{f})\prime(x)f(x)\,\mathrm{d}x$ satisfies $0 < I(f) < \infty$, the maximum likelihood estimator of τ, suitably regularised (e.g. Le Cam & Yang, 1988), has influence function,

$$\psi_{f,\eta}(Z, Y; \tau) = -\frac{1}{I(f)} \cdot \left(\frac{Z}{p} \cdot \frac{f'}{f}(Y - \eta - \tau) - \frac{1-Z}{1-p} \cdot \frac{f'}{f}(Y - \eta) \right). \tag{2.11}$$

There is an interesting alternative efficient estimator in this known $f(\cdot)$ case. Suppose f'/f is absolutely continuous. Then the weight function

$$w_f(F(x)) \equiv \frac{1}{I(f)}\left(-\frac{f'}{f}\right)'(x) \text{ or } w_f(u) = \frac{1}{I(f)}\left(-\frac{f'}{f}\right)'(F^{-1}(u)) \tag{2.12}$$

provides an efficient L estimate when substituted appropriately in (2.6), leading to

$$\hat{\tau}(F_0, F_1, W) = \int_0^1 (\hat{F}_1^{-1}(u) - \hat{F}_0^{-1}(u))w_f(u)\,\mathrm{d}u. \tag{2.13}$$

It is interesting to inspect the form of the weights $w_f(u)$. These weights are proportional to minus the second derivative of the logarithm of the density function. In other words, we can approximate the efficient estimator by first estimating a large number of quantile treatment effects. Under the model these quantile treatment effects are all identical. To efficiently estimate that common treatment effect we can simply use a weighted average of the estimated quantile treatment effects. It turns out the optimal weights simplify to minus the second derivative of the logarithm of the density. For the Normal distribution, that means the weights are constant. For the Double Exponential distribution the weights put point mass at the median. For the Cauchy distribution the weights are proportional to $-\cos(2\pi u)\sin(\pi u)^2$. Interestingly these weights are negative for some quantiles. One can of course see this by inspecting the estimated weights. If one is concerned by the negative weights one can also modify them by restricting them to be nonnegative. Finally, note that implicitly the influence function estimator also has the negative weights in such cases because the two estimators are first order equivalent.

A final comment connects this to common methods for dealing with thick-tailed distributions. In practice many researchers use winsorising to deal with these problems. This can be interpreted as using a waq estimator with a particular set of weights. Specifically, with winsorising at the q and $1-q$ quantiles, the implicit weights are constant on the interval $(q, 1-q)$, and then put additional point mass q on the qth and $(1-q)$th quantiles. As discussed in Bickel (1965), the asymptotic properties of the winsorising estimator depend delicately on the density at the winsorising quantiles. In our simulations this estimator does not perform particularly well. Like other settings, there is tension here between having an interpretable target that may not be precisely estimable (e.g. the average effect of the treatment), vs. a precisely estimable estimand whose interpretation is more complex (e.g. the waq effect). This tension arises also in other settings. An example is the estimation of average treatment effects under unconfoundedness where weighting by the confounders may affect both the interpretation of the estimand and the precision with which we can estimate it (Crump et al., 2009; Li et al., 2018). Another setting is that discussed in Vansteelandt and Dukes (2022). The use of quantile methods for estimating treatment effects in thick-tailed settings has been studied in Firpo (2007); Firpo et al. (2009), but unlike in those papers, our focus is on the overall treatment effect, rather than the effect at specific quantiles.

2.5 Fully adaptive estimation

As stated earlier, in practice we do not know the density $f(\cdot)$ up to location. However, in this case with constant quantile treatment effects this knowledge does not matter up to first order. Because of the orthogonality of the tangent space with respect to f, it follows from semiparametric theory (Bickel et al., 1993) that even if the density f is unknown, substituting a suitable estimate of f (and η) in (2.11) or (2.13), will yield estimators with influence functions given by (2.11), or equivalently by

$$\psi_{f_0}(Z, Y; \tau) = -\frac{1}{I(f_0)} \cdot \left(\frac{Z}{p} \cdot \frac{f_0'}{f_0}(Y-\tau) - \frac{1-Z}{1-p} \cdot \frac{f_0'}{f_0}(Y)\right), \tag{2.14}$$

where $f_0(\cdot) \equiv F_0'(\cdot) \equiv f(\,\cdot - \eta)$.

For our proposed estimator we split the data randomly into two parts, with the two sub-samples denoted by A, corresponding to $\{(Z_i, Y_i) : 1 \le i \le \frac{n}{2}\}$, and B, corresponding to $\{(Z_i, Y_i) : \frac{n}{2} < i \le n\}$.

The M estimate using the estimated $\hat{f}(\cdot)$ is of the form,

$$\hat{\tau}^{if} \equiv \frac{\tilde{\tau}_{(A)} + \tilde{\tau}_{(B)}}{2} + \frac{1}{n}\left\{\sum_{i=1}^{n/2} \psi_{\hat{f}_{0(B)}}(Z_i, Y_i; \tilde{\tau}_{(B)}) + \sum_{i=1+n/2}^{n} \psi_{\hat{f}_{0(A)}}(Z_i, Y_i; \tilde{\tau}_{(A)})\right\}, \tag{2.15}$$

where $\hat{f}_{0(A)}$ is an estimate of f_0 using $\{(Z_i, Y_i) : 1 \le i \le \frac{n}{2}\}$, and $\hat{f}_{0(B)}$ using $\{(Z_i, Y_i) : \frac{n}{2} < i \le n\}$, a one-step estimate using the sample splitting technique (Klaassen, 1987). $\tilde{\tau}_{(A)}$ and $\tilde{\tau}_{(B)}$ are initial $\sqrt{n}$ consistent estimates based on the two sub-samples, for example based on the difference in medians or other quantiles. Algorithm 1 shows the key steps; additional details are given in the online supplementary material.

Algorithm 1 Influence Function-Based Estimator $\hat{\tau}^{if}$

1: ▷ Input:
2: n_1 treated observations $Y_1^1, \ldots, Y_{n_1}^{(1)}$
3: n_0 control observations $Y_1^0, \ldots, Y_{n_0}^{(0)}$
4:
5: ▷ Randomly split sample into halves A and B:
6: $n_{1(A)} = \lceil n_1/2 \rceil$, $n_{1(B)} = \lfloor n_1/2 \rfloor$, $n_{0(A)} = \lceil n_0/2 \rceil$, $n_{0(B)} = \lfloor n_0/2 \rfloor$
7: denote treated in halves A and B by $Y_1^{(1,A)}, \ldots, Y_{n_{1(A)}}^{(1,A)}$ and $Y_1^{(1,B)}, \ldots, Y_{n_{1(B)}}^{(1,B)}$
8: denote control in halves A and B by $Y_1^{(0,A)}, \ldots, Y_{n_{0(A)}}^{(0,A)}$ and $Y_1^{(0,B)}, \ldots, Y_{n_{0(B)}}^{(0,B)}$
9:
10: ▷ Calculate a preliminary consistent estimator:
11: $\tilde{\tau}_{(B)} = \text{median}(Y_i^{(1,B)}) - \text{median}(Y_i^{(0,B)})$
12:
13: ▷ Estimate density and its derivatives:
14: $\hat{f}_{0(B)}(\cdot), \hat{f}'_{0(B)}(\cdot), \hat{f}''_{0(B)}(\cdot) \leftarrow$ estimated using data $Y_1^{(0,B)}, \ldots, Y_{n_{0(B)}}^{(0,B)}$
15:
16: ▷ Estimate the Fisher information I:
17: $\hat{I}_{(B)} \leftarrow -\frac{1}{n_{0(B)}} \sum_{i=1}^{n_{0(B)}} \frac{\hat{f}_{0(B)}(Y_i^{(0,B)})\hat{f}''_{0(B)}(Y_i^{(0,B)}) - \hat{f}'_{0(B)}(Y_i^{(0,B)})^2}{\hat{f}_{0(B)}(Y_i^{(0,B)})^2}$
18:
19: ▷ Estimate the effects:
20: $\hat{\tau}_{(A)} \leftarrow \tilde{\tau}_{(B)} + \frac{1}{n_{1(A)}} \sum_{i=1}^{n_{1(A)}} \frac{-1}{p\hat{I}_{(B)}} \frac{\hat{f}'_{0(B)}(Y_i^{(1,A)} - \tilde{\tau}_{(B)})}{\hat{f}_{0(B)}(Y_i^{(1,A)} - \tilde{\tau}_{(B)})} - \frac{1}{n_{0(A)}} \sum_{i=1}^{n_{0(A)}} \frac{-1}{(1-p)\hat{I}_{(B)}} \frac{\hat{f}'_{0(B)}(Y_i^{(0,A)})}{\hat{f}_{0(B)}(Y_i^{(0,A)})}$
21:
22: ▷ Repeat lines 10 through 20 reversing A and B, then average:
23: $\hat{\tau}^{if} \leftarrow (\hat{\tau}_{(A)} + \hat{\tau}_{(B)})/2$

We can also construct an L estimate based on an average of the quantile differences. This estimator is obtained by first estimating $F_0(\cdot)$, $f_0(\cdot)$, and $f'_0(\cdot)$, substituting that for $F(\cdot)$, $f(\cdot)$, and $f'(\cdot)$ into $w_f(u)$ in equation (2.12), followed by using this estimated set of weights in equation (2.13), leading to

$$\hat{\tau}^{waq} = \int_0^1 (\hat{F}_1^{-1}(u) - \hat{F}_0^{-1}(u))\hat{w}_f(u)\, \mathrm{d}u. \tag{2.16}$$

Formally, we would use the same sample splitting as above. Details are in Algorithm 2 and the online supplementary material.

