Efficient Estimation of Quantiles in Missing Data Models

Iván Díaz

Google Inc.

December 21, 2015

Abstract

We present a method to estimate the quantile of a variable subject to missingness, under the missing at random assumption. Our proposed estimator is locally efficient, $\sqrt{n}$-consistent, asymptotically normal, and doubly robust, under regularity conditions. We use Monte Carlo simulation to compare our proposal to the one-step and inverse-probability weighted estimators. Our estimator is superior to both competitors, with a mean squared error up to 8 times smaller than the one-step estimator, and up to 2.5 times smaller than an inverse probability weighted estimator. We develop extensions for estimating the causal effect of treatment on a population quantile among the treated. Our methods are motivated by an application with a heavy tailed continuous outcome. In this situation, the efficiency bound for estimating the effect on the mean is often large or infinite, ruling out $\sqrt{n}$-consistent inference and reducing the power for testing hypothesis of no treatment effect. Using quantiles (e.g., the median) may yield more accurate measures of the treatment effect, along with more powerful hypothesis tests. In our application, the proposed estimator of the effect on the median yields hypothesis tests of no treatment effect up to two times more powerful, and its variance is up to four times smaller than the variance of its mean counterpart.

1 Introduction

Estimation of quantiles in missing data models is a statistical problem with applications to a variety of research areas. For example, policy makers may be interested in evaluating the effect of an educational program on the tails of the skill distribution. In this case quantile treatment effects may be useful since they capture intervention effects that are heterogeneous across the outcome distribution. Quantiles may also be useful in economics research to compute inequality indicators such as the Gini coefficient, and may be used in adaptive clinical trials to estimate stopping rules in interim analyses since quantile estimation does not require completion of the study.

Our methods are motivated by an application to estimation of the causal effect of treatment on an outcome whose distribution exhibits heavy tails. The data we consider arises as part of various sales and services programs targeted to introduce new features to users of the AdWords advertisement platform at Google Inc. A important question for decision makers is thus to quantify the causal effect of these programs on the advertisers' spend through AdWords. The outcome we consider exhibits heavy tails, as there is a small but non-trivial number of advertisers who spend large quantities through on AdWords. Heavy tailed distributions are often characterized by large or infinite variance, which in turn yields a large or infinite efficiency bound for estimating the effect of treatment on the mean. As a consequence, the variance of all regular estimators is also large, possibly precluding $\sqrt{n}$-consistent inference and statistical significance at most plausible sample sizes.

As an alternative to estimation of the effect on the mean, in this document we present a methodology to estimate the causal effect on the q-th quantile. Our estimator is locally efficient in the non-parametric model and asymptotically linear, under certain regularity conditions. In our application, estimating a collection of quantiles of interest (e.g., 25%, 50% and 75%) allows us to make statements about treatment effects, even though we would have difficulty making similar statements for the mean, due to the large variability caused by the heavy tailed distribution.

Our goal is to estimate an unconditional quantile. An alternative goal, not considered here, is to estimate an outcome quantile conditional on the values of certain covariates. Though we do not estimate conditional quantiles, we use covariate information in order to correctly identify the unconditional quantiles under the missing at random assumption.

In order to assess the performance of the proposed estimator in our real data application, we use Monte Carlo simulations based on a real dataset to approximate the bias, variance, mean squared error, and coverage probability of the confidence interval estimators for our application. Our results corroborate the theoretical property that the proposed estimator has the best performance across various modeling scenarios in comparison to the available alternatives of one-step and inverse-probability weighted estimation. We also use the simulation study to demonstrate that estimation of the effect on the median has a smaller variance and improved power compared to the effect on the mean. In the worst case scenario, a hypothesis test for a zero effect using the effect on the mean as a test statistic yields a power of 0.16, whereas its mean counterpart yields a power of 0.84.

Various proposals exist that address the problem we consider. None of them, however, has the properties achieved by our estimator, which are outlined in the abstract. Wang and Qin (2010) consider pointwise estimation of the distribution function using the augmented inverse probability weighted estimator applied to an indicator function, where the missingness probabilities and observed outcome distribution functions are estimated via kernel regression. They propose to use the distribution function to estimate the relevant quantiles using a plug-in estimator (i.e., the inverse of the distribution function). Their approach suffers from various flaws stemming from the fact that the estimated distribution function may be ill-defined: direct inverse probability weighting may generate estimates outside $[0, 1]$, and pointwise estimation may yield a non-monotonic function. In addition, their approach may not be used in high dimensions since kernel estimators suffer from the curse of dimensionality. Zhao et al. (2013) propose similar estimators for non-ignorable missing data, under the assumption that the missingness mechanism is linked to the outcome through a parametric model that can be estimated from external data sources. Liu et al. (2011), Cheng and Chu (1996), and Hu et al. (2011) consider estimators that yield estimated distribution functions in the parameter space, relying either on kernel estimators for the outcome distribution function, or knowledge of the true missingness probabilities. Firpo (2007) proposes to estimate the quantiles by minimizing an inverse probability weighted check loss function. Their estimator achieves non-parametric consistency by means of a propensity score estimated as a logistic power series whose degree increases with sample size. Melly (2006), Frölich and Melly (2013), and Chernozhukov et al. (2013) consider estimation of the quantiles under a linear parametric model for the distribution and quantile functions, respectively. Unfortunately their parametric assumptions are seldom realistic and generally yield inconsistent and irregular estimators.

