Taylor & Francis
Taylor & Francis Group

Journal of the American Statistical Association

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

Unified unconditional regression for multivariate quantiles, M-quantiles and expectiles

Luca Merlo, Lea Petrella, Nicola Salvati & Nikos Tzavidis

To cite this article: Luca Merlo, Lea Petrella, Nicola Salvati & Nikos Tzavidis (2023): Unified unconditional regression for multivariate quantiles, M-quantiles and expectiles, Journal of the American Statistical Association, DOI: 10.1080/01621459.2023.2250512

To link to this article: https://doi.org/10.1080/01621459.2023.2250512

View supplementary material

Accepted author version posted online: 24 Aug 2023.

Submit your article to this journal

Article views: 58

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=uasa20

Check for updates

Unified unconditional regression for multivariate quantiles, M-quantiles and expectiles

Luca Merlo[a,*], Lea Petrella[b], Nicola Salvati[c], Nikos Tzavidis[d]

[a]Department of Human Sciences, European University of Rome

[b]MEMOTEF Department, Sapienza University of Rome

[c]Department of Economics and Management, University of Pisa

[d]Department of Social Statistics and Demography, Southampton Statistical Sciences Research Institute, University of Southampton

*luca.merlo@unier.it

Accepted Manuscript

Abstract

In this paper, we develop a unified regression approach to model unconditional quantiles, M-quantiles and expectiles of multivariate dependent variables exploiting the multidimensional Huber's function. To assess the impact of changes in the covariates across the entire unconditional distribution of the responses, we extend the work of Firpo et al. (2009) by running a mean regression of the recentered influence function on the explanatory variables. We discuss the estimation procedure and establish the asymptotic properties of the derived estimators. A data-driven procedure is also presented to select the tuning constant of the Huber's function. The validity of the proposed methodology is explored with simulation studies and through an application using the Survey of Household Income and Wealth 2016 conducted by the Bank of Italy.

Keywords: Influence Function, M-estimation, Multivariate Data, RIF Regression, Unconditional Partial Effect

1 Introduction

When researchers wish to determine the effect of relevant predictors across the entire distribution of the dependent variable of interest, Quantile Regression (QR), as introduced by Koenker and Bassett Jr (1978), plays a crucial role in providing a much more complete statistical analysis compared to the classical mean regression. Indeed, it allows to model conditional quantiles of a response as a function of explanatory variables and it has been greatly exploited for the study of non-Gaussian, heavy-tailed and highly skewed data. For a detailed survey and list of references of the most used QR techniques, please refer to Koenker (2005) and Koenker et al. (2017).

However, if one is interested in how the whole unconditional distribution of the outcome responds to changes in the covariates, QR methods would yield misleading inferences (see Firpo et al. 2009; Borah and Basu 2013; Maclean et al. 2014). As explained by Frölich and Melly (2013) in a simple example relating wages to years of education, the unconditional 90-th quantile refers to the high wage workers, whereas the 90-th quantile conditional on education refers to the high wage workers within each education class, who however may not necessarily be high earners overall. Presuming a strong positive correlation between education and wages, it may well be that the 90-th quantile among high school dropouts is lower than, say, the median of all Ph.D. graduates. The interpretation of the 90-th quantile is thus different for conditional and unconditional quantiles. From a policy perspective, while the welfare of highly educated people with relatively low wages catches little interest, the welfare of the poor, i.e., those located in the lower end of the unconditional distribution of wages, attracts a lot of attention in the political debate.

Accepted Manuscript

In medicine, the analysis of epidural analgesia on the duration of the second stage of labor (see Zhang et al. 2012) gives another example where researchers are interested in changes in the quantiles, q_τ, of the unconditional distribution of the response Y, $F_Y(y)$. Obstetricians are particularly concerned about the unconditional effect $dq_\tau(\text{p})/d\text{p}$ of increasing the proportion of patients receiving epidural analgesia, $\text{p} = Pr[X = 1]$, on the τ-th quantile of the unconditional distribution of the duration of second-stage labor, where X = 1 if the patient receives the treatment (epidural) and X = 0 otherwise. Unfortunately, the coefficient β_τ from a conditional quantile regression, $\beta_\tau = F_Y^{-1}(\tau \mid X = 1) - F_Y^{-1}(\tau \mid X = 0)$, is generally different from $dq_\tau(\text{p})/d\text{p} = (\Pr[Y > q_\tau \mid X = 1] - \Pr[Y > q_\tau \mid X = 0]) / f_Y(q_\tau)$, the effect of increasing the proportion of patients receiving epidural analgesia on the τ-th quantile of the unconditional distribution of Y.