Algorithm 2 Weighted Average Quantile Estimator $\hat{\tau}^{waq}$

1: ▷ Input:
2: n_1 treated observations $Y_1^1, \ldots, Y_{n_1}^{(1)}$
3: n_0 control observations $Y_1^0, \ldots, Y_{n_0}^{(0)}$
4:
5: ▷ Randomly split sample into halves A and B:
6: $n_{1(A)} = \lceil n_1/2 \rceil$, $n_{1(B)} = \lfloor n_1/2 \rfloor$, $n_{0(A)} = \lceil n_0/2 \rceil$, $n_{0(B)} = \lfloor n_0/2 \rfloor$
7: denote treated in halves A and B by $Y_1^{(1,A)}, \ldots, Y_{n_{1(A)}}^{(1,A)}$ and $Y_1^{(1,B)}, \ldots, Y_{n_{1(B)}}^{(1,B)}$
8: denote control in halves A and B by $Y_1^{(0,A)}, \ldots, Y_{n_{0(A)}}^{(0,A)}$ and $Y_1^{(0,B)}, \ldots, Y_{n_{0(B)}}^{(0,B)}$
9:
10: ▷ Estimate density and its derivatives:
11: $\hat{f}_{0(B)}(\cdot), \hat{f}'_{0(B)}(\cdot), \hat{f}''_{0(B)}(\cdot) \leftarrow$ estimated using data $Y_1^{(0,B)}, \ldots, Y_{n_{0(B)}}^{(0,B)}$
12:
13: ▷ Order and pair observations:
14: $n_{(A)} = \max(n_{1(A)}, n_{0(A)})$
15: duplicate treated or control observations as needed such that there are $n_{(A)}$ of both, evenly across the distribution, and order them (analogously for the B split):
16: $Y_{(1)}^{(0,A)} \leq Y_{(2)}^{(0,A)} \leq \cdots \leq Y_{(n_{(A)})}^{(0,A)}$; $Y_{(1)}^{(1,A)} \leq Y_{(2)}^{(1,A)} \leq \cdots \leq Y_{(n_{(A)})}^{(1,A)}$
17: $Y_{(1)}^{(0,B)} \leq Y_{(2)}^{(0,B)} \leq \cdots \leq Y_{(n_{(B)})}^{(0,B)}$; $Y_{(1)}^{(1,A)} \leq Y_{(2)}^{(1,B)} \leq \cdots \leq Y_{(n_{(B)})}^{(1,B)}$
18:
19: ▷ Estimate the weighted average quantile effect:
20: weights: $w_{(i)}^{(B)} \leftarrow -\frac{\hat{f}_{0(B)}(Y_{(i)}^{(0,B)})\hat{f}''_{0(B)}(Y_{(i)}^{(0,B)}) - \hat{f}'_{0(B)}(Y_{(i)}^{(0,B)})^2}{\hat{f}_{0(B)}(Y_{(i)}^{(0,B)})^2}$
21: $\hat{\tau}_{(A)} \leftarrow \sum_{i=1}^{n_{(A)}} w_{(i)}^{(B)}(Y_{(i)}^{(1,A)} - Y_{(i)}^{(0,A)}) / \sum_{i=1}^{n_{(A)}} w_{(i)}^{(B)}$
22:
23: ▷ Repeat lines 10 through 21 reversing A and B, then average:
24: $\hat{\tau}^{waq} \leftarrow (\hat{\tau}_{(A)} + \hat{\tau}_{(B)})/2$

Formally, we have for the unknown $f(\cdot)$ case:

Theorem 1 For all f such that f' exists and $0 < I(f) < \infty$:

(i) There exist a $\sqrt{n}$-consistent estimator $\hat{\tau}$.
(ii) Under mild conditions (see equations (A.9) and (A.10) in the online supplementary appendix), we can construct an M estimate $\hat{\tau}^{if}$ such that

$$\sqrt{n}(\hat{\tau}^{if} - \tau) \xrightarrow{d} \mathcal{N}\left(0, \frac{1}{p(1-p)I(f)}\right). \tag{2.17}$$

(iii) Under mild conditions (see Lemma A.2 in the online supplementary appendix), we can construct an L estimate $\hat{\tau}^{waq}$ (weighted average quantile) such that $\hat{\tau}^{waq}$ also satisfies equation (2.17).

The main insight is that the asymptotic variance for the proposed estimator is the same as the variance for the maximum likelihood estimator in the case where F_0 is known up to a shift. Our conditions are not optimal (see Stone, 1975 for minimal ones in the one-sample case). The $\sqrt{n}$-consistent estimator for τ can be based on any quantile treatment effect estimator.

Thus, both the estimators $\hat{\tau}^{waq}$ and $\hat{\tau}^{if}$ are adaptive to models in which the distribution of control and treated potential outcomes is known up to location. Theorem 1 implies that the influence function-based estimator is as efficient as the maximum likelihood estimator based on the true distribution function. For instance, if potential outcomes are normally distributed, the maximum

likelihood estimator is the difference in means, and the influence function estimator has the same limiting distribution. If, however, the potential outcomes follow a Double Exponential distribution, the difference in medians is the efficient estimator. Under this distribution, the influence function-based estimator adapts and has the same limiting distribution as the difference in medians. For the Cauchy distribution the optimal weights are more complicated, $w_f(u) \propto -\cos(2\pi u)\sin(\pi u)^2$, but the influence function-based estimator has the same limiting distribution, without requiring a priori knowledge about the distribution. This influence function-based estimator $\hat{\tau}^{if}$ is a special case of the estimator developed by Cuzick (1992a, 1992b) for the partial linear regression model setting.

Although these estimators are efficient under the constant additive treatment model, as we shall see in Section 3, if the the constant quantile treatment effect assumption is violated, estimates of the types derived from f known continue to estimate at rate $n^{-1/2}$, meaningful measures of the treatment effect as discussed in Section 2.1. This is unfortunately not the case for $\hat{\tau}^{if}$ and $\hat{\tau}^{waq}$ because estimation of f' and f'' introduces components of variance of order larger than $n^{-1/2}$. However, there is a partial remedy, that we discuss next.

2.6 Partial adaptation

An interesting alternative to fully adaptative estimation in the closely related one-sample symmetric case was studied by Jaeckel (1971b). See also Yu and Yao (2017). The estimator proposed by Jaeckel (1971b) is

$$\hat{\mu}_\alpha \equiv \frac{1}{1-2\alpha}\int_\alpha^{1-\alpha} \hat{F}^{-1}(u)\,\mathrm{d}u$$

for a sample from $F(x-\mu)$ with f symmetric. We can interpret this as restricting the class of weight functions to one indexed by a scalar parameter α:

$$w_\alpha(u) = \begin{cases} \dfrac{1}{1-2\alpha} & \text{if } \alpha \le u \le 1-\alpha \\ 0 & \text{otherwise.} \end{cases}$$

This weight function yields the α-trimmed mean. We can then choose the value of α that minimises the asymptotic variance of $\hat{\mu}_\alpha$. This asymptotic variance is equal to

$$\begin{aligned}\sigma_\alpha^2 \equiv \frac{1}{n(1-2\alpha)^2}&(\mathbb{E}(X-\overline{\mu})^2 1(F^{-1}(\alpha) \le X \le F^{-1}(1-\alpha)) \\ &+ \alpha\cdot(F^{-1}(1-\alpha)-\overline{\mu})^2 + \alpha(F^{-1}(\alpha)-\overline{\mu})^2),\end{aligned}$$

with

$$\overline{\mu} = \int_\alpha^{1-\alpha} F^{-1}(u)\,\mathrm{d}u + \alpha(F^{-1}(\alpha)+F^{-1}(1-\alpha)),$$

which can be estimated by replacing F with the empirical distribution, denoted as $\hat{\sigma}_\alpha^2$. Let $\hat{\alpha} = \arg\min_\alpha \hat{\sigma}_\alpha^2$, and let $\hat{\mu}_{\hat{\alpha}}$ be the corresponding estimator for μ. Jaeckel (1971b) shows that $\hat{\mu}_{\hat{\alpha}}$ is adaptive for estimating μ over a Huber family of densities. In the Huber family $(X-\mu)/\sigma$ has density f for varying μ, $\sigma > 0$:

$$\log f(x) = \begin{cases} -\dfrac{x^2}{2} - c(k), & \text{if } |x| \le k \\ -\dfrac{k|x|}{2} - c(k), & \text{if } |x| > k, \end{cases}$$

where $c(k)$ makes $\int f(x)\,\mathrm{d}x = 1$ and $k = -F^{-1}(\alpha)$. Adaptivity here means that using the trimming proportion optimising the variance estimate, in fact yields an estimate which is efficient for the member of the Huber family generating the data. He optimises $0 < \alpha_0 \leq \alpha \leq \alpha_1 < \frac{1}{2}$.

