Statistica Sinica Preprint No: SS-2025-0034	
Title	Assumption-Lean Quantile Regression
Manuscript ID	SS-2025-0034
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202025.0034
Complete List of Authors	Georgi Baklicharov, Christophe Ley, Vanessa Gorasso, Brecht Devleesschauwer and Stijn Vansteelandt
Corresponding Authors	Georgi Baklicharov
E-mails	georgi.baklicharov@ugent.be

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0034

Assumption-Lean Quantile Regression

Georgi Baklicharov[1], Christophe Ley[2], Vanessa Gorasso[3], Brecht Devleesschauwer[3] and Stijn Vansteelandt[1]

[1]*Ghent University,* [2]*University of Luxembourg and* [3]*Sciensano*

Abstract: Quantile regression is a powerful tool for detecting exposure-outcome associations given covariates across different parts of the outcome's distribution, but has two major limitations when the aim is to infer an exposure effect. Firstly, the exposure coefficient estimator may not converge to a meaningful quantity when the model is misspecified, and secondly, variable selection methods may induce bias and excess uncertainty, rendering inferences biased and overly optimistic. In this paper, we address these issues via partially linear quantile regression models which parametrize the conditional association of interest, but do not restrict the association with other covariates in the model. We propose consistent estimators for the unknown model parameter by mapping it onto a nonparametric main effect estimand that captures the (conditional) association of interest even when the quantile model is misspecified. This estimand is estimated using the efficient influence function under the nonparametric model, allowing for the incorporation of data-adaptive procedures such as variable selection and machine learning. Our approach provides a flexible and reliable method for detecting associations, robust to model misspecification and excess uncertainty induced by variable selection methods. The proposal is illustrated using simulation studies

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0034

and data on annual health care costs associated with excess body weight.

Key words and phrases: Debiased machine learning, Model misspecification, Non-parametric inference, Quantile treatment effect, Targeted learning.

1. Introduction

Quantile regression, originally introduced by Koenker and Bassett (1978), is a statistical method that provides a flexible framework for modeling the relationship between a response variable and predictors. Unlike traditional linear regression models that estimate the conditional mean of the response variable, quantile regression aims to estimate conditional quantiles, which can provide valuable and broader insights into the distributional characteristics of the response variable. It has gained popularity in recent years thanks to its ability to capture heterogeneity in the data and provide more robust estimates in the presence of outliers, skewed and/or heavy-tailed distributions (Koenker, 2005; Koenker et al., 2017). Quantile regression has been applied in many fields, including economics, finance, health, and environmental sciences, among others (e.g., Koenker and Hallock, 2001; Yu et al., 2003) but deserves to be used more frequently, because of the rich description it provides of various relevant aspects of data distributions, combined with appealing ease of interpretation (Das et al., 2019).

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0034

In this work, we focus on the estimation of causal quantile treatment or exposure effects under standard causal assumptions (Hernán and Robins, 2020). The quantile treatment effect has originally been defined by Doksum (1974) and Lehmann and D'Abrera (1975) in a two-sample setting as the horizontal distance (i.e., the distance along the horizontal axis between two fixed points of different distribution functions) between the outcome distribution of exposed versus unexposed. However, extension to allow for covariate adjustment via standard quantile regression raises concerns due to potential model misspecification and post-selection inference.

Model misspecification is a concern for most statistical methods. When the goal is solely prediction, this is often less problematic. Wrong models can still deliver good predictions, because predictions are commonly obtained by minimizing a measure of prediction error over the considered model class. Moreover, unbiased estimates of prediction error can be obtained, even when the model is misspecified. However, when estimating exposure effects, model misspecification typically causes more severe problems. For instance, in (generalized) linear regression models that have a main exposure effect and covariates, the ordinary least squares estimator generally fails to converge to a comprehensive measure of the conditional relationship between the exposure and outcome under model misspecifi-

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0034

cation (Vansteelandt and Dukes, 2022). In quantile regression models, the situation is worse as there is no good understanding of where even the effect of a randomized treatment, conditional on baseline covariates, converges to when the model is wrong. One possible solution is adding complexity to the model (e.g., by including interactions, higher-order terms, transformations). However, it is generally unclear how much additional complexity is needed. Moreover, adding complexity may be undesirable as it may make results obscure and difficult to communicate. In practice, one is therefore often torn between keeping the model simple so that it is comprehensible and making it complex so that it represents the ground truth (Breiman, 2001). A common way of finding a trade-off between model simplicity and model correctness is by using data-adaptive procedures, such as variable selection algorithms. However, this also raises concerns because data-adaptive procedures usually give rise to additional uncertainty that is not taken into account in standard error calculations. In particular, standard inference commonly ignores the fact that decisions on whether or not to add certain covariates were data-driven. This problem of post-selection inference renders standard inferences overly optimistic (Leeb and Pötscher, 2005).

In previous literature, attempts have been made to circumvent this issue by employing nonparametric approaches, focusing on marginal effects.

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0034

However, the majority of these approaches is limited to binary exposures (e.g., Melly, 2006; Firpo, 2007; Chernozhukov et al., 2013; Donald and Hsu, 2014; Caracciolo and Furno, 2017; Athey et al., 2023). While Ai et al. (2022) and Alejo et al. (2018) extend the analysis to continuous exposures, their focus remains on marginal quantile effects. Instead, our objective is to infer conditional quantile effects as these are less vulnerable to extrapolation (see the discussion section) and may give more refined insight into subject-specific effects.

Our work builds on the large literature on partially linear quantile models (e.g., Lee, 2003; Sun, 2005; Wu et al., 2010; Wu and Yu, 2014; Lv et al., 2015; Sherwood and Wang, 2016; Zhong and Wang, 2023). Traditionally, two approaches are being considered for estimating the nonparametric component of the partially linear quantile model: kernel estimators (e.g., Lee, 2003; Sun, 2005; Lv et al., 2015), and splines (e.g., He and Shi, 1996; Sherwood and Wang, 2016). Those approaches are also common in (partially linear) single-index quantile models (e.g., Wu et al., 2010; Wu and Yu, 2014). However, prior developments have a number of important shortcomings. First, in high-dimensional applications, they pose high computational demands and difficulties in handling high-dimensional covariates stemming from reliance on kernel weighting or splines. To accommodate

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0034

this, Zhong and Wang (2023) proposed neural networks to estimate the nonparametric component. However, a further concern with this and the aforementioned proposals is that the estimates have no straightforward interpretation when the partially linear quantile model is misspecified (e.g., due to exposure-covariate interactions). In that case, existing proposals also fail to provide valid data-adaptive inference (not even for the probability limit of the estimator). Our proposal provides a solution for both shortcomings by allowing the use of flexible estimators that are known to work well in high dimensions, and by ensuring an inference that is correct even under model misspecification. We take inspiration from related work for generalized linear models in a discussion paper by Vansteelandt and Dukes (2022) and for Cox models (Vansteelandt et al., 2022), which are both closely related to recent developments on Targeted Maximum Likelihood Estimation (TMLE) (van der Laan and Rose, 2011) and Debiased Machine Learning (DML) (Chernozhukov et al., 2018).