Our paper is organized as follows. In Section 2 we introduce the problem in terms of a closely related one: estimating the distribution function of an outcome missing at random. In Section 3 we present our proposed estimators for the quantiles of a variable missing at random as well as the effect of treatment on the quantiles, together with their asymptotic normality results and confidence interval estimators. In Section 4 we present a Monte Carlo simulation study based on a real dataset, where we illustrate the performance of our estimator and show the benefits of using the median as a location parameter for the counterfactual distribution in the presence of heavy tails. Finally, in Section 5, we discuss some concluding remarks.

2 Notation and Estimation Problem

Let Y denote an outcome observed only when a missingness indicator M equals one, and let X denote a set of observed covariates satisfying $Y \perp\!\!\!\perp M \mid X$. We use P_0 to denote the true joint distribution of the observed data $Z = (X, M, MY)$. We use the word *model* to refer to a set of probability distributions, and the expression *nonparametric model* to refer to the set of all distributions having a continuous density with respect to a dominating measure of interest. The word *estimator* is used to refer to a particular procedure or method for obtaining estimates of P_0 or functionals of it. We assume P_0 is in the nonparametric model $\mathcal{M}$, and use P to denote a general $P \in \mathcal{M}$. For a function $h(z)$, we denote $Ph = \int hdP$. For simplicity in the presentation we assume that X is finitely supported but the results generalize to infinite support by replacing the counting measure by an appropriate measure whenever necessary. Under the assumption that $P_0(M = 1 \mid X = x) > 0$ almost everywhere, the distribution $F_0(y) \equiv Pr(Y \leq y)$ is identified in terms of P_0 as

$$\begin{aligned} F_0(y) &= \sum_x Pr_0(Y \leq y \mid X = x) Pr_0(X = x) \\ &= \sum_x Pr_0(Y \leq y \mid M = 1, X = x) Pr_0(X = x) \\ &= \sum_x P_{Y,0}(y \mid 1, x) p_{X,0}(x), \end{aligned}$$

where we have denoted $P_Y(y \mid 1, x) \equiv Pr(Y \leq y \mid M = 1, X = x)$ and $p_X(x) \equiv Pr(X = x)$. We use f to denote the density corresponding to F and $e(x)$ to denote $Pr(M = 1 \mid X = x)$, following the convention in the propensity score literature. Consider the q-th quantile of the outcome distribution:

$$\chi = F^{-1}(q),$$

where we define the generalized inverse as $F^{-1}(q) = \inf\{y : F(y) \geq q\}$. We use the notation $\chi(P)$ to refer to the functional that maps an observed distribution P into a real number. Given a consistent estimator $\hat{P}$ of P_0, the plug-in estimator $\chi(\hat{P})$ is typically consistent, but it may be an inefficient and $\sqrt{n}$-inconsistent estimator. To remedy this, various methods exist in the semi-parametric statistics literature. The analysis of the asymptotic properties of such methods often relies on so-called von Mises expansions (von Mises, 1947) and on the theory of asymptotic lower bounds for estimation of regular parameters in semi=parametric models (see, e.g., Bickel et al., 1997; Newey, 1990).

The efficient influence function $D(Z)$ is one of the key concepts introduced by semi-parametric efficient estimation theory. This function characterizes all efficient, asymptotically linear estimators χ_n. Specifically, the following holds for any such estimator (see e.g., Bickel et al., 1997):

$$\sqrt{n}(\chi_n - \chi) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} D(Z_i) + o_P(1/\sqrt{n}). \tag{1}$$

This property of an estimator is very desirable since it allows the use of the central limit theorem to construct asymptotically valid confidence intervals and hypothesis tests. For our target of inference χ, the efficient influence function in the non-parametric models given below in Lemma 1.

Lemma 1 (Efficient Influence Function)**.** *The efficient influence function of χ at P in the non-parametric model is equal to*

$$D(Z) = -\frac{1}{f_Y(\chi)} \left[\frac{M}{e(X)} \left\{ I_{(-\infty,\chi]}(Y) - P_Y(\chi \mid 1, X) \right\} + P_Y(\chi \mid 1, X) - q \right], \tag{2}$$

where we have omitted the dependence of D on P. We add it explicitly whenever the omission may lead to confusion.

This lemma is a direct consequence of the functional delta method applied to the non-parametric estimator of $F_0(y)$, and the Hadamard derivative of the quantile functional given in Lemma 21.4 of van der Vaart (2000). Note that if there is no missingness, then $P(M=1)=1$, and $D(Z)$ reduces to

$$D(Z) = -\frac{1}{f_Y(\chi)}(I_{(-\infty,\chi]}(Y) - q).$$

Then (1) is the standard asymptotic linearity result for the sample median (see, e.g., corollary 21.5 of van der Vaart, 2000).

An equally important concept for the analysis of the estimators that we consider is the von Mises expansion of the parameter functional given in Lemma 2 below.

Lemma 2 (von Mises expansion)**.** *The quantile function $\chi(P)$ satisfies*

$$\chi(P) - \chi(P_0) = -P_0 D(P) + R_2(P, P_0), \tag{3}$$

where

$$R_2(P,P_0) = \frac{1}{f(\chi)} P_0 \left(\frac{e_0}{e} - 1\right)(h_{0,\chi} - h_\chi) + O(\chi_0 - \chi)^2.$$

Here we have denoted $\chi = \chi(P)$, $\chi_0 = \chi(P_0)$, $h_{0,\chi}(x) = P_0(Y \leq \chi \mid M = 1, X = x)$, and $h_\chi(x) = P(Y \leq \chi \mid M = 1, X = x)$. D is defined in (2).