Because this family includes among others the Gaussian ($k \to \infty$) and double exponential ($k = 0$), this family is very flexible. For more properties, see Huber (2011).

In the two-sample case, it is reasonable to consider asymmetric weight functions leading to the natural generalisation,

$$\hat{\tau}_{\alpha,\beta} = \frac{1}{1-(\alpha+\beta)} \int_{\alpha}^{1-\beta} (\hat{F}_1^{-1}(u) - \hat{F}_0^{-1}(u))\,\mathrm{d}u. \tag{2.18}$$

This estimator is partially adaptive, in a similar way to the symmetric trimmed mean in the one sample problem. In the online supplementary appendix, we extend Jaeckel's result on partial adaptation for our two-sample problem to a generalisation of the Huber family whose members are symmetric iff $k_1 = k_2$, defined by $(X-\mu)/\sigma \sim f$:

$$\log f(x) = \begin{cases} -\dfrac{x^2}{2} - c(k_1, k_2), & \text{if } -k_1 \leq x \leq k_2, \\ \dfrac{k_1 x}{2} - c(k_1, k_2), & \text{if } x < -k_1, \\ -\dfrac{k_2 x}{2} - c(k_1, k_2), & \text{if } x > k_2, \end{cases} \tag{2.19}$$

where $c(k_1, k_2) \equiv \log\left(2\left(\frac{\mathrm{e}^{-k_1/2}}{k_1} + \frac{\mathrm{e}^{-k_2/2}}{k_2}\right) + \sqrt{2\pi}(\Phi(k_2) - \Phi(-k_1))\right)$ and Φ is the cumulative distribution function of $\mathcal{N}(0, 1)$. See Figure 1 for illustration. This family can be equivalently parametrised by $F(-k_1)$ and $F(k_2)$. $f(\cdot)$ is symmetric if $k_1 = k_2$.

In the online supplementary material, we also discuss how inference can proceed in this setting.

3 The general parametric treatment effect case

In some settings, the assumption of an additive model may be too restrictive. In this section, we develop estimators given a general parametric model for this difference.

The starting point is a model governing the relation between the two potential outcomes:

Assumption 1 (Parametric model quantile treatment effects). The potential outcome distributions satisfy

$$F_1(h(y, \theta)) = F_0(y).$$

The constant quantile treatment effect case is a special case of this with $h(y, \theta) = y + \theta$. Another important special case is the proportional treatment effect case, $h(y, \theta) = \theta y$. For the general case, the waq estimator does not directly generalise, so we focus on the influence function-based estimator. For the general case the influence function is more complex.

This approach of modelling treatment effects has connections to the literature on structural nested models, which also imposes modelling restrictions on treatment effects, although for different reasons, largely based on the challenges in dynamic settings. See Robins (1986) for an early paper, and Vansteelandt and Joffe (2014) for a review.

As before, we initially assume F_0 known. In terms of the quantile treatment effects $\tau(u)$ Assumption 1 implies the restriction

$$\tau(u) = F_1^{-1}(u) - F_0^{-1}(u) = h(F_0^{-1}(u), \theta) - F_0^{-1}(u).$$

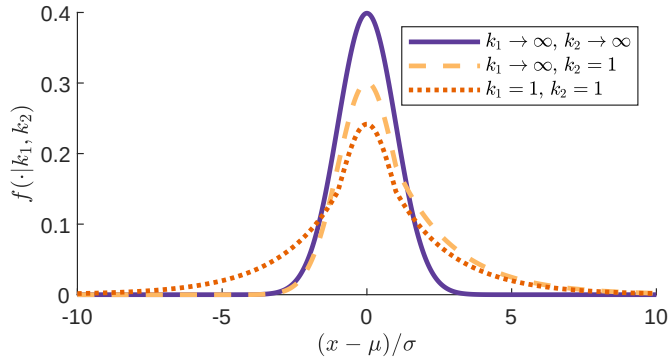

Figure 1. Example members of the generalised Huber family of distributions.

Given Assumption 1, the population average treatment effect can be characterised as

$$\tau^{\text{pop}} = \int_0^1 \left(h(F_0^{-1}(u), \theta) - F_0^{-1}(u)\right) \mathrm{d}u.$$

In practice, however, estimating τ^{pop} may still be subject to substantial sampling variance, even if $h(\cdot)$ is known. For example, suppose that $h(y, \theta) = \theta y$, so that the treatment effect is proportional. The average treatment effect is then $\theta E[Y_i \mid Z_i = 0]$. Even if θ is known, estimating the population mean $E[Y_i \mid Z_i = 0]$ could lead to a large standard error. As an alternative, we therefore focus on a different estimand. Specifically, we suggest to estimate the in-sample, as opposed to population, average treatment effect. This is still a well-defined average causal effect that is useful for decision makers. It is in the spirit of the typical analysis of randomised experiments based on convenience samples where the focus is on the average effect for the particular sample. A key insight is that estimators of this object can have a much lower variance. We define the in-sample average treatment effect as

$$\tau^{\text{is}} = \frac{1}{N}\sum_{i=1}^{N}\left(Y_i(1) - Y_i(0)\right) = \frac{1}{N}\sum_{i=1}^{N}\left\{Z_i(Y_i - h^{-1}(Y_i, \theta)) + (1 - Z_i)(h(Y_i, \theta) - Y_i)\right\}. \tag{3.1}$$

When h is not just an additive function, τ^{is} is sample-dependent, and thus stochastic. In particular, when the variance of Y_i is large because of thick tails for the potential outcome distributions, the variance of τ^{is} over repeated samples can be large, too. To give some intuition for this, suppose that θ is known. Then the variance of $\hat{\tau} - \tau^{\text{is}}$ is zero, but the variance of $\tau^{\text{is}} - \tau^{\text{pop}}$ over repeated samples can be large. We therefore focus on the variance of estimators $\hat{\tau}$ relative to τ^{is} for the particular sample at hand, rather than on the variance of $\hat{\tau}$ relative to the population average τ^{pop}.

If F_0 was known, we could estimate θ efficiently by some version of maximum likelihood to get an estimate $\hat{\theta}$ and

$$\hat{\tau} = \int_0^1 \left(h(F_0^{-1}(y), \hat{\theta}) - F_0^{-1}(y)\right) \mathrm{d}y$$

as an estimate of τ. The density of Y given Z is

$$f(y|z) = \left(f_0(y)\right)^{1-z}\left(f_1(y, \theta)\right)^{z}.$$

By Assumption 1

$$f_1(y, \theta) = f_0(h^{-1}(y, \theta))\frac{\partial h^{-1}(y, \theta)}{\partial y} \tag{3.2}$$

and the score function is

$$\dot{\ell}(y, z, \theta) \equiv z \cdot \frac{\partial}{\partial \theta} \log f_1(y, \theta)$$

yielding,

$$\hat{\theta} = \theta + \frac{1}{n} \sum_{i=1}^{n} I^{-1} \dot{\ell}(Y_i, Z_i, \theta) + o_P(n^{-1/2}),$$

where $I = E(\dot{\ell}(Y, Z, \theta))^2$.