There are, indeed, numerous applications of practical relevance where the ultimate research objective is the unconditional distribution of the dependent variable, as in the context of the earnings disparities between different groups of workers, the effect of education on earnings, or the distributional impacts of a particular treatment in a given population (Borah and Basu 2013; Frölich and Melly 2013; Huffman et al. 2017; Firpo et al. 2018). Therefore, a number of proposals has been introduced in the literature to estimate these unconditional effects (Machado and Mata 2005; Melly 2005).

Motivated by this interest, Firpo et al. (2009) proposed the Unconditional Quantile Regression (UQR) approach for estimating the impact of changes in the distribution of the explanatory variables on quantiles of the unconditional distribution of a dependent variable. This method builds upon the concept of Recentered Influence Function (RIF) which originates from a widely used tool in robust statistics, namely the Influence Function (IF, see Hampel 1974; Huber and Ronchetti 2009). In particular, the UQR of Firpo et al. (2009) consists of regressing the RIF of the unconditional quantile of the outcome variable on the explanatory variables using either Ordinary Least Squares (OLS), logistic regression or the nonparametric regression of Newey (1994).

Accepted Manuscript

Their approach represents an important contribution on quantile regression methods and its validity is also demonstrated by the growing scientific literature spanning from medicine, economics, social inequalities and agriculture.

These studies, however, focus on a univariate regression framework. In the analysis of multivariate data, univariate approaches are not appropriate for this purpose, as they provide only partial pictures of the phenomenon under investigation. In these cases, the research interest may focus not only on a regression model for each outcome, but also on accounting for the dependence structure between the responses. When the problem under investigation involves multivariate dependent variables, the method of Firpo et al. (2009) cannot be easily extended to higher dimensions due to the non existence of a natural ordering in a p-dimensional space, $p > 1$ (see Serfling 2002; Kong and Mizera 2012; Koenker et al. 2017; Petrella and Raponi 2019; Merlo et al. 2022).

With this paper, we contribute to the current literature extending the univariate UQR approach of Firpo et al. (2009) to a more general multivariate setting. Particularly, we propose to employ the multidimensional Huber's function defined in Hampel et al. (2011) to build a unified unconditional regression approach that encompasses multivariate quantiles, M-quantiles (Breckling and Chambers 1988) and expectiles (Newey and Powell 1987).

In the statistical literature, the Huber's function has been used to define the M-quantile for robust modeling of the entire distribution of univariate response variables, extending the ideas of M-estimation of Huber (1964) and Huber and Ronchetti (2009). This method provides a "quantile-like" generalization of the mean based on influence functions that combines in a common framework the robustness and efficiency properties of quantiles and expectiles, depending on the choice of the Huber's tuning constant; the latter offering higher estimation efficiency and computational advantages compared with the former when there are no outliers in the data. In the multivariate framework, the multidimensional Huber's function (Hampel et al. 2011) has been exploited by Breckling and Chambers (1988) to define the multivariate M-quantile using a direction vector in the Euclidean p-dimensional space in order to

Accepted Manuscript

establish a suitable ordering procedure for multivariate observations. Subsequently, Kokic et al. (2002) generalized their definition by introducing a class of multivariate M-quantiles based on weighted estimating equations, which includes multivariate quantiles and expectiles depending on the value of the tuning constant. This proposal provides a robust technique for summarizing the distribution of multidimensional data and overcomes the shortcomings of the definitions of multivariate geometric or spatial quantiles (Chaudhuri 1996) and expectiles (Herrmann et al. 2018) when extremes observations are of interest; see Girard and Stupfler (2017).

In our paper, we rely on the Kokic et al. (2002) approach using the multidimensional Huber's function to model unconditional quantiles, M-quantiles and expectiles of multivariate response variables in a unified regression framework, by choosing the tuning constant appropriately.