Proof. Consider a Taylor expansion of the function $F_0(y)$ around $\chi = \chi(P)$ as follows:

$$F_0(y) = F_0(\chi) + f_0(\chi)(y - \chi) + O(y - \chi)^2.$$

Then

$$\chi - \chi_0 = -\frac{q - F_0(\chi)}{f_0(\chi)} + O(\chi - \chi_0)^2,$$

for $\chi_0 = \chi(P_0)$. Substitute this in the expression

$$R_2(P, P_0) = P_0 D(P) + \chi(P) - \chi(P_0),$$

to find

$$\begin{aligned} R_2(P,P_0) &= \frac{1}{f(\chi)} P_0 \left\{\frac{e_0}{e}(h_{0,\chi} - h_\chi) + h_\chi - q\right\} - \frac{q - F_0(\chi)}{f_0(\chi)} + O(\chi - \chi_0)^2 \\ &= \frac{1}{f(\chi)} P_0 \left(\frac{e_0}{e} - 1\right)(h_{0,\chi} - h_\chi) + P_0 \left(\frac{1}{f(\chi)} - \frac{1}{f(\chi_0)}\right)(q - h_{0,\chi}) + O(\chi - \chi_0)^2. \end{aligned}$$

Because $q = P_0 h_{0,\chi_0}$, the term in the middle is $O(\chi - \chi_0)^2$, and the result follows. □

This expansion suggests a natural correction to the plug-in estimator:

$$\chi_{n,\mathrm{os}} = \chi(\hat{P}) + P_n D(\hat{P}), \tag{4}$$

often referred to as the one-step estimator. Using results from empirical process theory, and under the assumption that $R_2(\hat{P}, P_0) = o_P(1/\sqrt{n})$, it may be proved that this estimator satisfies (1). This estimator was first proposed for a general parameter $\chi(P)$ by Levit (1976). Estimator (4) may have sub-optimal performance in finite samples because computation of D involves inverse probability weighting, and thus may yield unstable estimates. Alternatively, in the next section we propose to use an estimator $\chi_n = \chi(P_n)$ for a suitable estimator P_n satisfying

$$\frac{1}{n}\sum_{i=1}^{n} D_n(Z_i) = o_P(1/\sqrt{n}), \tag{5}$$

where D_n denotes $D(P_n)$. Using M-estimation and empirical process theory we derive the conditions under which this estimator satisfies (1). We present the proposed estimation algorithm along with theoretical results establishing its asymptotic properties.

3 Targeted Minimum Loss Based Estimator

The proposed estimation algorithm is given by the following iterative procedure, and constitutes an application of the general targeted minimum loss based estimator (TMLE) developed by van der Laan and Rubin (2006).

1. *Initialize.* Obtain initial estimates e_n and $P_{Y,n}$ of e_0 and $P_{Y,0}$. We discuss possible options to estimate these quantities below.

2. *Compute* χ_n. For the current estimate $P_{Y,n}$, compute

$$F_n(y) = \frac{1}{n}\sum_{i=1}^{n} P_{Y,n}(y \mid 1, X_i),$$

and $\chi_n = F_n^{-1}(q)$.

3. *Update* $P_{Y,n}$. Let $p_{Y,n}$ denote the density associated to $P_{Y,n}$, and consider the exponential model

$$p_{Y,\epsilon}(y \mid 1, x) = c(\epsilon, p_{Y,n}) \exp\{\epsilon D_{Y,n}(z)\} p_{Y,n}(y \mid 1, x),$$

where $c(\epsilon, p_{Y,n})$ is a normalizing constant and

$$D_{Y,n}(z) = \frac{1}{e_n(x)}\{I_{(-\infty,\chi_n]}(y) - P_{Y,n}(y \mid 1, x)\}$$

is the score of the model. Estimate ϵ as

$$\hat{\epsilon} = \arg\max_{\epsilon} \sum_{i=1}^{n} M_i \log p_{Y,\epsilon}(Y_i \mid 1, X_i).$$

Note that this MLE solves the score equation

$$\sum_{i=1}^{n} \frac{M_i}{e_n(X_i)} \{I_{(-\infty,\chi_n]}(Y_i) - P_{Y,\hat{\epsilon}}(Y_i \mid 1, X_i)\} = 0, \quad (6)$$

which is key to attaining (5). The updated estimator of p_Y is given by

$$p^u_{Y,n}(y \mid 1, x) = \hat{p}_{Y,\hat{\epsilon}}(y \mid 1, x).$$

4. *Iterate.* Let $p_{Y,n} = p^u_{Y,n}$ and iterate steps 2-3 until convergence, i.e., until $\hat{\epsilon} \approx 0$.

The TMLE of $\chi_0 = \chi(P_0)$ is denoted by $\chi_{n,\text{TMLE}}$ and is defined as χ_n in the last iteration. We also use $P_n^\star$ to denote the estimate of P_0 obtained in the last iteration. In the remaining of this section we discuss the asymptotic properties of the resulting estimator.

Note that, by construction,

$$\frac{1}{n} \sum_{i=1}^{n} P^\star_{Y,n}(\chi_{n,\text{TMLE}} \mid 1, X_i) = q.$$

This, together with (6) shows that (5) follows. Empirical process theory may now be used to prove (1), under consistency of the initial estimators of e and P_Y. Specifically, under the assumptions:

i) $D(P_n)$ converges to $D(P_0)$ in $L_2(P_0)$ norm,

ii) there exists a Donsker class $\mathcal{H}$ so that $D(P_n) \in \mathcal{H}$ with probability tending to one,

Theorem 19.24 of van der Vaart (2000) and the von Mises expansion of Lemma 2 show that

$$\chi_n - \chi = \frac{1}{n} \sum_{i=1}^{n} D(P_0) + R_2(P_n^\star, P_0),$$

where R_2 is defined in the lemma. Under the assumption that $R_2(P_n^\star, P_0) = o_P(1/\sqrt{n})$, asymptotic linearity follows. This asymptotic linearity result together with the central limit theorem may be used to construct confidence intervals and hypothesis tests. In particular, we have $\sqrt{n}(\chi_{n,\text{TMLE}} - \chi_0)$ has asymptotic distribution $N(0, \sigma^2)$, where $\sigma^2 = V(D(P_0)(Z))$. This variance may be estimated using a plug-in estimator.