If F_0 is assumed unknown, to obtain an efficient influence function (eif) we must

1. Compute the tangent plane as f_0 varies with θ fixed. The tangent plane is

$$\dot{P}_f = \{u(Y, Z) = (1 - Z)v(Y, \theta) + Zv(h^{-1}(Y, \theta), \theta) : \int v^2(y, \theta) f_0(y)\, \mathrm{d}y < \infty,\ \int v(y, \theta) f_0(y)\, \mathrm{d}y = 0\}.$$

   (Note that both factors of the likelihood must be varied treating θ as fixed.)
2. Project $\dot{\ell}$ on the orthocomplement of the tangent plane to get

$$\dot{\ell}^*(Y, Z, \theta) = \dot{\ell}(Y, Z, \theta) - (ZQ(h^{-1}(Y, \theta), \theta) + (1 - Z)Q(Y, \theta)),$$

   where $ZQ(h^{-1}(Y, \theta), \theta) + (1 - Z)Q(Y, \theta)$ is the projection of $\dot{\ell}$ on $\dot{P}_f$.
3. The eif is given by

$$\psi(Y, Z, \theta) = \frac{\dot{\ell}^*(Y, Z, \theta)}{E\left(\dot{\ell}^*(Y, Z, \theta)\right)^2}.$$

Lemma 1 The eif for θ is

$$\psi_{f_0}(y, z, \theta) = I^{-1} \left\{ \frac{z}{p} \cdot g(y, \theta) - \frac{1 - z}{1 - p} \cdot g(h(y, \theta), \theta) \right\},$$

where

$$g(y, \theta) = \frac{\partial}{\partial \theta} \log f_1(y, \theta) = \frac{\partial}{\partial \theta} \log \left(f_0(h^{-1}(y, \theta)) \cdot \frac{\partial h^{-1}(y, \theta)}{\partial y} \right)$$

and

$$I = \int g^2(h(y, \theta), \theta) f_0(y)\, \mathrm{d}y.$$

To use the influence function approach, we need a $\sqrt{n}$-consistent initial estimator $\tilde{\theta}$. We can do so by a fixed number of quantiles, $u_1, \ldots, u_d$, where d is the dimension of θ, and find the θ that solves

$$\hat{F}_1^{-1}(u) = h(\hat{F}_0^{-1}(u), \theta),$$

for $u = u_1, \ldots, u_d$. For simplicity, we suggest using evenly-spaced quantiles, $u = \frac{1}{1+d}, \frac{2}{1+d}, \ldots, \frac{d}{1+d}$. Let $\tilde{\theta}$ denote the solution to this system of equations. Then:

Theorem 2 Under mild conditions on the estimation of density and its derivative, the estimator $\hat{\theta}^{if}$ below is efficient for θ, i.e. $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, \frac{1}{p(1-p)} I^{-1})$:

$$\hat{\theta}^{if} \equiv \tilde{\theta} + \frac{1}{n}\left\{\sum_{i=1}^{n/2} \psi_{\hat{f}_{0(2)}}(Y_i, Z_i; \tilde{\theta}) + \sum_{i=1+n/2}^{n} \psi_{\hat{f}_{0(1)}}(Y_i, Z_i; \tilde{\theta})\right\},$$

where $\hat{f}_{0(1)}$ is the estimate of f_0 using $\{(Y_i, Z_i) : 1 \leq i \leq \frac{n}{2}\}$ and $\hat{f}_{0(2)}$ is the estimate of f_0 using $\{(Y_i, Z_i) : \frac{n}{2} < i \leq n\}$, again a one-step estimate using the sample splitting technique (Klaassen, 1987), similar to equation (2.15).

Note that, unlike the constant treatment effect setting, the eif is not necessarily the one corresponding to F_0 known.

Given inference for $\hat{\theta}$, inference for $\hat{\tau}$ as an estimator of the in-sample average treatment effect is straightforward based on the Delta method and the representation in equation (3.1), taken as given the potential outcomes. The asymptotic variance for $\hat{\tau}$ is equal to the variance for $\hat{\theta}$, pre- and post-multiplied by the derivative of the expression in equation (3.1) with respect to θ.

4 Simulations

We evaluate the performance of the proposed estimators and conventional estimators in a Monte Carlo study. Throughout most of these simulations, the true unit-level treatment effects are all zero. We estimate the treatment effect using the proposed efficient estimators based on an additive model. We consider seven estimators: The (standard) difference in means, the difference in medians, the Hodges–Lehman (1963) estimator,[1] the adaptively trimmed mean, the adaptively winsorised mean,[2] the estimator based on the eif, and the waq estimator. For the latter two we report only results without sample splitting. Results for the case with sample splitting are very similar and are available in the online supplementary material. Although in our illustrations we use relatively simple estimators for the densities and their derivatives based on variable bandwidth kernels, an alternative would be to use methods directy aimed at estimating derivatives of the logarithm of the density as in Pinkse and Schurter (2023). Detailed descriptions of how the new estimators are implemented, as well as Matlab code implementing all estimators with performance optimisations for these simulations, are available in the online supplementary material.[3] We present results for three sets of simulations, one with a range of known distributions for the potential outcomes, so we can directly assess the ability of the proposed methods to adapt to different distributions, and two with simulations based on real data: one based on housing prices and one based on medical expenditures, both with thick-tailed distribution.

4.1 Simulations with known distributions

We simulate samples of $n = 20{,}000$ observations, half of which are treated, and report summary statistics based on 10,001 simulated samples (using an odd number so that the median is unique). We repeat the simulation study for standardised Normal, Double Exponential (Laplace), and Cauchy distributions for the potential outcomes. The difference in means is the maximum

[1] The Hodges–Lehmann estimator is equal to the median of all pairwise differences between treated and control observations.

[2] We apply the ideas of Jaeckel (1971b) for the optimal trimmed mean to choose the parameters for the estimators in Section 2.6, see Theorem A.3 in the online supplementary appendix. We allow anywhere between no trimming (difference in means) and the extreme of trimming all but the medians (difference in medians). While including the extremes is not covered by the theory, this approach appeared to work best in our simulations.

[3] The fully documented R package is available at https://github.com/michaelpollmann/parTreat. Details on the empirical implementation of our estimators are in the online supplementary material. For a sample of 1,000 treated and 1,000 control observations, the R package computes estimates and standard errors practically instantaneously. With very large samples, the derivatives of the log density can be precomputed on a random sub-sample of the data for similarly fast computation.

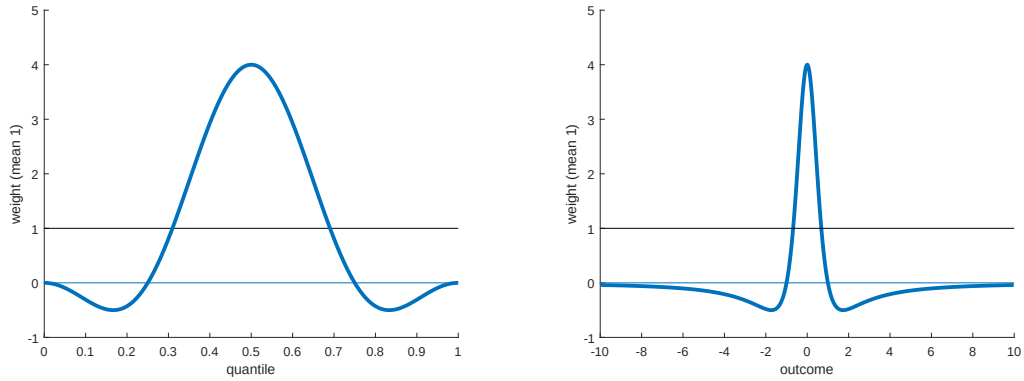

Figure 2. The efficient weight function for the Cauchy distribution. The weights, normalised to have mean 1, are plotted against the quantile (left) and against the value of the observations (right). For the figure on the right, we only plot the range from −10 to 10, which corresponds to approximately the 0.03–0.97 quantile. At this point, the weight is approximately −0.04, and weights for more extreme quantiles are closer to 0.

likelihood estimator for Normally distributed data, and so will do well there, but may perform poorly for thicker tailed distributions such as the Double Exponential distribution and in particular the Cauchy distribution. The difference in medians is the maximum likelihood estimator for the Double Exponential distribution, and relatively robust to thick tails and outliers, and so is expected to perform reasonably well across all specifications, but not as well as the efficient estimators for the Normal.

For the simulations with known distributions, we can derive the functional form for the optimal weights for the quantile-based estimator. The optimal weights for the waq estimator are proportional to the (estimated) second derivative of the log density. For the Normal distribution, $\frac{\partial^2 \ln f}{\partial y^2}(y) = -\frac{1}{\sigma^2}$, implying the optimal weights are constant. The density of the Double Exponential distribution is such that the optimal weights asymptotically place all weight close to the median. For the standard Cauchy distribution the efficient weights w on the difference in $u \in (0, 1)$ quantiles of treated and control distributions are $w_f(u) \propto -\cos(2\pi u)\sin(\pi u)^2$, shown in Figure 2. Most of the weight is concentrated around the median, with strictly negative weights outside the [0.25, 0.75] quantile range.