In order to analyze the impact of changes in the distribution of explanatory variables on the entire unconditional distribution of the responses, following Firpo et al. (2009), we regress the RIF of the proposed model on the covariates, producing the Unconditional Quantile, M-Quantile and Expectile Partial Effect (UQPE, UMQPE, UEPE) according to the selected value of the tuning constant. From the theoretical point of view, we establish the asymptotic properties of the corresponding estimators using the Bahadur representation (Bahadur 1966). Furthermore, we propose a data-driven method based on cross-validation for selecting the tuning constant that accounts for possible outliers in the data.

Using simulation studies, we illustrate the finite sample properties of the proposed methodology under different data generating processes. From an empirical standpoint, we demonstrate the usefulness of this method through the analysis of the Survey on Household Income and Wealth (SHIW) 2016 conducted by the Bank of Italy. In particular, we fit the proposed model to evaluate the effect of economic and socio-demographic characteristics of Italian households on the unconditional distributions of family wealth

Accepted Manuscript

and consumption, accounting both for the correlation between the outcomes and influential observations in the sample. The proposed multivariate approach allows us to consider consumption and wealth as part of a collective framework and it can be of great interest to investigate the unconditional effect of covariates on families' spending and wealth, with particular emphasis on those with jointly low or high consumption and wealth levels.

The remainder of the paper is organized as follows. In Section 2, we revise the RIF and its properties. Section 3 introduces the proposed unconditional regression model for multivariate response variables and provides a detailed discussion of the asymptotic properties of the introduced estimators. Finally, the empirical application is presented in Section 4, while Section 5 concludes. The simulation study and all the proofs are provided in the Supplementary Materials.

2 Notation and preliminary results

In this section, we present the main notation and concepts which we use throughout the paper. Specifically, we review the notion of Recentered Influence Function (RIF) which originates from the Influence Function (IF) of Hampel (1974). Then, we present the Unconditional Partial Effect (UPE) introduced by Firpo et al. (2009) that leads us to analyze the impact of changes in the distribution of covariates on the unconditional distribution of the response variable.

Let $\mathbf{Y}$ denote a vector-valued random variable belonging to an arbitrary sample space $\mathcal{Y}$, which can be either a subset or equal to $\mathbb{R}^p$, with absolutely continuous distribution function $F_{\mathbf{Y}}$ and consider a vector-valued functional $\nu(F_{\mathbf{Y}})$ where $\nu: \mathcal{F}_\nu \to \mathbb{R}^p$, such that $\mathcal{F}_\nu$ is the collection of all distributions on $\mathcal{Y}$ for which ν is defined. The functional ν can belong to a wide class of distributional statistics. For example, ν can be a location parameter characterizing the distribution of $\mathbf{Y}$, a measure of scatter, as well as many inequality measurements such as concentration functions.

Accepted Manuscript

The IF allows to study the effect of an infinitesimal contamination in the underlying distribution $F_{\mathbf{Y}}$ at a point $\mathbf{y}$ on the statistic $\nu(F_{\mathbf{Y}})$ we are interested in. Let us consider $\Delta_{\mathbf{y}}$ the probability measure that puts mass 1 at the value $\mathbf{y}$ and let $F_{\mathbf{Y},t\Delta_{\mathbf{y}}} = (1-t)F_{\mathbf{Y}} + t\Delta_{\mathbf{y}}$ represent the mixing distribution with $t \in [0,1]$. Following Hampel et al. (2011), the *p*-dimensional IF of $\nu(F_{\mathbf{Y}})$ is defined as:

$$IF(\mathbf{y};\nu) = \lim_{t\to 0} \frac{\nu(F_{\mathbf{Y},t\Delta_{\mathbf{y}}}) - \nu(F_{\mathbf{Y}})}{t}. \qquad (1)$$

Using the definition of IF in (1), Firpo et al. (2009) considered the RIF to analyze the statistic $\nu(F_{\mathbf{Y}})$ after a perturbation of $F_{\mathbf{Y}}$ in the direction of $\Delta_{\mathbf{y}}$. In particular, the RIF is defined as the first two terms of the von Mises linear approximation (Mises 1947) of the corresponding statistic $\nu(F_{\mathbf{Y},t\Delta_{\mathbf{y}}})$ with t = 1, namely:

$$RIF(\mathbf{y};\nu) = \nu(F_{\mathbf{Y}}) + \int IF(\mathbf{s};\nu)d\Delta_{\mathbf{y}}(\mathbf{s}) = \nu(F_{\mathbf{Y}}) + IF(\mathbf{y};\nu). \qquad (2)$$

The RIF in (2) can be interpreted as a linear approximation to a possibly complex and nonlinear statistic $\nu(F_{\mathbf{Y}})$ measuring how it is affected by an infinitesimal perturbation in $F_{\mathbf{Y}}$.