3.1 Discussion of Consistency Assumptions

The most important assumption is perhaps the consistency assumption that $R_2(P_n^\star, P_0) = o_P(1/\sqrt{n})$, which may be broken down into two assumptions:

A.1 $R_2^{(1)}(P_n^\star, P_0) = \sqrt{n}\frac{1}{f(\chi)}P_0\left(\frac{e_0}{e} - 1\right)(h_{0,\chi} - h_\chi) = o_P(1)$

A.2 $\sqrt{n}(\chi_n^{\text{TMLE}} - \chi_0)^2 = o_P(1)$

Assumption A.1 is standard in the analysis of doubly robust estimators. Using the Cauchy-Schwarz inequality repeatedly, $|R_2^{(1)}(P, P_0)|$ may be bounded as

$$|R_2^{(1)}(\hat{P}, P_0)| \leq ||1/e_n||_\infty \, ||e_n - e_0||_{P_0} \, ||h_{n,\chi_n^{\text{TMLE}}} - h_{0,\chi_n^{\text{TMLE}}}||_{P_0},$$

where $||f||_P^2 := \int f^2(o)dP(o)$, and $||f||_\infty := \sup\{f(o) : o \in \mathcal{O}\}$. A set of sufficient conditions for A.1 to hold is, for example, $||e_n - e_0||_{P_0} = O_P(1/\sqrt{n})$ (e.g., e_0 is estimated in a correctly specified parametric model) and $||h_{n,\chi_n^{\text{TMLE}}} - h_{0,\chi_n^{\text{TMLE}}}||_{P_0} = o_P(1)$.

Assumption A.2 is stronger and suggests that consistency of the initial estimator $P_{Y,n}$ at rate $n^{-1/4}$ or faster is required, thus ruling out double robustness. However, this requirement may be avoided by making use of the fact that the propensity score is a balancing score (Rosenbaum and Rubin, 1983). Specifically, if only e_n may be assumed consistent at rate $n^{-1/4}$ or faster, then A.2 may be arranged by making sure that the initial estimator $P_{Y,n}$ includes a non-parametric component adjusting for e_n. For example, if $P_{Y,n}$ is a kernel-based estimator of $P_0(Y \leq y \mid M = 1, e_n(X))$ using the optimal bandwidth, then the convergence rate of χ_n^{TMLE} is at least $n^{-2/5}$, which would satisfy A.2.

The consistency rate of the initial estimators e_n and $P_{Y,n}$ determines the rate at which $R_2(P_n^\star, P_0)$ convergence to zero, and thus determine the consistency and asymptotic linearity of $\chi_{n,\text{TMLE}}$. When the number of covariates is large, the curse of dimensionality precludes the use of non-parametric estimators. In those scenarios, we advocate for the use flexible, data adaptive estimators to fit these quantities, so that the assumption $R_2(P_n^\star, P_0) \xrightarrow{P} 0$ remains plausible. One such an estimator may be constructed by proposing a library of candidate algorithms and selecting a convex combination of them, where the weights are chosen to minimize the cross-validated risk. This algorithm is discussed by van der Laan et al. (2007), and is implemented in the R library `SuperLearner`.

3.2 Estimating the Causal Effect on the Treated

In this subsection we discuss estimation of the causal effect of treatment on an outcome quantile among the treated. Specifically, let X denote a set of pre-treatment variables, let T denote a binary variable indicating the treatment group, and let Y denote the outcome of interest. We adopt the structural causal model (Pearl, 2009)

$$\begin{aligned} X &= g_X(U_X) \\ T &= g_T(X, U_T) \\ Y &= g_Y(T, X, U_Y), \end{aligned}$$

where g_X, g_T, and g_Y are unknown deterministic functions, and U_X, U_T, and U_Y represent exogenous unmeasured variables with unrestricted distributions. As in the previous section, we denote the true distribution of the observed data with P_0. We define the potential outcomes $Y_t := g_Y(t, X, U_Y) : t \in \{0, 1\}$ as the outcome that would have been observed if, contrary to the fact, $P(T = t) = 1$. We assume that (i) $T \perp\!\!\!\perp Y_1 \mid X$, and that (ii) $e(x) = P(T = 1 \mid X = x) < 1$ almost everywhere. Assumption (i) is often referred to as the *no unmeasured confounders* or *ignorability* assumption, and states that all factors that are simultaneous causes of T and Y must be measured. Assumption (ii) is referred to as the *positivity* assumption, and ensures that all units have a non-zero chance of falling in the control arm $T = 0$ so that there is enough experimentation. Let $F^{(t)}(y) := P(Y_t \leq y \mid T = 1)$ denote the distribution function of Y_t conditional on $T = 1$, then our target estimand is given by $\chi = \chi^{(1)} - \chi^{(0)}$, where

$$\chi^{(t)} = \inf\{y : F^{(t)}(y) \geq q\}.$$

That is, χ quantifies the causal effect of setting $T = 1$ vs $T = 0$ on the q-th quantile, restricted to treated units. Note that $Y_1 = Y$ on the event $T = 1$, so that $F^{(1)}(y) = P(Y \leq y \mid T = 1)$ and $\chi^{(1)}$ may be optimally estimated by the sample quantile of Y among treated units, which we denote with $\chi_n^{(1)}$. Thus, we focus on estimation of $\chi^{(0)}$. Under assumptions (i) and (ii) above, the distribution function $F^{(0)}$ identified as