The efficient estimators perform well across distributions, and confidence intervals based either on estimates of the analytic variance formulas or on the bootstrap achieve their nominal coverage levels, as shown in Table 1. Their standard deviations are close to the theoretical efficiency bound, as shown in column 4, labelled relative efficiency, where values larger than one imply standard deviations of the estimator in excess of the efficiency bound. The eif and waq estimators are close to the most efficient estimator for the Normal, Double Exponential, and Cauchy distributions. The last columns show that the confidence intervals are close to their nominal coverage for each distribution. For computational convenience in the simulations, the confidence intervals based on the bootstrap variance use the m-out-of-n bootstrap (Bickel et al., 2012), with $m = 2{,}000$ (half treated, half control), to estimate the variance of the estimators. Even with these smaller sample sizes, the density estimates calculated within each bootstrap sample appear to be sufficiently good to yield reasonable confidence intervals for the estimators.

4.2 Simulations with house price data

In the second set of simulations, we use house price data from the replication files of Linden and Rockoff (2008) available at Linden and Rockoff (2019). They obtained property sales data for Mecklenburg County, North Carolina, between January 1994 and December 2004. They dropped sales below \$5,000 and above \$1,000,000, such that 170,239 observations remain, which we take as our population of interest. Despite the trimming the distribution is noticeably skewed (skewness 2.2) and thick tailed (kurtosis 9.5). Even after taking logs, the distribution is heavy-tailed with kurtosis equal to 5.1. Figure 3 plots a histogram for house prices, both in levels and in logs, along with

Table 1. Summary statistics for simulations with different distributions: Normal, Double Exponential, Cauchy

						95% C.I.		Boot. var. C.I.	
Estimator	Bias	Standard deviation	Relative efficiency	RMSE	MAD	Coverage	Median length	Coverage	Median length
Normal distribution:									
Diff. in means	0.000	0.014	1.01	0.014	0.010	0.95	0.055	0.95	0.055
Diff. in medians	−0.000	0.018	1.26	0.018	0.012	0.95	0.069	0.95	0.069
Hodges–Lehmann	0.000	0.015	1.03	0.015	0.010	0.95	0.057	0.95	0.057
Adaptive trim	0.000	0.015	1.03	0.015	0.010	0.94	0.055	0.95	0.057
Adaptive wins.	0.000	0.014	1.01	0.014	0.010	0.95	0.055	0.95	0.055
eif	0.000	0.014	1.02	0.014	0.010	0.95	0.056	0.95	0.057
waq	0.000	0.014	1.02	0.014	0.010	0.95	0.056	0.95	0.056
Double Exponential distribution:									
Diff. in means	−0.000	0.020	1.43	0.020	0.014	0.95	0.078	0.95	0.078
Diff. in medians	0.000	0.014	1.01	0.014	0.010	0.97	0.061	0.95	0.057
Hodges–Lehmann	−0.000	0.016	1.17	0.016	0.011	0.95	0.064	0.95	0.064
Adaptive trim	0.000	0.014	1.02	0.014	0.010	0.95	0.059	0.95	0.057
Adaptive wins.	−0.000	0.018	1.29	0.018	0.012	0.96	0.076	0.95	0.069
eif	−0.000	0.015	1.06	0.015	0.010	0.95	0.060	0.96	0.060
waq	−0.000	0.015	1.07	0.015	0.010	0.95	0.060	0.96	0.061
Cauchy distribution:									
Diff. in means	0.462	127.149	6357.47	127.144	2.047	0.98	9.324	0.98	9.292
Diff. in medians	0.000	0.022	1.11	0.022	0.015	0.97	0.093	0.95	0.087
Hodges–Lehmann	0.000	0.026	1.28	0.026	0.018	0.95	0.101	0.95	0.101
Adaptive trim	0.000	0.022	1.08	0.022	0.015	0.95	0.092	0.95	0.086
Adaptive wins.	0.000	0.024	1.19	0.024	0.016	0.98	0.111	0.97	0.100
eif	0.000	0.020	1.01	0.020	0.014	0.97	0.085	0.97	0.088
waq	0.000	0.021	1.05	0.021	0.014	0.96	0.085	0.98	0.099

Note. Statistics shown are based on 10,001 simulated samples. Each sample has 10,000 treated observations and 10,000 control observations. Relative efficiency is the ratio of standard deviation to the square root of the efficiency bound. Columns labelled length show the median length of the confidence intervals. RMSE = root mean squared error. MAD = median absolute deviation.

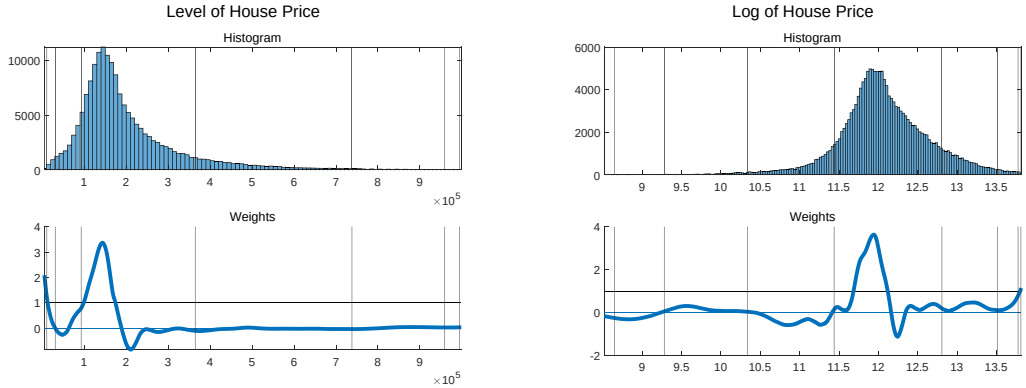

Figure 3. Histogram of house prices, in levels and logs, as well as estimated optimal weights (minus the second derivative of the log density) based on all 170,239 observations. The weights are normalised to be mean 1 (thick horizontal line). Some weights are below 0 (thin horizontal line). Vertical lines indicate the 0.0001, 0.001, 0.01, 0.1, 0.9, 0.99, 0.999, 0.9999 quantiles.

the estimated optimal weights (minus the second derivative of the log density) based on all 170,239 observations.

We base simulations on this data by drawing samples of size $n = 20{,}000$, and randomly assigning exactly half of each sample to the treatment group and the remaining half to the control group, with a zero treatment effect. Within each sample, observations are drawn from the population without replacement, but sampling is independent across samples, such that observations may appear in multiple samples. We estimate the efficiency bound using density estimates based on all observations. We compute the same estimators as in the simulations of the previous section, with a small adjustment to the adaptively trimmed and winsorised means where we fix the trimming and winsorising percentiles on the left to 0% (no trimming/winsorising), and only adaptively choose the threshold on the right.

Table 2 summarises the simulation results based on 10,001 simulated samples. When the house prices are in levels, the standard deviation of the difference in means estimator is twice as large as that of efficient estimators. For the difference in medians and the Hodges–Lehmann estimators, which are less affected by outliers in the data, the standard deviation is larger by approximately 30% and 20%, respectively. Confidence intervals, based on estimated variances and asymptotic normal approximations, have close to nominal coverage throughout, and are meaningfully shorter for the efficient estimators we propose.

In Figure 4, we show the root mean squared error and coverage of 95% confidence intervals *both relative to the average treatment effect* under deviations from the constant treatment effect model.[4] In the top panel, the unit-level treatment effects are independent draws from a normal distribution with mean equal to 0.1 standard deviations of the (population) standard deviation of the potential outcomes in the absence of treatment. On the horizontal axis, we vary the standard deviation of the normal distribution as a fraction q of the (population) standard deviation of the control potential outcomes, simulating 10,001 samples for each value. When $q > 0$, the constant treatment effect model is misspecified. In the bottom panel, the unit-level treatment effects are 0 with probability q and t with probability $1 - q$, where t is chosen as a function of q such that the average treatment effect is constant across all simulations and the same as in the top panel. The difference in means estimator is unbiased for the average treatment effect regardless of the value of q, so its large root mean squared error is due to its variance. In both panels, the influence function-based and waq estimators are only (asymptotically) unbiased for the average treatment effect when $q = 0$.

We also estimate a proportional treatment effect (multiplicative) model and translate the estimated coefficients into level effects. Under the multiplicative model, $Y_i(1) = \theta Y_i(0)$. When

[4] The design of these simulations was kindly suggested by one of the referees.