In particular, throughout this work, we will focus on the conditional association between an outcome Y and a possibly continuous exposure or treatment A, given a vector of measured covariates denoted by L that are sufficient to control for confounding. Therefore, we consider the following

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0034

conditional working model

$$Q_\tau(Y^{A+a}|A,L) - Q_\tau(Y^A|A,L) = \beta_\tau(L)a \quad \text{for all } a, \tag{1.1}$$

where $Q_\tau(Y^a|L)$ denotes the τ-quantile of the conditional distribution of the counterfactual Y^a (i.e., the outcome that would have been seen, if the individual would have had exposure value $A = a$) given L for a given value of $0 < \tau < 1$, and $\beta_\tau(.)$ is an unknown, quantile-specific function. In this model, $\beta_\tau(L)$ reflects the quantile effect of a unit shift in exposure. Under consistency (i.e., $Y^a = Y$ if $A = a$) and conditional exchangeability (i.e., when A is independent of the counterfactual outcome Y^a to exposure level a given L, $Y^a \perp\!\!\!\perp A|L$), (1.1) is equivalent to

$$Q_\tau(Y|A,L) = \beta_\tau(L)A + Q_\tau(Y|A=0,L). \tag{1.2}$$

This reduces to a partially linear quantile model when $\beta_\tau(L) = \beta_\tau$ is constant. We avoid this assumption by developing debiased machine learning estimators for summary statistics of $\beta_\tau(L)$, which reduce to the parameter β_τ indexing the partially linear quantile model whenever the model holds.

2. Summary measures of $\beta_\tau(L)$

First, assume A is dichotomous. In that case, $\beta_\tau(L)$ equals the causal quantile exposure effect $Q_\tau(Y^1|L) - Q_\tau(Y^0|L)$. One way to summarize

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0034

this quantity across levels of L is via its mean and variance. Alternatively, one may consider a weighted average $E\left[w(L)\{Q_\tau(Y^1|L) - Q_\tau(Y^0|L)\}\right]$, and weighted variance with propensity overlap weights $w(L) = \pi(L)\{1-\pi(L)\}/E[\pi(L)\{1-\pi(L)\}]$, where $\pi(L) = P(A = 1|L)$. This choice of weights is motivated as follows. Since interest lies in the causal effect of A given L, one may wish to upweight those strata where $\text{Var}(A|L) = \pi(L)\{1 - \pi(L)\}$ is relatively large, emphasizing individuals for whom there is most uncertainty about the choice of A. Strata of L where all subjects are (un)exposed will not contribute to the estimand as they do not contain information about the causal effect. In the special case of a randomized treatment A, the weights equal one.

For general A, we first introduce the following loss function

$$E\left[\int \left\{Q_\tau(Y^{A+a}|A, L) - Q_\tau(Y^A|A, L) - \Psi_\tau a\right\}^2 f(A + a|L)da\right], \quad (2.1)$$

which is a translation of the loss considered by Vansteelandt (2025) to quantile regression. Minimizing (2.1) selects the best value of Ψ_τ for which $\Psi_\tau a$ best approximates the real quantile effect $Q_\tau(Y^{A+a}|A, L) - Q_\tau(Y^A|A, L)$ of an a unit shift in exposure, averaging over shifts a of magnitudes at which exposure level $A + a$ is likely for the considered subgroup L. This averaging prevents unrealistic extrapolations. For binary A, the minimizer recovers the propensity overlap weighted average of $Q_\tau(Y^1|L) - Q_\tau(Y^0|L)$.

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0034

For general A, it equals the weighted average $E\{w(L)\Psi_\tau(L)\}$, with $w(L) = \text{Var}(A|L)/E\{\text{Var}(A|L)\}$ and where $\Psi_\tau(L) = \text{Cov}\{A, Q_\tau(Y|A,L)|L\}\text{Var}(A|L)^{-1}$ is the population least squares projections of $Q_\tau(Y|A,L)$ onto A, given L which equals $\beta_\tau(L)$ under model (1.2) (see the online appendix for a proof). Thus, the minimizer of (2.1) equals

$$\Psi_\tau = \frac{E\left(\{A - E(A|L)\}\left[Q_\tau(Y|A,L) - E\{Q_\tau(Y|A,L)|L\}\right]\right)}{E[\{A - E(A|L)\}^2]}. \tag{2.2}$$

Alternatively, one can focus on weighted averages with other weights, or on weighted variances $\Phi_\tau = E\left[w(L)\left\{\Psi_\tau(L) - \Psi_\tau\right\}^2\right]$ of $\Psi_\tau(L)$.

The main focus of this paper is estimation and inference for Ψ_τ. If $\beta_\tau(L) = \beta_\tau$ in (1.2) were independent of L, then model (1.2) reduces to

$$Q_\tau(Y|A,L) = \beta_\tau A + Q_\tau(Y|A=0,L), \tag{2.3}$$

and $\Psi_\tau = \beta_\tau$ and $\Phi_\tau = 0$ (see the online appendix for a proof). This suggests that Ψ_τ has a standard interpretation in model (2.3), but remains well interpretable as summary statistic or loss minimizer (2.2) when this model is misspecified. Estimators for the summary statistics with weights equal to 1 are provided in the online appendix.

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0034

3. Data-adaptive inference

In this section, we will develop nonparametric inference for the nonparametric estimand Ψ_τ. We first propose a plug-in estimator and caution for its potential flaws. Next, we explain how to resolve these by making use of the estimand's efficient influence function (EIF) for debiasing. Then, we describe an asymptotically equivalent estimator based on Targeted Learning. Finally, we present simplified versions of the proposed estimators involving parametric quantile regression with variable selection.

3.1 A plug-in estimator

A so-called plug-in estimator is straightforwardly obtained by substituting $E(A|L)$, $Q_\tau(Y|A,L)$ and $E\{Q_\tau(Y|A,L)|L\}$ by consistent estimators $\hat{E}(A|L)$, $\hat{Q}_\tau(Y|A,L)$ and $\hat{E}\{\hat{Q}_\tau(Y|A,L)|L\}$, respectively, and by replacing population expectations by sample averages. This results in

$$\frac{\sum_{i=1}^n \left(\{A_i - \hat{E}(A_i|L_i)\}\left[\hat{Q}_\tau(Y_i|A_i,L_i) - \hat{E}\{\hat{Q}_\tau(Y_i|A_i,L_i)|L_i\}\right]\right)}{\sum_{i=1}^n \{A_i - \hat{E}(A_i|L_i)\}^2}, \quad (3.1)$$

with nuisance parameters estimated using flexible data-adaptive procedures such as variable selection techniques for quantile regression or more advanced algorithms such as random forests or neural networks.

However, plug-in estimators like (3.1) typically give rise to two prob-

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0034

lems. First, data-adaptive procedures are optimized towards minimal prediction error, but prediction is not our aim. Instead, our aim is to obtain well-behaved estimators for this estimand. For example, a poor selection of variables may induce confounding or selection bias. Second, data-adaptive estimators $\hat{E}(A|L)$, $\hat{Q}_\tau(Y|A,L)$ and $\hat{E}\{\hat{Q}_\tau(Y|A,L)|L\}$ have a nonstandard and usually unknown distribution and unknown convergence rates, making uncertainty quantification extremely difficult. Our proposal addresses these concerns.

3.2 Debiased Machine Learning (DML)

In view of the aforementioned concerns, we aim to establish inference for main effect estimand (2.2) under the nonparametric model. This enables us to avoid dependence on modeling constraints, which could result in excessively optimistic inferences as well as computational difficulties. Like other debiased machine learning strategies, we will make use of the EIF of the estimand, given in Theorem 1 (with proof in the online appendix), to accomplish this goal (Pfanzagl, 1990; Bickel et al., 1993).