$$F^{(0)}(y) = \sum_x P_Y(y \mid 0, x) p_X(x \mid 1),$$

where $P_Y(y \mid 0, x) := Pr(Y \leq y \mid T = 0, X = x)$ and $p_X(x \mid 1) := Pr(X = x \mid T = 1)$. The efficient influence function for estimation of $\chi^{(0)}$ in the non-parametric model may be found using similar techniques as in the previous section as

$$\begin{aligned} D^{(0)}(Z) = -\frac{1}{f^{(0)}(\chi^{(0)})} \Bigg[\frac{1-T}{E(T)} \frac{e(X)}{1-e(X)} \Big\{ I_{(-\infty, \chi^{(0)}]}(Y) - P_Y(\chi^{(0)} \mid 0, X) \Big\} \\ + \frac{T}{E(T)} \{ P_Y(\chi^{(0)} \mid 0, X) - q \} \Bigg], \quad (7) \end{aligned}$$

where $f^{(0)}$ is the probability density function associated to $F^{(0)}$. An expansion similar to that of Lemma 3 may be shown to hold, with a corresponding remainder term R_2 containing only second order terms.

The estimation algorithm involves the following steps:

1. *Initialize.* Obtain initial estimates e_n and $P_{Y,n}$ of e_0 and $P_{Y,0}$.

2. *Compute* $\chi_n^{(1)}$. For the current estimate $P_{Y,n}$, compute

$$F_n^{(0)}(y) = \frac{1}{\sum_i T_i} \sum_{i=1}^n T_i P_{Y,n}(y \mid 0, X_i),$$

   and $\chi_n^{(0)} = \inf\{y : F_n^{(0)}(y) \geq q\}$.

3. *Update* $P_{Y,n}$. Let $p_{Y,n}$ denote the density associated to $P_{Y,n}$, and consider the exponential model $p_{Y,\epsilon}(y \mid 0, x) = c(\epsilon, p_{Y,n}) \exp\{\epsilon D_{Y,n}(z)\} p_{Y,n}(y \mid 0, x)$, where $c(\epsilon, \hat{p}_Y)$ is a normalizing constant and

$$\hat{D}_{Y,n}(z) = \frac{e_n(X)}{1 - e_n(x)} \{I_{(-\infty, \chi_n^{(0)}]}(y) - P_{Y,n}(y \mid 0, x)\}$$

   is the score of the model. Estimate ϵ as $\hat{\epsilon} = \arg\max_\epsilon \sum_{i=1}^n (1 - T_i) \log p_{Y,\epsilon}(Y_i \mid 0, X_i)$. The updated estimator of p_Y is given by $p_{Y,n}^u(y \mid 0, x) = \hat{p}_{Y,\hat{\epsilon}}(y \mid 0, x)$.

4. *Iterate.* Let $\hat{p}_{Y,n} = \hat{p}_{Y,n}^u$ and iterate steps 2-3 until convergence, i.e., until $\hat{\epsilon} \approx 0$.

The TMLE of $\chi_0^{(0)}$ is denoted by $\chi_{n,\text{TMLE}}^{(0)}$ and is defined as the value of $\chi_n^{(0)}$ in the last iteration. The estimator of χ_0 is then defined as $\chi_{n,\text{TMLE}} = \chi_n^{(1)} - \chi_{n,\text{TMLE}}^{(0)}$. Arguments in the previous section may be used to show that, under analogous consistency and regularity conditions on the initial estimators $P_{Y,n}$ and e_n, the estimator satisfies

$$\sqrt{n}(\chi_{n,\text{TMLE}} - \chi) = \frac{1}{\sqrt{n}} \sum_{i=1}^n D(Z_i) + o_P(1),$$

with $D(Z) = D^{(1)}(Z) - D^{(0)}(Z)$, where

$$D^{(1)}(Z) = \frac{1}{f_Y(\chi^{(1)} \mid T = 1)} \frac{T}{E(T)} \{I_{(-\infty, \chi^{(1)}]}(Y) - q\}$$

is the influence function of the empirical quantile among the treated $\chi_n^{(1)}$. Thus, $\chi_{n,\text{TMLE}}$ has asymptotic distribution $N(\chi_0, \sigma^2/n)$, where $\sigma^2 = V(D(P_0)(Z))$. The latter variance may be estimated using a plug-in estimator to construct confidence intervals and hypothesis tests.

4 Simulation Studies

In this section we illustrate the properties of the proposed estimation algorithm using a real data set. We compare the performance of our proposed estimator to the performance of the one-step estimator in expression (4), and the inverse-probability weighted estimator proposed by Firpo (2007) defined as the minimum of

$$f(\chi^{(0)}) = \sum_{i=1}^{n} \frac{(1-T_i)e_n(X_i)}{1-e_n(X_i)}(Y_i - \chi^{(0)})(q - I_{(-\infty,\chi^{(0)}]}(Y_i)),$$

which we denote $\chi^{(0)}_{n,\text{IPW}}$, and compute using the `quantreg` R package (Koenker, 2013). We estimate the effect on the 25%, 50%, and 75% quantiles.