Table 2. Summary statistics for simulations based on house price data from Linden and Rockoff (2008)

						95% C.I.		Boot. var. C.I.	
Estimator	Bias	Standard deviation	Relative efficiency	RMSE	MAD	Coverage	Length	Coverage	Length
Effect in levels based on additive model in levels									
Diff. in means	−28	1,871	1.92	1,871	1,251	0.95	7,343	0.95	7,339
Diff. in medians	8	1,259	1.30	1259	855	0.95	4,989	0.95	4,893
Hodges–Lehmann	−5	1,138	1.17	1,138	757	0.95	4,487	0.95	4,487
Adaptive trim	5	989	1.02	989	651	0.98	4,560	0.95	3,914
Adaptive wins.	4	1,114	1.15	1,114	738	0.99	6,378	0.97	4,760
eif	13	941	0.97	941	628	0.95	3,819	0.95	3,768
waq	13	946	0.97	946	626	0.95	3,819	0.95	3,859
Multiplicative parameter: additive model in logs									
Diff. in means	−0.0000	0.0083	1.35	0.0083	0.0055	0.95	0.033	0.95	0.032
Diff. in medians	0.0000	0.0075	1.23	0.0075	0.0051	0.95	0.030	0.95	0.029
Hodges–Lehmann	−0.0000	0.0072	1.17	0.0072	0.0048	0.95	0.028	0.95	0.028
Adaptive trim	−0.0000	0.0083	1.35	0.0083	0.0055	0.95	0.033	0.95	0.032
Adaptive wins.	−0.0000	0.0083	1.35	0.0083	0.0055	0.95	0.033	0.95	0.032
eif	−0.0000	0.0067	1.08	0.0067	0.0045	0.95	0.026	0.95	0.026
waq	−0.0000	0.0067	1.08	0.0067	0.0044	0.95	0.026	0.95	0.027
Effect in levels based on additive model in logs									
diff. in means	−7	1,695	1.35	1,695	1,125	0.95	6,658	0.95	6,657
Diff. in medians	9	1545	1.23	1,545	1,048	0.95	6,128	0.95	6,001
Hodges–Lehmann	−6	1,468	1.17	1,468	976	0.95	5,786	0.95	5,783
Adaptive trim	−7	1,695	1.35	1,695	1,125	0.95	6,658	0.95	6,657
Adaptive wins.	−7	1,695	1.35	1,695	1,125	0.95	6,658	0.95	6,657
eif	−5	1,363	1.08	1,363	914	0.95	5,374	0.95	5,415
waq	−3	1,364	1.08	1,364	910	0.95	5,374	0.95	5,462

Note. Statistics shown are based on 10,001 simulated samples. Each sample has 10,000 treated observations and 10,000 control observations. Relative efficiency is the ratio of standard deviation to the square root of the efficiency bound, which we calculate from density estimates based on the full sample. Columns labelled length show the median length of the confidence intervals. The adaptively trimmed mean and adaptively winsorised mean only apply trimming at the top, but not at the bottom. RMSE = root mean squared error. MAD = median absolute deviation.

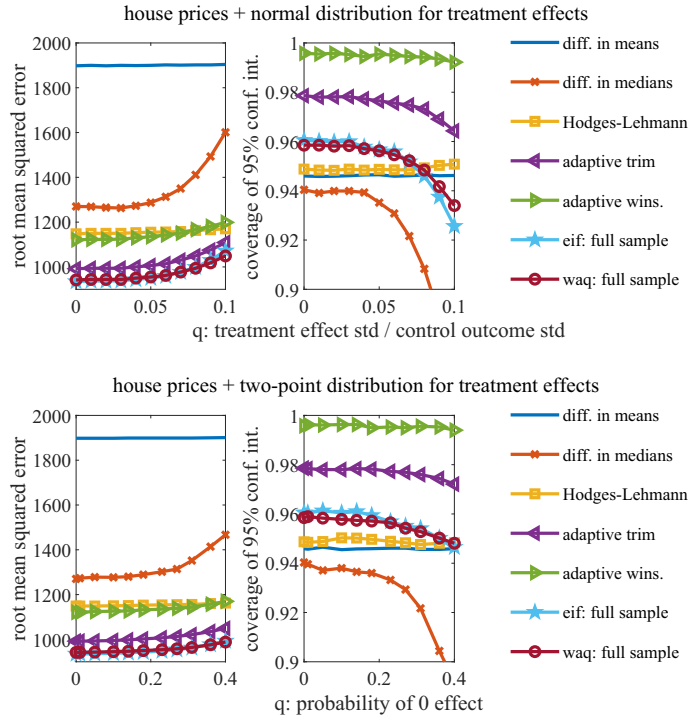

Figure 4. Root mean squared error and coverage of 95% confidence intervals relative to the average treatment effect (fixed at 0.1 standard deviations of the outcome in the absence of treatment) in simulations with heterogeneous treatment effects in the house price data. The horizontal axis varies the amount of heterogeneity. When $q = 0$, there is no heterogeneity such that the constant additive treatment effect model is correctly specified.

outcomes are strictly positive, this is identical to an additive model for log outcomes; $\log(Y_i(1)) = \log(\theta) + \log(Y_i(0))$. Given an estimate $\hat{\tau}_{\log}$ of the additive model with the outcome in logs, the estimate of the level treatment effect is then

$$\hat{\tau} = \frac{1}{n}\left(\sum_{i=1}^{n}(1 - Z_i)\big((\exp(\hat{\tau}_{\log}) - 1)Y_i\big) + Z_i\big((1 - \exp(-\hat{\tau}_{\log}))Y_i\big)\right).$$

For estimates of the in-sample treatment effect, we therefore apply estimates $\hat{\tau}_{\log}$ to a population of interest with known means μ_{Y_0} and μ_{Y_1} and fixed treatment probability p as

$$\hat{\tau} = (1 - p)((\exp(\hat{\tau}_{\log}) - 1)\mu_{Y_0}) + p((1 - \exp(-\hat{\tau}_{\log}))\mu_{Y_1}).$$

Using the Delta method, if V is the asymptotic variance of $\hat{\tau}_{\log}$, then the asymptotic variance of $\hat{\tau}$, holding μ_{Y_0}, μ_{Y_1}, and p fixed, is

$$\big((1 - p)\exp(\tau_{\log})\mu_{Y_0} + p\exp(-\tau_{\log})\mu_{Y_1}\big)^2 V$$

which we estimate by replacing $\tau_{\log}$ with $\hat{\tau}_{\log}$ and V by the estimate of the variance, $\hat{V}$. For the purpose of these simulations, we set $\mu_{Y_0} = \mu_{Y_1}$ equal to the population mean of house prices, and $p = 1/2$.

When treatment effects are assumed to be proportional to potential outcomes, the proposed estimators for this multiplicative model are still more efficient than alternative estimators, but the gains are smaller. The middle panel of Table 2 shows the quality of estimates of the multiplicative

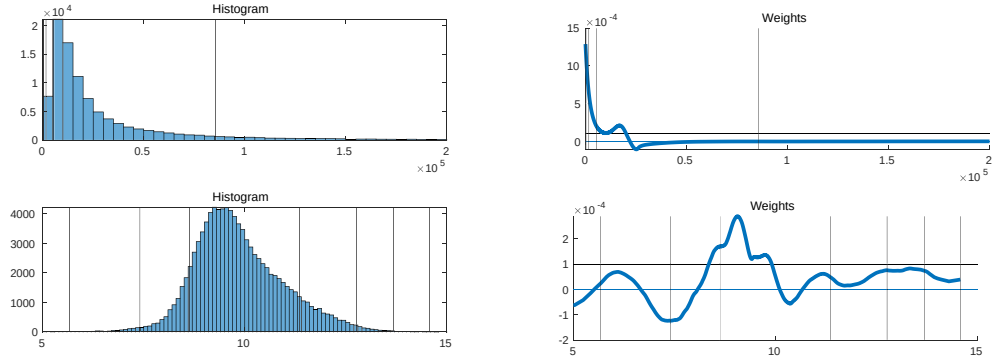

Figure 5. Histogram of medical expenditures per admission, in levels and logs, as well as estimated optimal weights (minus the second derivative of the log density) based on the 98,155 observations in the control group. For the figure in levels, vertical lines indicate the 0.001, 0.01, 0.1, and 0.9 quantiles; the figure is limited to below $200,000, such that the 0.99 and higher quantiles do not appear. For the figure in logs, vertical lines indicate the 0.001, 0.01, 0.1, 0.99, 0.999, 0.9999 quantiles. The weights are normalised to be mean 1 (black horizontal line). Some weights are below 0 (blue horizontal line).

parameter obtained by transforming outcomes into logs, $\log(\theta)$. As can be seen in Figure 3, the distribution of log house prices appears closer to the normal distribution with fewer 'outliers'. Consequently, the difference in means estimator, which is efficient for the normal distribution, comes noticeably closer to the efficiency bound than when outcomes are in levels. Nevertheless, the efficient estimators further reduce the variance. The bottom panel of Table 2 shows the same summary statistics when the multiplicative parameter is translated back into a level effect. The efficient estimators of the multiplicative parameter lead to treatment effects with smaller variance and shorter confidence intervals than the difference in means, irrespective of whether the latter is estimated in levels (top panel) or in logs and then translated into levels (bottom panel).