Theorem 1. *The efficient influence function of the main effect estimand*

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0034

Ψ_τ, defined by (2.2), under the nonparametric model is

$$\phi(Y,A,L) = \frac{A - E(A|L)}{E[\{A - E(A|L)\}^2]}\Big[Q_\tau(Y|A,L) - E\left\{Q_\tau(Y|A,L)|L\right\} + \frac{\tau - I\{Y \le Q_\tau(Y|A,L)\}}{f\{Q_\tau(Y|A,L)|A,L\}} - \Psi_\tau\{A - E(A|L)\}\Big], \tag{3.2}$$

where $I(.)$ and $f(.|A,L)$ denote an indicator function and the conditional density of Y given A and L, respectively.

If the infinite-dimensional nuisance parameters $E(A|L)$, $Q_\tau(Y|A,L)$, $E\left\{Q_\tau(Y|A,L)|L\right\}$ and $f\{Q_\tau(Y|A,L)|A,L\}$ indexing the EIF were known, then it would follow from its mean zero property that a consistent estimator $\tilde{\Psi}_\tau$ of Ψ_τ could be obtained as the value of Ψ_τ that makes the sample average of the influence functions zero. From standard results on M-estimation, the resulting estimator's asymptotic distribution would be governed by this influence function in the sense that $\sqrt{n}(\tilde{\Psi}_\tau - \Psi_\tau)$ equals

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{A_i - E(A_i|L_i)}{E[\{A_i - E(A_i|L_i)\}^2]}\Big[Q_\tau(Y_i|A_i,L_i) - E\left\{Q_\tau(Y_i|A_i,L_i)|L_i\right\} + \frac{\tau - I\{Y_i \le Q_\tau(Y_i|A_i,L_i)\}}{f\{Q_\tau(Y_i|A_i,L_i)|A_i,L_i\}} - \Psi_\tau\{A_i - E(A_i|L_i)\}\Big] + o_p(1). \tag{3.3}$$

It follows that the estimator $\tilde{\Psi}_\tau$ is asymptotically normal by the central limit theorem. Moreover, its bias approaches zero at a faster rate than the standard error, and its variance can be estimated by computing the sample variance of the EIF. However, the estimator $\tilde{\Psi}_\tau$ is practically infeasible because the EIF involves unknown nuisance parameters. We will therefore

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0034

substitute these with consistent data-adaptive estimators and denote the resulting estimator $\hat{\Psi}_\tau$. In this way, we obtain the closed-form estimator

$$\hat{\Psi}_\tau = \frac{1}{n}\sum_{i=1}^{n} \frac{A_i - \hat{E}(A_i|L_i)}{\frac{1}{n}\sum_{i=1}^{n}\{A_i - \hat{E}(A_i|L_i)\}^2}\Bigg[\hat{Q}_\tau(Y_i|A_i, L_i) - \hat{E}\{\hat{Q}_\tau(Y_i|A_i, L_i)|L_i\} + \frac{\tau - I\{Y_i \leq \hat{Q}_\tau(Y_i|A_i, L_i)\}}{\hat{f}\{\hat{Q}_\tau(Y_i|A_i, L_i)|A_i, L_i\}}\Bigg]. \tag{3.4}$$

From the general theory on nonparametric estimation (Pfanzagl, 1990; Bickel et al., 1993) and the results in the online appendix, it follows that, when the nuisance parameters are estimated on a sample independent to the one on which $\hat{\Psi}_\tau$ is evaluated (to mitigate overfitting bias) and under specific convergence rate assumptions detailed in Theorem 2, the estimator is asymptotically linear in the sense that (3.3) holds for $\hat{\Psi}_\tau$ instead of $\tilde{\Psi}_\tau$. Consequently, its variance can be consistently estimated as $1/n^2 \sum_{i=1}^{n} \hat{\phi}(O_i)^2$, where $\hat{\phi}$ is defined like ϕ but with all nuisance parameters substituted by consistent estimates.

Estimating the nuisance parameters on an independent sample can be attained via sample splitting (Zheng and van der Laan, 2011) or cross-fitting (Chernozhukov et al., 2018). In particular, the data can be split into K disjoint subsets I_k, $k = 1, \ldots, K$. Let I_k^C denote the complement of the subset I_k. For each k, one can train data-adaptive nuisance parameter estimators using only the data in I_k^C. In calculating $\hat{\Psi}_\tau$, when a subject i

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0034

belongs to subset I_k, then the nuisance parameter estimators trained on I_k^C should be used to plug in the nuisance parameter estimates in (3.4). When flexible data-adaptive estimates are used for the nuisance parameters, then this cross-fitting is necessary for the asymptotic validity of the proposal.

The advantage of using the EIF as the foundation for the estimator is that it exhibits consistent asymptotic behavior, regardless of whether it is based on the known nuisance parameters or consistent estimates thereof. This is contingent on the convergence rate conditions detailed in the following theorem. A proof is given in the online appendix. Let $\mathbb{P}_n$ denote the empirical measure (sample average). For a function $f(O)$ of the data O, we write $\mathbb{P}\{f(O)\} = \int f(O)\mathbb{P}(O)dO$, where $\mathbb{P}(O)$ is the data density. Note: $\mathbb{P}\{\hat{f}(O)\}$ averages over O, not over the estimator $\hat{f}$.

Theorem 2. *The estimator $\hat{\Psi}_\tau$ is asymptotically linear as in (3.3) when*

1. *all nuisance parameters are estimated on a sample independent to the one on which $\hat{\Psi}_\tau$ is evaluated (e.g., using cross-fitting),*

2. *$A - \hat{E}(A|L) = O_p(1)$ and the following terms are $o_p(1)$:*

$$\mathbb{P}\left(\left[\hat{f}\{\hat{Q}_\tau(Y|A,L)|A,L\} - f\{Q_\tau(Y|A,L)|A,L\}\right]^2\right),$$

$$\mathbb{P}\left(\left[E(Q_\tau(Y|A,L)|L) - \hat{E}\{\hat{Q}_\tau(Y|A,L)|L\}\right]^2\right),$$

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0034

3. *all of the following terms are* $o_p(n^{-1/2})$*:*

$$\mathbb{P}\left[\{\hat{Q}_\tau(Y|A,L) - Q_\tau(Y|A,L)\}^2\right], \tag{3.5}$$

$$\mathbb{P}\left(\left[1 - \frac{f\{Q_\tau(Y|A,L)|A,L\}}{\hat{f}\{\hat{Q}_\tau(Y|A,L)|A,L\}}\right]^2\right)^{1/2} \mathbb{P}\left[\{\hat{Q}_\tau(Y|A,L) - Q_\tau(Y|A,L)\}^2\right]^{1/2},$$

$$\mathbb{P}\left[\{E(A|L) - \hat{E}(A|L)\}^2\right]^{1/2} \mathbb{P}\left(\left[E(Q_\tau(Y|A,L)|L) - \hat{E}\{\hat{Q}_\tau(Y|A,L)|L\}\right]^2\right)^{1/2},$$

$$\mathbb{P}\left[\{E(A|L) - \hat{E}(A|L)\}^2\right],$$

where the last one only needs to be $o_p(1)$ *when* $\Psi_\tau = 0$*,*

4. *and* $\mathbb{P}[\{A - E(A|L)\}^2] > \epsilon$*,* $\mathbb{P}_n[\{A - \hat{E}(A|L)\}^2] > \epsilon$ *and* $\hat{f}\{\hat{Q}_\tau(Y|A,L)|A,L\} > \epsilon$ *with probability 1 for some* $\epsilon > 0$*.*

The rate conditions stated in Theorem 2 will be satisfied, e.g., if all nuisance parameters are consistently estimated at a rate faster than $n^{1/4}$, but are slightly less restrictive because of the rate conditions on product terms. For instance, the third condition in Theorem 2 states that (3.5) needs to be $o_p(n^{-1/2})$. However, if a faster convergence rate is achieved, then a slower convergence rate for $\hat{f}\{\hat{Q}_\tau(Y|A,L)|A,L\}$ may suffice, as long as the product below (3.5) converges sufficiently fast. Similarly, a much faster estimation rate for $E(A|L)$ (e.g., in a randomized experiment) allows for a slower estimation rate for $E\{Q_\tau(Y|A,L)|L\}$. Under sufficient smoothness/sparsity assumptions, many data-adaptive methods can achieve these

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0034

convergence rates (a summary is given in Chernozhukov et al. (2018)).