We also compare the performance of our proposed estimator $\chi_{n,\text{TMLE}}$ for the effect on the median among the treated versus the performance of an analogous estimator of the effect on the mean among the treated in terms of mean squared error and power for testing the null hypothesis of no treatment effect. As a measure of the causal effect, we use the effect on the quantiles among the treated, which is defined in the previous section. We use data from one of the AdWords programs at Google. Treatment consists of proactive consultations by sales representatives that help identify advertisers' business goals and suggest changes to improve performance. Since advertisers do not always adopt the proposed changes, a unit is considered treated if it is offered and accepts treatment. As a result, treatment is not randomized and we must use methods for observational data to assess the effect of such programs. We have standardized the outcome to a variable with mean 10 and standard deviation 5 before carrying out our analyses. These values are selected arbitrarily and do not reflect any particular feature of the data. Figure 1 shows the distribution of the logarithm of the standardized outcome, which can be seen to exhibit heavy tails and a large variability, even in the logarithmic scale. The original dataset consists of 40,303 units, with 29,362 being treated. To adjust for confounders of the relation between treatment and spend through AdWords, we use 93 variables containing baseline characteristics of the customer as well as activity on their AdWords account.

4.1 Estimator Performance

In order to compare the performance of the estimators, we simulated 1,000 datasets from the observed data as follows. First, we fit a logistic regression to estimate the probability of treatment conditional on the covariates. Second, we fit a linear quantile regression to the outcome separately for the control and the treated group, for 500 quantiles, using the `quantreg` R package (Koenker, 2013). We then generate a sample by first drawing covariates from the empirical distribution (i.e., sampling with replacement), and then using the parametric fits to draw a treatment indicator and an outcome. First, for each generated dataset, we estimated the effect of treatment on the 25%, 50%, and 75% quantiles. We approximate the bias, variance, mean squared error, coverage of a 95% Gaussian-based confidence interval using empirical means across the 1,000 simulated datasets. We compare the performance of the estimators in 3 scenarios using different initial estimators for e and P_Y:

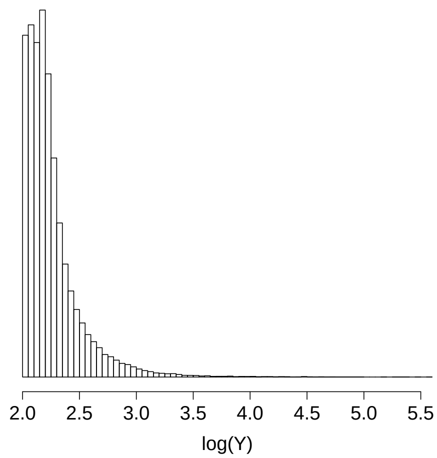

Figure 1: Histogram of the natural logarithm of the standardized outcome.

Scenario 1: Correctly specified parametrization for e and P_Y.

Scenario 2: Incorrectly specified parametrization for e but correctly specified for P_Y.

Scenario 3: Correctly specified parametrization for e but incorrectly specified for P_Y. According to the discussion about double robustness in Section 3.1, P_Y is estimated using a correctly specified parametrization for the distribution of the outcome conditional only on the propensity score.

If both parametrizations are misspecified, all estimators are expected to be inconsistent, with unknown asymptotic bias depending on the true data generating mechanism and the limit of the initial estimators e_n and $P_{Y,n}$. As a consequence, simulation results from that scenario would be of little practical guidance and the scenario is not considered simulations. Misspecification of the above parametrizations was performed by completely ignoring all covariates and using marginal distributions as estimators. This type of misspecification is highly unlikely to occur in practice and provides a worst case scenario. Estimation of $\hat{P}_Y$ is carried out by fitting a parametric quantile regression algorithm on m equally spaced quantiles using the R package `quantreg`. Then the initial density estimate $\hat{p}_Y$ has point mass $1/m$ at each of the initial quantiles, and the effect of the MLE in step 3 of Section 3 is to update the probability mass of each point. We simulated data using four sample sizes: 5,000, 10,000, 20,000 and 40,000. We compare the estimator performance in terms of percent bias, variance, MSE, and coverage probability of a 95% normal-based confidence interval. The results are presented in Table 1.

Results First of all, we note that the TMLE outperforms the one-step estimator in almost all situations, particularly for smaller sample sizes. We conjecture this is due to the fact that the TMLE is a plug-in estimator that falls within bounds of the observed outcome space and therefore has enhanced finite sample properties, compared to the inverse-probability weighting involved in the one-step estimator. In the worst case scenario, for $q = 0.25$ and $n = 5,000$, the TMLE is 10 times more efficient than the one-step estimator. As sample size increases, the performance of both estimators seems to be more similar.