4.3 Medical expenditures data

Next, we present simulation results based on confidential medical expenditure data from the IBM MarketScan Research Database, following the sample construction of Koenecke et al. (2021, Figure 2). We restrict the sample to males, age 45–64, with pneumonia inpatient diagnosis and at least 1 year of continuous medical enrolment. For each patient, we consider the first inpatient admission only to abstract away from any dynamics. We focus on medical expenditure as the outcome variable. For each patient, we sum the payments recorded by MarketScan for this admission. In total, we use data on 103,662 admissions.[5] Figure 5 plots a histogram for medical expenditure in levels and in logs along with the estimated optimal weights (minus the second derivative of the log density). We also observe a treatment variable in this data set, the (prior) use of alpha blockers, which Koenecke et al. (2021) find may improve health outcomes during respiratory distress by preventing hyperinflammation.

We design a simulation study similar to those of the previous sections, treating the receipt of alpha blockers as randomly assigned in our population. This allows us to study (coverage) properties of the estimators and inference procedures in settings where the parametric treatment effect model is not (necessarily) correctly specified. For these simulations, each sample is a draw, without replacement, of 200 of the 5,507 observations in the treatment group and 3,565 of the 98,155 observations in the control group. While the control group is smaller than in our other simulations, it remains sufficiently large to estimate the density and its derivatives required for our estimators. In this simulation design, F_0 is given by the empirical distribution of the control group, and F_1 is given by the empirical distribution of the treatment group, of the full sample. Although it is not necessarily correct, the additive treatment effect model may offer a reasonable approximation, and we are interested in the performance of inference methods when the conditions for our theoretical

[5] The sample size differs slightly from that reported by Koenecke et al. (2021) due to missing expenditure data for a small number of admissions.

Table 3. Summary statistics for simulations based on the medical expenditures data in logs, across 1,000 simulated samples

					95% C.I.		Boot. var. C.I.	
Estimator	Bias	Standard deviation	RMSE	MAD	Coverage	Length	Coverage	Length
Effect in levels based on additive model in levels								
Error and coverage relative to population difference in means								
Diff. in means	−108	4,567	4,566	3,016	0.93	16,935	0.93	16,888
Diff. in medians	3,345	1,271	3,578	3,222	0.46	6,161	0.28	5,259
Hodges–Lehmann	3,490	891	3,602	3,464	0.03	3,664	0.01	3,534
Adaptive trim	3,877	641	3,929	3,860	0.00	3,052	0.00	2,778
Adaptive wins.	2,816	1,384	3,138	2,786	0.52	5,668	0.57	6,023
eif	3,928	682	3,987	3,922	0.01	4,878	0.00	3,240
waq	3,846	792	3,927	3,848	0.03	4,878	0.01	4,018
Multiplicative parameter: additive model in logs								
Error and coverage relative to population difference in means of outcomes in logs								
Diff. in means	−0.001	0.077	0.077	0.052	0.95	0.30	0.95	0.30
Diff. in medians	0.005	0.080	0.080	0.051	0.97	0.33	0.95	0.32
Hodges–Lehmann	0.008	0.071	0.071	0.048	0.95	0.29	0.95	0.28
Adaptive trim	0.016	0.074	0.076	0.053	0.95	0.29	0.95	0.29
Adaptive wins.	−0.002	0.078	0.078	0.052	0.94	0.30	0.95	0.31
eif	0.029	0.066	0.072	0.049	0.95	0.27	0.94	0.26
waq	0.028	0.066	0.071	0.047	0.95	0.27	0.95	0.27
Effect in levels based on additive model in logs								
Error and coverage relative to population difference in means								
Diff. in means	2,423	2,871	3,756	2,522	0.90	11,167	0.91	11,138
Diff. in medians	2,647	3,009	4,006	2,499	0.92	12,318	0.92	11,951
Hodges–Lehmann	2,747	2,673	3,832	2,672	0.88	10,730	0.87	10,390
Adaptive trim	3,062	2,809	4,154	2,983	0.86	11,055	0.85	10,850

(*continued*)

Table 3. Continued

Estimator	Bias	Standard deviation	RMSE	MAD	95% C.I. Coverage	95% C.I. Length	Boot. var. C.I. Coverage	Boot. var. C.I. Length
Adaptive wins.	2,365	2,915	3,752	2,544	0.90	11,084	0.92	11,389
eif	3,532	2,525	4,341	3,481	0.78	10,452	0.76	10,014
waq	3,464	2,523	4,285	3,413	0.79	10,441	0.77	10,147

Note. Each sample has 200 treated observations and 3,565 control observations. Coverage and median length refer to 95% confidence intervals. RMSE = root mean squared error. MAD = median absolute deviation.

results are not (quite) met in this particular application. The simulation results reported in Table 3, for the same estimators as in Section 4.2, suggest that the proposed estimators perform reasonably well in this setting despite mis-specification.

5 Conclusion

In many modern settings where randomised experiments are used to estimate treatment effects, the presence of heavy-tailed distributions can lead to larger standard errors. Often researchers use winsorising with ad hoc thresholds to address this. Here, we develop systematic methods for obtaining more precise inferences using parametric models for the treatment effects, while avoiding the specification of models for the potential outcomes. We present results for semi-parametric effiency bounds, suggest efficient estimators, and show in simulations that these methods can be effective in realistic settings.

In particular, we recommend the semi-parametrically efficient estimator under the constant additive treatment effect model. Although one may not think the constant additive treatment effect assumption holds exactly, the fact that the estimator can be interpreted as estimating a weighted average of the quantile treatment effects make this an attractive choice.

In this discussion, we do not incorporate covariates or pre-treatment variables. One could combine the ideas exposited in the current paper with models for the control potential outcome. One may also wish to incorporate covariates in the model for the quantile treatment effects and thus allow for heterogenous treatment effects, e.g. Chen and Au (2022) and Wager and Athey (2018). Another interesting avenue is to consider alternative estimators for the current model by first estimating the unrestricted quantile functions, followed by minimum distance methods, as in L. Alvarez (2022) and L. A. Alvarez and Biderman (2022).

Conflict of interests: None declared.

Funding

Generous support from the Office of Naval Research through ONR grants N00014-17-1-2131 and N00014-19-1-2468 is gratefully acknowledged. Data for the health application were accessed using the Stanford Center for Population Health Sciences Data Core.

Data availability

The source for the house price data is Linden and Rockoff (2019). The relevant columns of the data are also included in the online supplementary material of this paper. The medical data are proprietary and confidential. Researchers at some institutions, in particular medical schools, may have access to the IBM MarketScan Research Database from which the data are drawn following Koenecke et al. (2021, Figure 2).

Supplementary material

Supplementary material is available online at *Journal of the Royal Statistical Society: Series B*.

References

Alvarez L., Chiann C., & Morettin P. (2022). Inference in parametric models with many L-moments. arXiv, arXiv:2210.04146, preprint: not peer reviewed.

Alvarez L. A., & Biderman C. (2022). Semiparametric analysis of randomised experiments using L-moments. Fundação Getulio Vargas, working paper: not peer reviewed.

Bickel P. J. (1965). On some robust estimates of location. *The Annals of Mathematical Statistics*, *36*(3), 847–858. https://doi.org/10.1214/aoms/1177700058

Bickel P. J. (1967). Some contributions to the theory of order statistics. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics* (pp. 575–591). University of California Press.

Bickel P. J., & Doksum K. A. (2015). *Mathematical statistics: Basic ideas and selected topics* (Vol. 2). CRC Press.

Bickel P. J., Götze F., & van Zwet W. R. (2012). Resampling fewer than *n* observations: Gains, losses, and remedies for losses. In *Selected works of Willem van Zwet* (pp. 267–297). Springer.

Bickel P. J., Klaassen C. A., Ritov Y., & Wellner J. A. (1993). *Efficient and adaptive estimation for semiparametric models* (Vol. *4*). Johns Hopkins University Press.