The proposed estimator thus mitigates the plug-in bias that arises from the use of flexible data-adaptive techniques. The bias-variance tradeoff made by these techniques is usually fine-tuned towards prediction error, leading to plug-in bias. This bias, approximately equal to minus the sample average of the influence functions (van der Vaart, 1998), can be removed by choosing the estimator (3.4) as the value of Ψ_τ that makes this sample average zero, ensuring approximately zero bias by construction. Our proposal further takes into account the excess uncertainty induced by the use of data-adaptive procedures (e.g., variable selection or machine learning algorithms). By removing the bias induced by these procedures, our estimator behaves asymptotically like an oracle estimator based on the known nuisance parameter values. Thus, the proposed standard errors accurately reflect the uncertainty that arises from these data-driven model choices.

3.3 Targeted Learning

Like other debiased machine learning strategies, targeted learning also removes plug-in bias based on the EIF, but it does so by 'targeting' the initial nuisance parameter estimates (van der Laan and Rubin, 2006; van der Laan and Rose, 2011). It thereby delivers an estimator that is asymptotically

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0034

equivalent to the earlier defined debiased estimator under the conditions of Theorem 2, though it may have different finite sample performance (Yadlowsky, 2022). Especially the last term in (3.4) can be unstable because of the density in the denominator. We therefore define a retargeted estimator $\widetilde{Q}_\tau(Y|A,L)$ which forces

$$\frac{1}{n}\sum_{i=1}^{n}\left\{A_i - \hat{E}(A_i|L_i)\right\}\left[\frac{\tau - I\{Y_i \le \widetilde{Q}_\tau(Y_i|A_i,L_i)\}}{\hat{f}\{\widetilde{Q}_\tau(Y_i|A_i,L_i)|A_i,L_i\}}\right] = 0. \tag{3.6}$$

This is justified because (3.6) is an unbiased estimating equation for $Q_\tau(Y|A,L)$. If (3.6) holds and $\hat{Q}_\tau$ is replaced by $\widetilde{Q}_\tau$ in (3.4), the resulting estimator for Ψ_τ equals

$$\frac{1}{n}\sum_{i=1}^{n}\frac{A_i - \hat{E}(A_i|L_i)}{\frac{1}{n}\sum_{i=1}^{n}\{A_i - \hat{E}(A_i|L_i)\}^2}\left[\widetilde{Q}_\tau(Y_i|A_i,L_i) - \hat{E}\{\widetilde{Q}_\tau(Y_i|A_i,L_i)|L_i\}\right],$$

which is a plug-in estimator of estimand (2.2).

Constructing an estimator that satisfies (3.6) turns out to be rather challenging. For convenience, we will write

$$w(Q_{\tau,i}) = \frac{A_i - \hat{E}(A_i|L_i)}{\hat{f}\{Q_\tau(Y_i|A_i,L_i)|A_i,L_i\}}. \tag{3.7}$$

Using the weights defined in (3.7), we can write equation (3.6) as

$$\frac{1}{n}\sum_{i=1}^{n} w(\widetilde{Q}_{\tau,i})\left[\tau - I\{Y_i \le \widetilde{Q}_\tau(Y_i|A_i,L_i)\}\right] = 0.$$

An approximated solution to this highly nonlinear equation for $\widetilde{Q}_\tau(Y_i|A_i,L_i)$ can be found by updating the initial estimator $\hat{Q}_\tau(Y_i|A_i,L_i)$ (e.g., obtained

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0034

Algorithm 1 Iterative Targeting of $\hat{Q}_\tau(Y_i|A_i, L_i)$

1: Initialize $Q^{(0)}_{\tau,i} = \hat{Q}_\tau(Y_i \mid A_i, L_i)$

2: **for** $k = 1, 2, \ldots$ **do**

3: Compute ϵ_{k-1} defined by:

$$\epsilon_{k-1} = \arg\min_{\epsilon} \left| \frac{1}{n} \sum_{i=1}^{n} w\left(Q^{(k-1)}_{\tau,i}\right) \left[\tau - I\left\{Y_i \le Q^{(k-1)}_{\tau,i} + \epsilon w\left(Q^{(k-1)}_{\tau,i}\right)\right\}\right] \right|.$$

4: Update: $Q^{(k)}_{\tau,i} = Q^{(k-1)}_{\tau,i} + \epsilon_{k-1} w\left(Q^{(k-1)}_{\tau,i}\right)$

5: **if** the quantity

$$\left| \frac{1}{n} \sum_{i=1}^{n} w(Q^{(k-1)}_{\tau,i}) \left[\tau - I\{Y_i \le Q^{(k-1)}_{\tau,i} + \epsilon\, w(Q^{(k-1)}_{\tau,i})\}\right] \right|$$

is no longer decreasing **then**

6: **break**

7: **end if**

8: **end for**

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0034

by using quantile regression forests) as explained in Algorithm 1. The resulting estimator for $Q_\tau(Y_i|A_i, L_i)$ is denoted $\widetilde{Q}_\tau(Y_i|A_i, L_i)$. Substituting $\hat{Q}_\tau(Y_i|A_i, L_i)$ with $\widetilde{Q}_\tau(Y_i|A_i, L_i)$ in the estimator (3.4) yields a Targeted Maximum Likelihood Estimation (TMLE) estimator

$$\hat{\Psi}_\tau^{TMLE} = \frac{1}{n}\sum_{i=1}^{n} \frac{A_i - \hat{E}(A_i|L_i)}{\frac{1}{n}\sum_{i=1}^{n}\{A_i - \hat{E}(A_i|L_i)\}^2} \Big[\widetilde{Q}_\tau(Y_i|A_i, L_i) - \hat{E}\{\widetilde{Q}_\tau(Y_i|A_i, L_i)|L_i\} + \frac{\tau - I\{Y_i \le \widetilde{Q}_\tau(Y_i|A_i, L_i)\}}{\hat{f}\{\widetilde{Q}_\tau(Y_i|A_i, L_i)|A_i, L_i\}}\Big]. \tag{3.8}$$

The targeting step results in the last term of (3.8) being approximately zero, such that $\hat{\Psi}_\tau^{TMLE}$ is approximately equal to

$$\sum_{i=1}^{n} \frac{A_i - \hat{E}(A_i|L_i)}{\sum_{i=1}^{n}\{A_i - \hat{E}(A_i|L_i)\}^2} \Big[\widetilde{Q}_\tau(Y_i|A_i, L_i) - \hat{E}\{\widetilde{Q}_\tau(Y_i|A_i, L_i)|L_i\}\Big]. \tag{3.9}$$

Comparing (3.9) with (3.1), it is clear that we removed (approximately) all bias from the initial plug-in estimator by cleverly targeting the nuisance parameter estimator for $Q_\tau(Y_i|A_i, L_i)$. In the online appendix, we give a sketch of a proof that (3.8) and (3.9) are asymptotically equivalent, even with just one iteration of targeting. Nevertheless, we opt to use (3.8) to mitigate reliance on the accuracy of approximation (3.9).