q	Sc.	n	Rel. MSE			% Bias			Rel. Var.			Cov. Prob.		
			IPW	OS	TMLE	IPW	OS	TMLE	IPW	OS	TMLE	IPW	OS	TMLE
0.25	1	5,000	1.77	10.03	1.13	-3.24	9.89	7.28	1.76	9.91	1.06	0.95	0.95	0.95
		10,000	1.79	2.09	0.96	-1.20	3.11	2.74	1.79	2.07	0.94	0.97	0.96	0.96
		20,000	1.69	1.05	0.99	-1.37	1.90	1.23	1.68	1.03	0.98	0.96	0.94	0.95
		40,000	1.70	0.98	0.89	-0.89	1.01	0.17	1.69	0.97	0.89	0.95	0.95	0.96
	2	5,000	–	1.02	0.79	–	10.32	4.89	–	0.89	0.77	–	0.95	0.97
		10,000	–	1.00	0.78	–	5.89	1.90	–	0.92	0.77	–	0.94	0.97
		20,000	–	0.96	0.72	–	3.94	0.93	–	0.88	0.72	–	0.95	0.98
		40,000	–	0.97	0.76	–	2.63	0.53	–	0.90	0.76	–	0.95	0.97
	3	5,000	1.77	2.29	1.93	-3.24	-5.23	-5.14	1.76	2.26	1.89	0.95	0.98	0.98
		10,000	1.79	2.21	1.92	-1.20	-4.56	-3.40	1.79	2.16	1.89	0.97	0.99	0.99
		20,000	1.69	1.83	1.78	-1.37	-3.18	-1.86	1.68	1.78	1.76	0.96	0.98	0.98
		40,000	1.70	1.70	1.57	-0.89	-2.83	-0.85	1.69	1.62	1.56	0.95	0.98	0.98
0.5	1	5,000	2.52	2.30	1.18	-3.77	0.08	2.18	2.50	2.30	1.17	0.95	0.94	0.93
		10,000	2.18	3.25	1.16	-1.75	0.97	0.79	2.17	3.25	1.16	0.94	0.93	0.93
		20,000	2.00	0.98	0.98	-0.89	0.64	0.68	2.00	0.98	0.97	0.95	0.94	0.94
		40,000	2.03	0.98	0.95	0.19	0.70	0.49	2.03	0.97	0.95	0.96	0.95	0.94
	2	5,000	–	0.81	0.80	–	-0.75	-1.19	–	0.81	0.80	–	0.95	0.95
		10,000	–	0.91	0.85	–	0.60	-0.06	–	0.91	0.85	–	0.94	0.95
		20,000	–	0.80	0.80	–	0.02	-0.32	–	0.80	0.80	–	0.95	0.95
		40,000	–	0.83	0.80	–	0.39	0.23	–	0.83	0.80	–	0.95	0.95
	3	5,000	2.52	3.25	2.76	-3.77	-6.34	-7.99	2.50	3.20	2.67	0.95	0.98	0.98
		10,000	2.18	2.44	2.28	-1.75	-3.77	-4.20	2.17	2.41	2.24	0.94	0.97	0.97
		20,000	2.00	2.01	1.99	-0.89	-1.35	-1.16	2.00	2.00	1.99	0.95	0.98	0.98
		40,000	2.03	2.05	2.03	0.19	-1.54	-1.00	2.03	2.02	2.02	0.96	0.98	0.99
0.75	1	5,000	3.23	1.75	1.58	-2.26	-6.78	-1.51	3.23	1.70	1.58	0.95	0.93	0.86
		10,000	2.89	1.85	1.32	-0.27	-2.64	-0.56	2.89	1.84	1.32	0.94	0.94	0.91
		20,000	3.01	1.27	1.29	-2.07	-1.57	-0.19	3.00	1.26	1.29	0.95	0.93	0.91
		40,000	2.99	1.01	1.19	-0.44	-0.55	-0.14	2.99	1.01	1.19	0.96	0.93	0.90
	2	5,000	–	0.93	1.08	–	-11.46	-8.58	–	0.80	1.00	–	0.91	0.89
		10,000	–	0.93	1.09	–	-5.55	-3.93	–	0.87	1.06	–	0.92	0.88
		20,000	–	0.89	1.03	–	-2.22	-0.94	–	0.86	1.03	–	0.91	0.90
		40,000	–	0.89	1.04	–	-1.20	-0.75	–	0.88	1.03	–	0.92	0.89
	3	5,000	3.23	3.38	3.61	-2.26	7.88	-14.52	3.23	3.32	3.39	0.95	0.98	0.96
		10,000	2.89	2.69	2.95	-0.27	9.48	-6.40	2.89	2.50	2.86	0.94	0.97	0.97
		20,000	3.01	2.90	3.07	-2.07	9.08	-5.24	3.00	2.56	2.96	0.95	0.96	0.96
		40,000	2.99	3.48	2.94	-0.44	10.63	-2.26	2.99	2.55	2.90	0.96	0.94	0.96

Table 1: Simulation results for different scenarios for the initial estimators (Sc.) and sample sizes (n). % Bias is the bias relative to the true parameter value. Rel. Var. and Rel. MSE are the variance and MSE scaled by n relative to the non-parametric efficiency bound, respectively. Cov. Prob. is the coverage probability of a 95% confidence interval.

When the parametrizations for the initial estimators are correctly specified but the propensity score is incorrectly estimated by the sample proportion of treated units (scenario 2), the TMLE and OS estimators are sometimes super-efficient, having a relative MSE that is smaller than the efficiency bound. This is a particularity of this dataset and the initial estimators used, and should not be expected to hold in general. This super-efficiency arises because the misspecification of the treatment mechanism as the sample proportion makes the TMLE and one-step estimators asymptotically equivalent to the MLE in a correctly specified parametric model, which is often a more efficient estimator, though not doubly robust. Another possible explanation is that the asymptotic properties expected for the estimators have not yet taken effect for the finite sample sizes we analyzed, but would be observed in larger sample sizes not considered here for computational constraints. In any case, theoretical results show that this is not to be expected uniformly across data generating mechanisms and different estimators.

The bias of the IPW estimator when the propensity score is incorrectly specified (scenario 2) was always larger then 50% and often larger than 100%, therefore results for that scenario are not presented. This is expected since this estimator is only consistent under consistent estimation of the propensity score. When the outcome estimator is incorrect (scenario 3), the IPW has similar performance to the OS estimator and the TMLE. This is also expected since in this case the latter estimators are expected to behave asymptotically like the IPW. However, the efficiency gains obtained by using an outcome regression (scenario 1) are evident, with a mean square error up to 2.5 times smaller for the TMLE.