In the presence of a set of covariates $\mathbf{X} \in \mathcal{X}$, with $\mathcal{X} \subset \mathbb{R}^k$ being the support of $\mathbf{X}$, Firpo et al. (2009) suggested the use of the RIF in (2) for analyzing the impact on $\nu(F_{\mathbf{Y}})$ due to changes in the distribution of $\mathbf{X}$, $F_{\mathbf{X}}$. In particular, in order to incorporate the effect of the explanatory variables, by the law of iterated expectations it follows from (2) that:

$$\nu(F_{\mathbf{Y}}) = \int RIF(\mathbf{y};\nu)dF_{\mathbf{Y}}(\mathbf{y}) = \int \mathbb{E}[RIF(\mathbf{Y};\nu) \mid \mathbf{X} = \mathbf{x}]dF_{\mathbf{X}}(\mathbf{x}), \quad (3)$$

Accepted Manuscript

where in the first equality we used the fact that $\int IF(\mathbf{y};\nu)dF_{\mathbf{Y}}(\mathbf{y})=\mathbf{0}$ (see Hampel et al. (2011), p. 226) and in the second one we substituted in $F_{\mathbf{Y}}(\mathbf{y})=\int F_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}\,|\,\mathbf{x})dF_{\mathbf{X}}(\mathbf{x})$.

From (3) it can be seen that when one is interested in the impact of a change in the covariates $\mathbf{X}$ on a specific distributional statistic $\nu(F_{\mathbf{Y}})$, the $\mathbb{E}[RIF(\mathbf{Y};\nu)\,|\,\mathbf{X}=\mathbf{x}]$ can be modeled as a function of $\mathbf{X}$, which can be easily implemented using regression methods for the conditional mean (see Firpo et al. 2009, 2018 and Rios-Avila 2020). More formally, in this work we are mainly interested in a small location shift t in the distribution of covariates $\mathbf{X}$ from $F_{\mathbf{X}}$ to the distribution $G_{\mathbf{X}}$ of the k-dimensional vector $\mathbf{X}$, where $\tilde{X}_l=X_l$ for $l\neq j, l=1,\ldots,k$, and $\tilde{X}_j=X_j+t$. In this case, let $\alpha_j(\nu)$ denote the partial effect of a small change in the distribution of covariates from $F_{\mathbf{X}}$ to $G_{\mathbf{X}}$ on the functional $\nu(F_{\mathbf{Y}})$ and let $\boldsymbol{\alpha}(\nu)=(\alpha_1(\nu),\ldots,\alpha_k(\nu))$ be the matrix collecting all j entries. Under the assumption that the conditional distribution of $\mathbf{Y}$ given $\mathbf{X}$, $F_{\mathbf{Y}|\mathbf{X}}$, is unaffected by changes in the law of $\mathbf{X}$, the matrix of partial effects of small location shifts in the distribution of $\mathbf{X}$ is given by (see Corollary 1 in Firpo et al. 2009):

$$\boldsymbol{\alpha}(\nu)=\int\frac{d\mathbb{E}[RIF(\mathbf{Y};\nu)\,|\,\mathbf{X}=\mathbf{x}]}{d\mathbf{x}}dF_{\mathbf{X}}(\mathbf{x}), \qquad (4)$$

where $d\mathbb{E}[RIF(\mathbf{Y};\nu)\,|\,\mathbf{X}=\mathbf{x}]/d\mathbf{x}$ is understood to indicate the Jacobian matrix of all its first order partial derivatives with respect to $[\mathbf{x}_j]_{j=1}^{k}$. Firpo et al. (2009) call the quantity $\boldsymbol{\alpha}(\nu)$ in (4) as the Unconditional Partial Effect (UPE). By analogy with standard conditional regression coefficients, $\boldsymbol{\alpha}(\nu)$ corresponds to the effect of a small increase in the location of the distribution of the explanatory variables on the functional $\nu(F_{\mathbf{Y}})$, holding everything else constant. It is worth noting that our approach requires $F_{\mathbf{Y}|\mathbf{X}}$ to be " structural" in the sense of " invariant to a class of modifications" (Heckman and Vytlacil 2007, p. 4848 and Section 4.8) and rules out the presence of potential sources of endogeneity like selection bias in inference.