The targeted learning estimator with 5-fold cross-fitting can be obtained as follows. After splitting the data into 5 equal parts I_k ($k = 1, \ldots, 5$), five models are trained for each nuisance parameter, where each time one subset I_k is left out from the data. In particular, a nonparametric model is trained

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0034

for $E(A|L)$, $Q_\tau(Y|A,L)$, $E\{Q_\tau(Y|A,L)|L\}$ and $f\{Q_\tau(Y|A,L)|A,L\}$ on I_k^C (i.e., 80% of the data). Then, those models are used to estimate the nuisance parameters for subjects belonging to I_k. These estimates are then used to calculate the contribution to Ψ_τ (which is a sample average) for the same subjects. The targeting algorithm is then applied once to the entire dataset using the pooled cross-fitted estimates of the nuisance parameters. Then, Ψ_τ is estimated as in (3.8). The standard error can be estimated similarly to the debiased estimator (3.4), i.e., as the square root of

$$\frac{1}{n^2}\sum_{i=1}^{n}\left(\frac{A_i-\hat{E}(A_i|L_i)}{\frac{1}{n}\sum_{i=1}^{n}\{A_i-\hat{E}(A_i|L_i)\}^2}\left[\widetilde{Q}_\tau(Y_i|A_i,L_i)-\hat{E}\{\widetilde{Q}_\tau(Y_i|A_i,L_i)|L_i\}\right.\right.$$
$$\left.\left.+\frac{\tau-I\{Y_i\le\widetilde{Q}_\tau(Y_i|A_i,L_i)\}}{\hat{f}\{\widetilde{Q}_\tau(Y_i|A_i,L_i)|A_i,L_i\}}-\hat{\Psi}_\tau^{TMLE}\{A_i-E(A_i|L_i)\}\right]\right)^2.$$

3.4 Parametric quantile regression with variable selection

When using quantile regression, possibly with (stepwise) variable selection, to estimate the nuisance parameter $Q_\tau(Y|A,L)$, the debiased estimator (3.4) and TMLE estimator (3.8) can be further simplified. In particular, suppose we model $Q_\tau(Y|A,L)$ as $\beta_\tau A+\alpha^T L^*$, with L^* a subset of variables of L. This is a working model rather than an assumption on the true data-generating process, which would otherwise change the EIF of Ψ_τ. Under

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0034

this assumption, the EIF of Ψ_τ simplifies to

$$\frac{A-E(A|L)}{E[\{A-E(A|L)\}^2]}\left[\beta_\tau\{A-E(A|L)\}+\frac{\tau-I\{Y\leq Q_\tau(Y|A,L)\}}{f\{Q_\tau(Y|A,L)|A,L\}}-\Psi_\tau\{A-E(A|L)\}\right].$$

In that case we can estimate Ψ_τ as

$$\hat{\beta}_\tau+\frac{1}{n}\sum_{i=1}^{n}\frac{A_i-\hat{E}(A_i|L_i)}{\frac{1}{n}\sum_{i=1}^{n}\{A_i-\hat{E}(A_i|L_i)\}^2}\left[\frac{\tau-I\{Y_i\leq\hat{Q}_\tau(Y_i|A_i,L_i)\}}{\hat{f}\{\hat{Q}_\tau(Y_i|A_i,L_i)|A_i,L_i\}}\right]. \quad (3.10)$$

Here, $\hat{\beta}_\tau$ is the estimate of β_τ from fitting a quantile regression model of Y on A and L via standard software (e.g., the R package 'quantreg'), possibly with variable selection. The second term eliminates bias due to variable selection errors, which is needed to achieve valid post-selection inference.

An asymptotically equivalent (one-step) TMLE estimator is

$$\hat{\beta}+\frac{1}{n}\sum_{i=1}^{n}\frac{A_i-\hat{E}(A_i|L_i)}{\frac{1}{n}\sum_{i=1}^{n}\{A_i-\hat{E}(A_i|L_i)\}^2}\left[\left(\frac{A_i-\hat{E}(A_i|L_i)}{\hat{f}\{\widetilde{Q}_\tau(Y_i|A_i,L_i)|A_i,L_i\}}-\hat{E}\left[\frac{A_i-\hat{E}(A_i|L_i)}{\hat{f}\{\widetilde{Q}_\tau(Y_i|A_i,L_i)|A_i,L_i\}}|L_i\right]\right)\hat{\epsilon}+\frac{\tau-I\{Y_i\leq\widetilde{Q}_\tau(Y_i|A_i,L_i)\}}{\hat{f}\{\widetilde{Q}_\tau(Y_i|A_i,L_i)|A_i,L_i\}}\right], \quad (3.11)$$

where the last term is approximately zero. More details on the construction of this estimator can be found in the online appendix.

4. Simulation experiments

To evaluate the performance of the proposed estimators for Ψ_τ, we use four simulation experiments. In all experiments, we report Monte Carlo bias

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0034

and standard deviation (SD), as well as the averaged estimated standard error (SE) and coverage of corresponding 95% Wald confidence intervals. For all experiments, we show results for 1000 simulation runs for a correctly specified quantile regression estimator (CSQR), obtained by fitting a correctly specified standard quantile regression model that includes the exposure A and all necessary confounders, the naive plug-in (Plug-in) estimator (3.1), the DML estimator (3.4) based on quantile random forests without cross-fitting (DML) and with 5-fold cross-fitting (DML-CF), and the targeted learning estimator (3.8) without cross-fitting (TMLE) and with 5-fold cross-fitting (TMLE-CF). The sample size is fixed at $n = 500$ unless otherwise stated. The standard error of the plug-in estimator is always calculated similarly as for the debiased machine learning estimator, but removing the last term of (3.4) from the EIF. Experiment-specific details are provided below, and data-generating mechanisms and most of the tables with results are given in the online appendix. Code for the proposed estimators and reproducing the simulation studies is available on `https://github.com/gbaklicharov/ALQRsimulations`.

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0034

4.1 Experiment 1: Homoscedasticity, heteroscedasticity, and continuity

This experiment consisted of three simulation settings, where the data generating mechanism was each time slightly changed. In setting 1, we studied the proposed estimators' behaviour for a binary exposure and a homoscedastic outcome where the exposure effect had the same magnitude for all quantiles. Setting 2 is the same, except that the outcome variability differed for both exposure groups. This implies that the exposure effect is different for different values of τ. In setting 3, the exposure was continuous. The outcome was each time generated according to $Y = 1 + A + \sin(L_1) + L_2^2 + L_3 + L_4 + L_3L_4 + \epsilon$, with ϵ gamma-distributed. The results for CSQR, Plug-in and our best performing estimator TMLE-CF are shown in Table 1. The results for the other estimators and for different quantiles are included in the online appendix.