Bickel P. J., & Lehmann E. L. (1975a). Descriptive statistics for nonparametric models I. Introduction. *The Annals of Statistics*, *3*(5), 1038–1044. https://doi.org/10.1214/aos/1176343239

Bickel P. J., & Lehmann E. L. (1975b). Descriptive statistics for nonparametric models II. Location. *The Annals of Statistics*, *3*(5), 1045–1069. https://doi.org/10.1214/aos/1176343240

Bickel P. J., & Lehmann E. L. (1976). Descriptive statistics for nonparametric models. III. Dispersion. *The Annals of Statistics*, *4*(6), 1139–1158. https://doi.org/10.1214/aos/1176343648

Bickel P. J., & Lehmann E. L. (2012). Descriptive statistics for nonparametric models IV. Spread. In *Selected Works of EL Lehmann* (pp. 519–526). Springer.

Chen A., & Au T. C. (2022). Robust causal inference for incremental return on ad spend with randomized paired geo experiments. *The Annals of Applied Statistics*, *16*(1), 1–20. https://doi.org/10.1214/21-AOAS1493

Chernoff H., Gastwirth J. L., & Johns M. V. (1967). Asymptotic distribution of linear combinations of functions of order statistics with applications to estimation. *The Annals of Mathematical Statistics*, *38*(1), 52–72. https://doi.org/10.1214/aoms/1177699058

Crump R. K., Hotz V. J., Imbens G. W., & Mitnik O. A. (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, *96*(1), 187–199. https://doi.org/10.1093/biomet/asn055

Cuzick J. (1992a). Efficient estimates in semiparametric additive regression models with unknown error distribution. *The Annals of Statistics*, *20*(2), 1129–1136. https://doi.org/10.1214/aos/1176348675

Cuzick J. (1992b). Semiparametric additive regression. *Journal of the Royal Statistical Society. Series B (Methodological)*, *54*(3), 831–843. https://doi.org/10.1111/j.2517-6161.1992.tb01455.x

Doksum K. (1974). Empirical probability plots and statistical inference for nonlinear models in the two-sample case. *The Annals of Statistics*, *2*(2), 267–277. https://doi.org/10.1214/aos/1176342662

Doksum K. A., & Sievers G. L. (1976). Plotting with confidence: Graphical comparisons of two populations. *Biometrika*, *63*(3), 421–434. https://doi.org/10.1093/biomet/63.3.421

Firpo S. (2007). Efficient semiparametric estimation of quantile treatment effects. *Econometrica*, *75*(1), 259–276. https://doi.org/10.1111/j.1468-0262.2007.00738.x

Firpo S., Fortin N. M., & Lemieux T. (2009). Unconditional quantile regressions. *Econometrica*, *77*(3), 953–973. https://doi.org/10.3982/ECTA6822

Fisher R. A. (1937). *The design of experiments*. Oliver and Boyd.

Govindarajulu Z., Le Cam L., & Raghavachari M. (1967). Generalizations of theorems of Chernoff and Savage on the asymptotic normality of test statistics. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics* (pp. 609–638). University of California Press.

Gupta S., Kohavi R., Tang D., Xu Y., Andersen R., Bakshy E., Cardin N., Chandran S., Chen N., Coey D., Curtis M., Deng A., Duan W., Forbes P., Frasca B., Guy T., Imbens G. W., Saint Jacques G., Kantawala P., … Yashkov I. (2019). Top challenges from the first practical online controlled experiments summit. *ACM SIGKDD Explorations Newsletter*, *21*(1), 20–35. https://doi.org/10.1145/3331651.3331655

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

Hodges Jr J. L., & Lehmann E. L. (1963). Estimates of location based on rank tests. *The Annals of Mathematical Statistics*, *34*(2), 598–611. https://doi.org/10.1214/aoms/1177704172

Huber P. J. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, *35*(1), 73–101. https://doi.org/10.1214/aoms/1177703732

Huber P. J. (2011). *Robust statistics*. Springer.

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

Jaeckel L. A. (1971a). Robust estimates of location: Symmetry and asymmetric contamination. *The Annals of Mathematical Statistics*, *42*(3), 1020–1034. https://doi.org/10.1214/aoms/1177693330

Jaeckel L. A. (1971b). Some flexible estimates of location. *The Annals of Mathematical Statistics*, *42*(5), 1540–1552. https://doi.org/10.1214/aoms/1177693152

Klaassen C. A. J. (1987). Consistent estimation of the influence function of locally asymptotically linear estimators. *The Annals of Statistics*, *15*(4), 1548–1562. https://doi.org/10.1214/aos/1176350609

Koenecke A., Powell M., Xiong R., Shen Z., Fischer N., Huq S., Khalafallah A. M., Trevisan M., Sparen P., Carrero J. J., Nishimura A., Caffo B., Stuart E. A., Bai R., Staedtke V., Thomas D. L., Papadopoulos N., Kinzler K. W., Vogelstein B., …Athey S. (2021). Alpha-1 adrenergic receptor antagonists to prevent hyperinflammation and death from lower respiratory tract infection. *eLife*, *10*, e61700. https://doi.org/10.7554/eLife.61700

Kohavi R., Tang D., & Xu Y. (2020). *Trustworthy online controlled experiments: A practical guide to a/b testing*. Cambridge University Press.

Le Cam L., & Yang G. L. (1988). On the preservation of local asyptotic normality under information loss. *The Annals of Statistics*, *16*(2), 483–520. https://doi.org/10.1214/aos/1176350817

Lehmann E. L., & D'Abrera H. J. (1975). *Nonparametrics: Statistical methods based on ranks*. Holden-Day.
Lewis R. A., & Rao J. M. (2015). The unfavorable economics of measuring the returns to advertising. *The Quarterly Journal of Economics*, *130*(4), 1941–1973. https://doi.org/10.1093/qje/qjv023
Li F., Morgan K. L., & Zaslavsky A. M. (2018). Balancing covariates via propensity score weighting. *Journal of the American Statistical Association*, *113*(521), 390–400. https://doi.org/10.1080/01621459.2016.1260466
Linden L., & Rockoff J. E. (2008). Estimates of the impact of crime risk on property values from Megan's laws. *American Economic Review*, *98*(3), 1103–1127. https://doi.org/10.1257/aer.98.3.1103
Linden L., & Rockoff J. E. (2019). *Replication data for: Estimates of the impact of crime risk on property values from Megan's laws*. American Economic Association. https://doi.org/10.3886/E113243V1
Neyman J. (1923/1990). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statistical Science*, *5*(4), 465–472. https://doi.org/10.1214/ss/1177012031
Pinkse J., & Schurter K. (2023). Estimates of derivatives of (log) densities and related objects. *Econometric Theory*, *39*(2), 1–36. https://doi.org/10.1017/S0266466621000529
Robins J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period application to control of the healthy worker survivor effect. *Mathematical Modelling*, *7*(9–12), 1393–1512. https://doi.org/10.1016/0270-0255(86)90088-6
Rosenbaum P. R. (1993). Hodges–Lehmann point estimates of treatment effect in observational studies. *Journal of the American Statistical Association*, *88*(424), 1250–1253. https://doi.org/10.1080/01621459.1993.10476405
Rubin D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*(5), 688–701. https://doi.org/10.1037/h0037350
Stigler S. M. (1974). Linear functions of order statistics with smooth weight functions. *The Annals of Statistics*, *2*(4), 676–693. https://doi.org/10.1214/aos/1176342756
Stone C. J. (1975). Adaptive maximum likelihood estimators of a location parameter. *The Annals of Statistics*, *3*(2), 267–284. https://doi.org/10.1214/aos/1176343056
Taddy M., Lopes H. F., & Gardner M. (2016). 'Scalable semiparametric inference for the means of heavy-tailed distributions', arXiv, arXiv:1602.08066, preprint: not peer reviewed.
Tripuraneni N., Madeka D., Foster D., Perrault-Joncas D., & Jordan M. I. (2021). 'Meta-analysis of randomized experiments with applications to heavy-tailed response data', arXiv e-prints, arXiv–2112, preprint: not peer reviewed.
Vansteelandt S., & Dukes O. (2022). Assumption-lean inference for generalised linear model parameters. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *84*(3), 657–685. https://doi.org/10.1111/rssb.12504
Vansteelandt S., & Joffe M. (2014). Structural nested models and G-estimation: The partially realized promise. *Statistical Science*, *29*(4), 707–731. https://doi.org/10.1214/14-STS493
Wager S., & Athey S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, *113*(523), 1228–1242. https://doi.org/10.1080/01621459.2017.1319839
Wu C. J., & Hamada M. S. (2011). *Experiments: Planning, analysis, and optimization* (Vol. *552*). John Wiley & Sons.
Yu C., & Yao W. (2017). Robust linear regression: A review and comparison. *Communications in Statistics-Simulation and Computation*, *46*(8), 6261–6282. https://doi.org/10.1080/03610918.2016.1202271