Accepted Manuscript

In the case of a dummy variable $X \in \{0,1\}$, the UPE represents the effect of a small increase in the probability that X = 1, namely:

$$\alpha(\nu) = \mathbb{E}[RIF(\mathbf{Y};\nu) \mid X = 1] - \mathbb{E}[RIF(\mathbf{Y};\nu) \mid X = 0]. \quad (5)$$

In addition, as discussed by Firpo et al. (2009) and Firpo et al. (2018), the approach based on the RIF can also be used to assess the effect of more general changes in the distribution of covariates on the functional $\nu(F_{\mathbf{Y}})$. For example, if the goal of the analysis is to assess how such modifications affect the dependence of $\mathbf{Y}$, ν can be a measure of multivariate scatter or a well-known correlation coefficient, provided that the corresponding influence function can be derived.

In the following section we exploit these properties for the analysis of unconditional quantile, M-quantile and expectile regressions associated to multivariate response variables.

3 Methodology

In this section, we generalize the univariate approach of Firpo et al. (2009) by developing a unifying regression method to model unconditional quantiles, M-quantiles and expectiles of vector-valued responses for exploring the effects of covariates using the RIF approach illustrated in Section 2. Firstly, we introduce the Huber's multidimensional M-function for estimating multivariate quantiles, M-quantiles and expectiles by varying the value of the tuning constant in an appropriate way. Then, in order to assess the effect of the covariates at different parts of the unconditional distribution of the dependent variable, we introduce the Unconditional Quantile, M-Quantile and Expectile Partial Effect (UQPE, UMQPE, UEPE) and establish the asymptotic properties of the corresponding estimators.

Accepted Manuscript

The Huber's function, in the univariate case, has been used by Breckling and Chambers (1988) to define the M-quantile, extending the concept of M-estimation of Huber (1964) to the quantile framework. Huber M-quantiles are a generalized form of M-estimators that include in a single modeling approach, quantiles and expectiles through a tuning constant to adjust the robustness of the estimator in the presence of outliers. In higher dimensions, extending these univariate notions to multivariate data is not a trivial task since there does not exist a natural ordering in p dimensions, $p > 1$. Already in their proposal, Breckling and Chambers (1988) considered the multivariate extension of Huber's function to estimate M-quantiles by considering a directional unit norm vector to set up a suitable ordering procedure for multidimensional data. Further, Kokic et al. (2002) generalized their approach by introducing a weighted estimating equation based on the multidimensional Huber's influence function that encompasses multivariate quantiles, M-quantiles and expectiles, depending on the value of the related tuning constant. Their method aimed to provide a robust and simple to implement technique of summarising and conveying valuable information about multidimensional data. Here and in what follows, in order to present a unified unconditional regression approach to model multivariate quantiles, M-quantiles and expectiles, we adopt the approach of Kokic et al. (2002) based on the multidimensional Huber's function.

Formally, the multidimensional Huber's influence function in Hampel et al. (2011) is defined as:

$$\Psi(\mathbf{r}) = \begin{cases} \dfrac{\mathbf{r}}{c}, & \|\mathbf{r}\| < c \\ \dfrac{\mathbf{r}}{\|\mathbf{r}\|}, & \|\mathbf{r}\| \geq c \end{cases}, \qquad \mathbf{r} \in \mathbb{R}^p, \mathbf{r} \neq \mathbf{0} \qquad (6)$$

with additionally $\Psi(\mathbf{0}) = \mathbf{0}$ and where $c \geq 0$ is the tuning constant that can be adjusted to trade robustness for efficiency, with increasing robustness when it is chosen to be close to 0 and increasing efficiency when it is chosen to be large. To show how one can use the $\Psi(\mathbf{r})$ function in (6) to estimate multivariate quantiles, M-quantiles and expectiles, we introduce the following additional