The simulations all demonstrate the poor performance of the naive plug-in estimator. It has a high bias and the standard error is poorly estimated and overly optimistic, leading to a coverage of almost 0. When comparing the debiased machine learning and targeted learning estimators with versus without cross-fitting, the advantage of cross-fitting becomes clearly visible. In general, the estimators with cross-fitting perform better,

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0034

4.1 Experiment 1: Homoscedasticity, heteroscedasticity, and continuity

Estimator	Setting 1				Setting 2				Setting 3			
	Bias	**SD**	**SE**	**Cov**	**Bias**	**SD**	**SE**	**Cov**	**Bias**	**SD**	**SE**	**Cov**
CSQR	-0.17	19	20	96.6	-0.59	26	26	93.7	-0.13	3.5	3.6	95.6
Plug-in	-70	12	1.5	0.1	-89	26	2.7	0.6	-17	6.4	1.6	0.5
TMLE-CF	1.2	22	25	97.2	5.0	29	28	93.9	-1.1	4.4	4.2	92.9

Table 1: Simulation results for experiment 1: sample size $n = 500$, quantile $\tau = 0.5$, Monte Carlo bias, Monte Carlo standard deviation (SD), average standard error (SE), and coverage of 95% Wald confidence intervals (Cov). All values have been multiplied by 10^2.

especially for $\tau = 0.9$ and in the setting with continuous exposure. The use of cross-fitting leads to a bias reduction and to estimated standard errors that better approximate the Monte Carlo standard deviation, resulting in a better coverage. For $\tau = 0.5$, we sometimes see the estimators without cross-fitting performing slightly better in terms of bias, likely because the nuisance parameters are estimated on larger sample sizes. When comparing DML-CF with TMLE-CF, we often observe the latter having a larger SD. However, in terms of bias and coverage, we always observe targeted learning outperforming the other estimator. As expected, all estimators perform worse for more extreme quantiles.

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0034

4.2 Experiment 2: Extreme propensity scores

In this experiment, the propensity scores can take extreme values (i.e., close to 0 or 1), leading to limited overlap between exposed and unexposed individuals. We evaluated the same estimators as in experiment 1 and included a naive (misspecified) main effects quantile regression model (QR). The results are shown in Table 2 and in the online appendix.

Estimator	Sample size 500				Sample size 1000			
	Bias	**SD**	**SE**	**Cov**	**Bias**	**SD**	**SE**	**Cov**
CSQR	-0.30	29	30	94.6	-0.089	20	21	95.6
QR	16	36	36	92.3	17	25	26	91.1
Plug-in	-72	14	1.5	0.0	-66	12	1.1	0.0
TMLE-CF	-6.5	32	34	95.4	-6.8	22	23	94.3

Table 2: Simulation results for experiment 2: sample sizes $n = 500$ (left) and $n = 1000$ (right), quantile $\tau = 0.5$, Monte Carlo bias, Monte Carlo standard deviation (SD), average standard error (SE), and coverage of 95% Wald confidence intervals (Cov). All values have been multiplied by 10^2.

The proposed estimators behave similarly as in the first experiment, with TMLE-CF again showing the best performance. It clearly performs better than a naive quantile regression, particularly in terms of bias. The

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0034

TMLE estimator has a large bias, which is likely caused by overfitting, given that the bias of TMLE-CF is much smaller. As in the first experiment, cross-fitting provides a major improvement.

4.3 Experiment 3: Randomized dichotomous exposure

In this experiment, we first chose the exposure to be randomized with $\text{pr}(A = 1) = 0.5$, independently from all other variables. We compared our estimators with two estimators recently proposed in Athey et al. (2023). Both estimators compare marginal quantiles $Q_\tau(Y^a)$ of counterfactual distributions Y^a, $a = 0, 1$. They are designed for a constant quantile treatment effect between two groups, meaning that there is a location shift between the outcome distribution of exposed versus unexposed subjects. A brief decsription of these estimators is given in the online appendix. The results are shown in Table 3 (left).

For the proposed estimators we see similar results as in experiment 1. TMLE-CF outperforms the other estimators with a bias of 0.013 and coverage of 95.1%. While being assumption-lean, it still performs slightly better than the estimators defined in Athey et al. (2023), despite all assumptions being met in this simulation experiment.

However, when applying the two estimators from Athey et al. (2023) in

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0034

Estimator	Randomized $A \in \{0,1\}$				Setting 2 of Table 1			
	Bias	**SD**	**SE**	**Cov**	**Bias**	**SD**	**SE**	**Cov**
TMLE-CF	0.013	0.23	0.23	95.1	0.050	0.29	0.28	93.9
IFB	0.016	0.22	0.23	96.4	-0.21	0.23	0.25	88.0
WAQ	0.018	0.22	0.23	95.8	-0.19	0.25	0.25	88.5

Table 3: Simulation results for (left) randomized dichotomous exposure and (right) setting 2 of experiment 1: sample size $n = 500$, quantile $\tau = 0.5$, Monte Carlo bias, Monte Carlo standard deviation (SD), average standard error (SE) and coverage of 95% Wald confidence intervals (Cov). IFB and WAQ are the Influence Function Based estimator and Weighted Average Quantile estimator from Athey et al. (2023).

the heteroscedasticity setting from experiment 1, we observe a much worse performance (see Table 3 (right)). Here the location shift assumption is violated, leading to an inflated bias. As expected, our proposal outperforms their estimators. The results of the TMLE-CF estimator are taken from Table 1 and included in Table 3 (right) for comparison.

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0034

4.4 Experiment 4: Variable selection in higher dimensions

In this simulation experiment, we considered a covariate vector of dimension 50, and we also compare sample size $n = 250$ with $n = 500$.

Furthermore, we compare the performance of a quantile regression with AIC-based backward stepwise variable selection (QRvs) with the TMLE estimator defined in (3.11) with 5-fold cross-fitting (TMLEvs-CF) for $\tau = 0.5$. The results are shown in Table 4. The online appendix contains the entire table with additional results for all estimators.

Estimator	Sample size 250				Sample size 500			
	Bias	**SD**	**SE**	**Cov**	**Bias**	**SD**	**SE**	**Cov**
CSQR	0.66	15	16	96.0	0.25	10	11	95.7
QRvs	2.4	17	16	91.9	0.36	12	11	92.8
TMLEvs-CF	1.3	20	19	94.3	-0.32	13	12	94.8

Table 4: Simulation results for variable selection: sample size $n = 250$ (left) and $n = 500$ (right), quantile $\tau = 0.5$, Monte Carlo bias, Monte Carlo standard deviation (SD), average standard error (SE) and coverage of 95% Wald confidence intervals (Cov). All values have been multiplied by 10^2.

We observe that the standard quantile regression with variable selection (QRvs) performs well in terms of bias. However, there is a slight underesti-

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0034

mation of the standard errors, leading to a coverage slightly below 95%. In contrast, the TMLE estimator achieves nominal coverage. As expected, its standard errors are more accurate because it takes the excess uncertainty of the variable selection procedure into account.