		% Bias		$\sqrt{n} \times$ MSE		Power	
Scenario	n	Mean	Median	Mean	Median	Mean	Median
1	5,000	-3.541	2.178	5.202	2.621	0.649	0.944
	10,000	-4.082	0.787	4.750	2.605	0.894	0.999
	20,000	-4.283	0.683	5.302	2.387	0.982	1.000
	40,000	-4.566	0.490	5.843	2.355	0.995	1.000
2	5,000	-1.601	-1.186	4.503	2.166	0.704	0.981
	10,000	-3.415	-0.064	4.165	2.224	0.922	1.000
	20,000	-4.411	-0.323	4.346	2.157	0.990	1.000
	40,000	-4.364	0.226	4.848	2.164	1.000	1.000
3	5,000	-180.361	-7.989	45.416	4.010	0.279	0.458
	10,000	-30.632	-4.203	15.514	3.650	0.162	0.848
	20,000	-30.990	-1.156	13.175	3.410	0.366	0.995
	40,000	-30.427	-1.001	17.052	3.443	0.571	1.000

Table 2: Simulation results comparing TMLE of the effect on the mean vs the effect on the median as a measure of the causal effect of treatment among the treated.

4.2 Comparison with the Effect on the Mean

In order to compare the effect on the median to the effect on the mean as a measure of the causal effect of treatment, we use the TMLE for the average treatment effect on the treated presented in Chapter 8 of van der Laan and Rose (2011). This estimator provides a fair comparison since it is also doubly robust and locally efficient in the non-parametric model. Table 2 contains a comparison between both estimators in terms of their percent bias, the squared root of the mean squared error scaled by n, and the power for testing the hypothesis of no treatment effect.

Note, in particular, the loss of power for the test based on the mean as compared to its median counterpart. In all scenarios, the hypothesis test based on the mean requires at least four times the sample size to achieve the power obtained with the test based on the median. This is very relevant in our setting since the sample size is not subject to modification but rather fixed by the number of AdWords customers available in a certain time period. In addition, note that the MSE of the estimator for the mean effect scaled by n seems to be increasing in scenario 1. This could be an indication that the effect on the mean is not estimable at a consistency rate of $\sqrt{n}$.

For all the sample sizes we considered, the estimator of the effect on the mean has a very large bias under scenario 3. This bias is due to inverse probability weighting extreme values of the outcome, and it vanishes as $n \to \infty$, as predicted by theory. When estimated with $n = 10^5$, this percent bias is equal to 11.06%, and it reduces to 0.38% when $n = 10^6$.

5 Concluding Remarks

We have proposed an estimator of the quantile function in missing data problems, with extensions to estimation of causal effects. Our estimator has properties not achieved by other estimators proposed in the literature; it is doubly robust, locally efficient, and asymptotically linear. Our double robustness result is not analogous to the standard double robustness of other estimators, in the sense that the initial estimator for the outcome regression may not be arbitrarily misspecified. Rather, double robustness in our setting means that it is always possible to construct an estimator that, though misspecified, yields a consistent estimator of the target parameter, under correct specification of the propensity score. The asymptotic properties of our estimator have been demonstrated analytically, and its finite sample superiority has been illustrated empirically using data arising from one of the applications that motivated the development of our methods.

References

Peter J. Bickel, Chris A.J. Klaassen, Ya'acov Ritov, and Jon A. Wellner. *Efficient and Adaptive Estimation for Semiparametric Models.* Springer-Verlag, 1997.

P.E. Cheng and C.K. Chu. Kernel estimation of distribution functions and quantiles with missing data. *Statistica Sinica*, 6:63–78, 1996.

Victor Chernozhukov, Iván Fernández-Val, and Blaise Melly. Inference on counterfactual distributions. *Econometrica*, 81(6):2205–2268, 2013.

Sergio Firpo. Efficient semiparametric estimation of quantile treatment effects. *Econometrica*, pages 259–276, 2007.

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

Zonghui Hu, Dean A Follmann, and Jing Qin. Dimension reduced kernel estimation for distribution function with incomplete data. *Journal of statistical planning and inference*, 141(9):3084–3093, 2011.

Roger Koenker. *quantreg: Quantile Regression*, 2013. URL `http://CRAN.R-project.org/package=quantreg`. R package version 5.05.

B. Ya Levit. On the efficiency of a class of non-parametric estimates. *Theory of Probability & Its Applications*, 20(4):723–740, 1976.

Xu Liu, Peixin Liu, and Yong Zhou. Distribution estimation with auxiliary information for missing data. *Journal of Statistical Planning and Inference*, 141(2):711–724, 2011.

Blaise Melly. Estimation of counterfactual distributions using quantile regression. *Review of Labor Economics*, 68(4):543–572, 2006.

Whitney K. Newey. Semiparametric efficiency bounds. *Journal of applied econometrics*, 5 (2):99–135, 1990.

Judea Pearl. *Causality: Models, Reasoning and Inference.* Cambridge University Press, New York, NY, USA, 2nd edition, 2009.

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

Mark J. van der Laan and Sherri Rose. *Targeted learning: causal inference for observational and experimental data*. Springer Science & Business Media, 2011.

Mark J. van der Laan and Daniel Rubin. Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2(1), 2006.

Mark J. van der Laan, Eric Polley, and Alan Hubbard. Super learner. *Statistical Applications in Genetics & Molecular Biology*, 6(25), 2007. ISSN 1.

Aad W. van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

Richard von Mises. On the asymptotic distribution of differentiable statistical functions. *The annals of mathematical statistics*, pages 309–348, 1947.

Qihua Wang and Yongsong Qin. Empirical likelihood confidence bands for distribution functions with missing responses. *Journal of Statistical Planning and Inference*, 140(9): 2778–2789, 2010.

Pu-Ying Zhao, Man-Lai Tang, and Nian-Sheng Tang. Robust estimation of distribution functions and quantiles with non-ignorable missing data. *Canadian Journal of Statistics*, 41(4):575–595, 2013.