Accepted Manuscript

notation. Let $\mathcal{Y} = \mathbb{R}^p$, consider a continuous *p*-dimensional random variable $\mathbf{Y}$ and let $\mathbf{u}$ denote a unit norm direction vector ranging over the *p*-dimensional unit sphere $\mathcal{S}^{p-1} = \{\mathbf{z} \in \mathbb{R}^p : \| \mathbf{z} \| = 1\}$, where $\|\cdot\|$ denotes the Euclidean norm. Following Kokic et al. (2002), for a general value of *c*, we obtain the τ-th multivariate M-quantile of $\mathbf{Y}$ in the direction of $\mathbf{u}$, $\boldsymbol{\theta}_{\tau,\mathbf{u}}$, with $\tau \in (0, \frac{1}{2}]$, by satisfying the equation:

$$\int \eta_\delta(\varphi)\Psi(\mathbf{y} - \boldsymbol{\theta}_{\tau,\mathbf{u}})dF_{\mathbf{Y}}(\mathbf{y}) = \mathbf{0}, \qquad (7)$$

where

$$\eta_\delta(\varphi) = \begin{cases} (1-\cos\varphi)^\delta \zeta + 2\tau, & \varphi \in (-\frac{\pi}{2}, \frac{\pi}{2}) \\ -(1-\cos\varphi)^\delta \zeta + 2(1-\tau), & \varphi \in [-\pi, -\frac{\pi}{2}] \cup [\frac{\pi}{2}, \pi], \end{cases}$$

is a weighting function with $\zeta = 1 - 2\tau, \delta > 0$ and φ being the angle between $\mathbf{Y} - \boldsymbol{\theta}$ and $\mathbf{u}$, so $\cos\varphi = \frac{(\mathbf{Y}-\boldsymbol{\theta})'\mathbf{u}}{\| \mathbf{Y} - \boldsymbol{\theta} \|}$. Evidently, one can also consider non-normalized directions by explicitly writing $\cos\varphi = \frac{(\mathbf{Y}-\boldsymbol{\theta})'\mathbf{u}}{\| \mathbf{Y} - \boldsymbol{\theta} \| \| \mathbf{u} \|}$.

The function $\eta_\delta(\varphi)$ gives asymmetric weights to the residual $\mathbf{Y} - \boldsymbol{\theta}$ depending both on its length and the angle it forms with $\mathbf{u}$. Most importantly, the tuning constant *c* determines where the weighted scheme based on $\eta_\delta(\varphi)$ defines multivariate quantiles when *c* = 0 and yields multivariate expectiles as $c \to \infty$, which could be particularly fruitful when the use of outlier-robust estimation methods is not justified but there is still interest in modeling the entire distribution of $\mathbf{Y}$ (Tzavidis et al. 2010). Hence, multivariate expectiles

Accepted Manuscript

inherit the efficiency properties of standard univariate expectiles and, at the same time, consider the dependence structure between the components of the variable analyzed. Computationally, an estimate of $\boldsymbol{\theta}_{\tau,\mathbf{u}}$ in (7) can be efficiently obtained by using Iteratively Reweighted Least Squares (IRLS, Breckling et al. 2001). Clearly, the multivariate M-quantile in (7) includes the traditional notion of univariate M-quantile. Indeed, if $p = 1$, $u = 1$ and $\delta = 1$, then (7) reduces to the estimating equation of the univariate M-quantile, θ_τ, because $\cos\varphi = \operatorname{sgn}(Y - \theta_\tau)$, which implies that $\eta_\delta(\varphi) = 1 - \zeta\operatorname{sgn}(Y - \theta_\tau)$. Throughout the rest of the paper we set $\delta = 1$, but other values of δ are possible (see Kokic et al. 2002 and the Supplementary Materials for this article).