5. Data analysis

To illustrate our proposal, we reanalyze an observational study to estimate annual health care (HC) and lost productivity costs associated with excess weight among the adult population in Belgium, using national health data. Gorasso et al. (2022) estimated the average difference in HC costs between individuals with underweight (BMI $<$ 18.5) and normal weight ($18.5 \leq$ BMI $<$ 25) versus overweight ($25 \leq$ BMI $<$ 30) and obesity (BMI $\geq$ 30) using a linear main effect model. They used a double-selection approach to select potential confounders (Belloni et al., 2013) and estimated a population level effect using g-computation (Robins, 1986). Candidate variables included age groups, gender, household educational level, household level of income, and some behavioral risk factors such as smoking, alcohol misuse, unhealthy eating behavior, and physical inactivity. In the online appendix we show summary statistics for these variables in the original population and in the propensity overlap population. Estimation for

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0034

the latter is based on debiased machine learning techniques, which may be of independent methodological interest.

Given the right-skewed nature of HC costs, average effects may be driven by extremes, making a quantile analysis more informative. We used the proposed TMLE estimator (3.11) with 5-fold cross-fitting to assess the conditional association between HC cost and body weight (dichotomized as underweight and normal weight vs. overweight and obesity), considering the same set of candidate baseline variables as Gorasso et al. (2022). This estimator makes use of parametric quantile regression with AIC-based backward stepwise variable selection. We compared the results with a traditional parametric quantile model that uses the same variable selection procedure. All analyses include sampling weights to adjust for the multi-stage sampling design (Demarest et al., 2013).

With access to complete data of 3818 individuals, the traditional parametric quantile regression with stepwise variable selection delivers an estimate of the median incremental HC cost per capita of a high BMI of €187.70 (SE: 53.95) and a 90th percentile equal to €521.13 (SE: 259.05). The proposed TMLE approach (with cross-fitting) delivers an estimated median incremental HC cost per capita of excess body weight of €205.09 (SE: 75.94) and an estimate of €1142.42 (SE: 220.15) for the 90th percentile. These

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0034

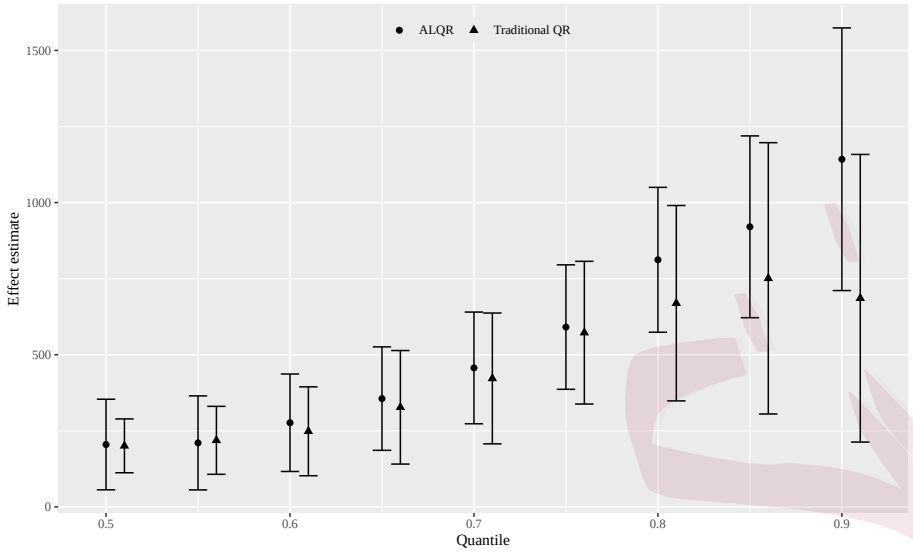

Figure 1: Effect estimate of overweight and obesity on annual health care cost per capita, with 95% confidence intervals for the proposal (3.11) (ALQR) and a traditional quantile regression approach (Traditional QR).

differences are likely caused by misspecification of the parametric quantile regression model, which did not include interactions involving BMI class. For instance, there is a significant interaction between BMI class and education level ($p = 0.0023$). Results for more quantiles are shown in Fig. 1. We observe that the difference in annual HC cost for high versus low body weight is much larger when looking at higher quantiles of the conditional outcome distribution. A sensitivity analysis, which examines additional variability for extreme quantiles related to randomness in the cross-fitting,

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0034

is included in the online appendix.

6. Discussion

In this article, inspired by Vansteelandt and Dukes (2022), we have focused on providing a scalar quantile-based summary of the conditional association between an outcome and an exposure, given covariates. Our results are primarily of interest in settings where averages are less informative and interest rather lies in median (or other quantile) effects. The main motivation to focus on conditional quantiles is because marginal quantiles of counterfactual outcomes often demand strong extrapolations by considering a hypothetical scenario where everyone got (or did not get) treatment. This is problematic when some covariate groups are highly unlikely to get treatment. We accommodate this by downweighting those covariate groups. These groups contain very little information on the exposure effect and there may often be less interest in the effect for those individuals. However, we faced challenges because of the conditional density $f\{Q_\tau(Y|A,L)|A,L\}$ appearing as a nuisance parameter in the estimator. Estimation of this density is generally difficult. Like Zhong and Wang (2023), we assume homoscedastic residuals to estimate it. The 'FKSUM' estimator for instance assumes this conditional density to be equal for all subjects. One way

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0034

to sidestep this assumption is to estimate the (marginal) distribution of $\{Y - E(Y|A, L)\}\text{Var}(Y|A, L)^{-1/2}$ and then transform an estimate of this distribution to estimate $f\{Q_\tau(Y|A, L)|A, L\}$. However, this approach resulted in an extremely unstable behaviour of our estimator for Ψ_τ (not shown) because of estimates of this density close to zero. The alternative strategy, shown in the online appendix, is more flexible, does not suffer from instability, but is computationally very expensive. Since the performance of the TMLE estimator did not significantly improve when using the latter density estimator (even in the absence of a location shift), we opted for the simpler one. A broader exploration of different (conditional) density estimators may lead to further improvements in the proposed estimator. Unfortunately, the existing literature on inverse density weighting is not directly applicable (Huling et al., 2023), because we need to evaluate the density in a quantile instead of in the observed outcomes. In future work, we will examine if Auto-DML (Chernozhukov et al., 2020, 2022) could offer a solution to directly estimate the inverse density rather than first estimating the density and then taking the reciprocal.

From the simulation results for more extreme quantiles, it was often observed that the average estimated standard error is less than the Monte Carlo standard deviation of the estimator. This could be the result of

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0034

ignoring the randomness in cross-fitting, for which rank-transformed subsampling (Guo and Shah, 2023) could offer improvements. Alternatively, one could repeat the cross-fitting multiple times and aggregate the results.

Our proposed estimators allow for estimating the nuisance parameters through data-adaptive techniques. To estimate $Q_\tau(Y|A,L)$, we used either quantile regression forests or parametric quantile regression with variable selection. A potential improvement could be achieved via ensemble learners, e.g., by constructing an ensemble through the R package 'SuperLearner', including the aforementioned methods in its library, potentially supplemented with other estimation methods (e.g., Lasso penalized quantile regression). In this way, the user no longer needs to manually choose which algorithm to use. However, to our knowledge such ensemble learners for quantile estimation are currently unavailable.

In future work, estimand (2.2) can be generalized by including a link function and replacing model (2.3) by $g\{Q_\tau(Y|A,L)\} = \beta_\tau A + g\{Q_\tau(Y|A=0,L)\}$, with link function g. More details about this proposal can be found in the online appendix.

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0034

Supplementary Material

Supplementary material available online includes details on the construction of adjusted estimators based on parametric quantile regression with variable selection, proofs, additional simulation results, a sensitivity analysis for the data analysis, and a generalization of the estimand with link function.

Acknowledgments

We thank Statbel for BHIS sample selection and fieldwork coordination, and the InterMutualistic Agency (IMA) for supporting data linkage. We are also grateful to all BHIS participants. BHIS is financed by the Federal and Inter-Federated Belgian Public Health authorities. The linkage between BHIS data and the Belgian Compulsory Health Insurance data is financed by the National Institute for Health and Disability Insurance. This work is supported by Advanced ERC grant ACME (101141305) and the Fonds Professor Frans Wuytack (Ghent University).

References

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

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0034

REFERENCES

Alejo, J., A. F. Galvao, and G. Montes-Rojas (2018). Quantile continuous treatment effects. Econometrics and Statistics 8, 13–36.

Athey, S., P. J. Bickel, A. Chen, G. W. Imbens, and M. Pollmann (2023, 07). Semi-parametric estimation of treatment effects in randomised experiments. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 85(5), 1615–1638.

Belloni, A., V. Chernozhukov, and C. Hansen (2013, 11). Inference on Treatment Effects after Selection among High-Dimensional Controls†. The Review of Economic Studies 81(2), 608–650.

Bickel, P. J., C. A. Klaassen, P. J. Bickel, Y. Ritov, J. Klaassen, J. A. Wellner, and Y. Ritov (1993). Efficient and adaptive estimation for semiparametric models, Volume 4. Baltimore: John Hopkins University Press.

Breiman, L. (2001). Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). Statistical Science 16(3), 199–231.

Caracciolo, F. and M. Furno (2017). Quantile treatment effect and double robust estimators: an appraisal on the italian labor market. Journal of Economic Studies 4(2), 585–604.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2018, 01). Double/debiased machine learning for treatment and structural parameters. The Econometrics Journal 21(1), C1–C68.

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

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0034

REFERENCES

Chernozhukov, V., W. Newey, R. Singh, and V. Syrgkanis (2020). Adversarial estimation of Riesz representers. arXiv preprint arXiv:2101.00009.

Chernozhukov, V., W. K. Newey, and R. Singh (2022). Automatic debiased machine learning of causal and structural effects. Econometrica 90(3), 967–1027.

Das, K., M. Krzywinski, and N. Altman (2019). Quantile regression. Nature Methods 16(6), 451–452.

Demarest, S., J. Van der Heyden, R. Charafeddine, S. Drieskens, L. Gisle, and J. Tafforeau (2013). Methodological basics and evolution of the Belgian health interview survey 1997–2008. Archives of Public Health 71, 1–10.

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.

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

Firpo, S. (2007). Efficient Semiparametric Estimation of Quantile Treatment Effects. Econometrica 75(1), 259–276.

Gorasso, V., I. Moyersoen, J. Van der Heyden, K. De Ridder, S. Vandevijvere, S. Vansteelandt, D. De Smedt, and B. Devleesschauwer (2022). Health care costs and lost productivity costs related to excess weight in Belgium. BMC Public Health 22(1), 1–11.

Guo, F. R. and R. D. Shah (2023). Rank-transformed subsampling: inference for multiple data

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0034

splitting and exchangeable p-values. arXiv preprint arXiv:2301.02739.

He, X. and P. Shi (1996). Bivariate tensor-product b-splines in a partly linear model. Journal of Multivariate Analysis 58(2), 162–181.

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

Huling, J. D., N. Greifer, and G. Chen (2023). Independence weights for causal inference with continuous treatments. Journal of the American Statistical Association 119(546), 1657–1670.

Koenker, R. (2005). Quantile Regression, Volume 38. Cambridge University Press.

Koenker, R. and G. Bassett (1978). Regression Quantiles. Econometrica 46(1), 33–50.

Koenker, R., V. Chernozhukov, X. He, and L. Peng (2017). Handbook of quantile regression (1st ed.). New York: Chapman & Hall/CRC.

Koenker, R. and K. F. Hallock (2001). Quantile Regression. Journal of Economic Perspectives 15(4), 143–156.

Lee, S. (2003). Efficient semiparametric estimation of a partially linear quantile regression model. Econometric Theory 19(1), 1–31.

Leeb, H. and B. M. Pötscher (2005). Model selection and inference: Facts and fiction. Econometric Theory 21(1), 21–59.

Lehmann, E. L. and H. J. D'Abrera (1975). Nonparametrics: Statistical methods based on

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0034

ranks. Holden-Day.

Lv, Y., R. Zhang, W. Zhao, and J. Liu (2015). Quantile regression and variable selection of partial linear single-index model. Annals of the Institute of Statistical Mathematics 67(2), 375–409.

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

Pfanzagl, J. (1990). Estimation in semiparametric models, pp. 17–22. New York, NY: Springer US.

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.

Sherwood, B. and L. Wang (2016). Partially linear additive quantile regression in ultra-high dimension. The Annals of Statistics 44(1), 288–317.

Sun, Y. (2005). Semiparametric efficient estimation of partially linear quantile regression models. Annals of Economics and Finance 6(1), 105–127.

van der Laan, M. J. and S. Rose (2011). Targeted learning: causal inference for observational and experimental data, Volume 4. New York: Springer Series in Statistics.

van der Laan, M. J. and D. Rubin (2006). Targeted Maximum Likelihood Learning. U.C. Berkeley Division of Biostatistics Working Paper Series, Working Paper 213. Available at:

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0034

https://biostats.bepress.com/ucbbiostat/paper213.

van der Vaart, A. W. (1998). Functional Delta Method, pp. 291–303. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Vansteelandt, S. (2025). Towards efficient and interpretable assumption-lean generalized linear modeling of continuous exposure effects. Biometrics 81(2), ujaf071. 10.1093/biomtc/ujaf071.

Vansteelandt, S. and O. Dukes (2022, 07). Assumption-lean Inference for Generalised Linear Model Parameters (with discussion). Journal of the Royal Statistical Society Series B: Statistical Methodology 84(3), 657–685.

Vansteelandt, S., O. Dukes, K. V. Lancker, and T. Martinussen (2022). Assumption-Lean Cox Regression. Journal of the American Statistical Association 119(545), 475–484. 10.1080/01621459.2022.2126362.

Wu, C. and Y. Yu (2014). Partially linear modeling of conditional quantiles using penalized splines. Computational Statistics & Data Analysis 77, 170–187.

Wu, T. Z., K. Yu, and Y. Yu (2010). Single-index quantile regression. Journal of Multivariate Analysis 101(7), 1607–1621.

Yadlowsky, S. (2022). Explaining practical differences between treatment effect estimators with high dimensional asymptotics. arXiv preprint arXiv:2203.12538.

Yu, K., Z. Lu, and J. Stander (2003). Quantile regression: applications and current research

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0034

areas. Journal of the Royal Statistical Society: Series D (The Statistician) 52(3), 331–350.

Zheng, W. and M. J. van der Laan (2011). Cross-Validated Targeted Minimum-Loss-Based Estimation, pp. 459–474. New York, NY: Springer New York.

Zhong, Q. and J.-L. Wang (2023). Neural networks for partially linear quantile regression. Journal of Business & Economic Statistics 42(2), 603–614. 10.1080/07350015.2023.2208183.

Ghent University

E-mail: georgi.baklicharov@ugent.be

E-mail: stijn.vansteelandt@ugent.be

University of Luxembourg

E-mail: christophe.ley@uni.lu

Sciensano

E-mail: vanessa.gorasso@sciensano.be

E-mail: brecht.devleesschauwer@sciensano.be