In comparison with other settings in the literature, the considered approach remedies the shortcoming of the Chaudhuri (1996) and Herrmann et al. (2018) definitions of geometric quantile and expectile which have recently been criticized because they can be situated outside the support of $\mathbf{Y}$ for extreme levels of τ (see Girard and Stupfler 2017; Konen and Paindaveine 2021). Moreover, Breckling and Chambers (1988) M-quantiles, which can be shown to coincide with the definition in Chaudhuri (1996) as a particular case, are also subject to the same problem. The definition in (7), on the contrary, lead to multivariate M-quantiles always situated within the convex hull of the sample data (for a comparative analysis between the considered definition of multivariate (M-)quantile and the geometric quantile, see the Supplementary Materials).

A particular issue in this context may be the choice of the direction $\mathbf{u}$, which is often selected on the basis of the empirical problem at hand to produce meaningful results (see Paindaveine and Šiman 2011; Kong and Mizera 2012 and Merlo et al. 2022). For example, Fraiman and Pateiro-López (2012) and Torres et al. (2017) consider a set of directions using the principal components obtained from a PCA to build a useful tool for exploratory data analysis and visualize important features of multidimensional data. In classification analysis, Farcomeni et al. (2022) focus on a single optimal direction that minimizes the misclassification error

Accepted Manuscript

meanwhile, Geraci et al. (2020) define the “allometric direction” for risk classification of abnormal multivariate anthropometric measurements.

When the directional approach is adopted, considering theoretically all directions in $\mathcal{S}^{p-1}$ simultaneously yields multivariate M-quantiles centrality regions, which allow us to provide a visual description of the location, spread, shape and dependence between the responses distribution. These quantities are of crucial interest as they are able to adapt to the underlying shape of the distribution of $\mathbf{Y}$ without being constrained to particular shapes, such as convex bodies or ellipses (Breckling et al. 2001).

Specifically, for a given level $\tau \in (0, \frac{1}{2}]$, by moving the direction $\mathbf{u}$ around the whole $\mathcal{S}^{p-1}$, the resulting set of corresponding multivariate M-quantiles generates the τ-th M-quantile region embedded within the p-dimensional Euclidean space, $R_\tau \subset \mathbb{R}^p$, defined as the collection whose vertices are:

$$R_\tau = \{\boldsymbol{\theta}_{\tau,\mathbf{u}} \big| \mathbf{u} \in \mathcal{S}^{p-1}\}. \quad (8)$$

The region in (8) is a closed surface and the corresponding M-quantile contour of order τ is defined as the boundary ∂R_τ of R_τ. The position, curvature, spread and orientation of such objects all reflect important characteristics of the data that are relevant to researchers for exploratory data analysis. M-quantile curves can thus be used for exploratory data analysis as well as a statistical diagnostic tool to evaluate the results to different values of the tuning constant. For fixed τ, when $c = 0$ it defines quantile contours and it generates expectile contours when $c \to \infty$. Meanwhile, for any $c \geq 0$, the contours are nested as τ increases. As $\tau \to 0$, instead, the τ-th M-quantile contour approaches the convex hull of the sample data providing valuable information about the extent of extremality of points (see Serfling 2002 and Kokic et al. 2002).

Accepted Manuscript

To build our model, it follows from (7) and Hampel et al. (2011) that the IF for the unconditional multivariate M-quantile $\boldsymbol{\theta}_{\tau,\mathbf{u}}$ is defined as:

$$IF(\mathbf{y};\boldsymbol{\theta}_{\tau,\mathbf{u}}) = \mathbf{M}(\boldsymbol{\theta}_{\tau,\mathbf{u}})^{-1}\eta_{\delta}(\varphi)\Psi(\mathbf{y}-\boldsymbol{\theta}_{\tau,\mathbf{u}}) \qquad (9)$$

with $\mathbf{M}(\boldsymbol{\theta}_{\tau,\mathbf{u}})$ being the $p \times p$ matrix given by:

$$\mathbf{M}(\boldsymbol{\theta}_{\tau,\mathbf{u}}) = -\int \nabla_{\theta_{\tau,\mathbf{u}}}\left(\eta_{\delta}(\varphi)\Psi(\mathbf{y}-\boldsymbol{\theta})\right)dF_{\mathbf{Y}}(\mathbf{y}), \quad (10)$$

where $\nabla_{\theta_{\tau,\mathbf{u}}}(\cdot)$ is the $p \times p$ matrix of first order derivatives of $\eta_{\delta}(\varphi)\Psi(\mathbf{y}-\boldsymbol{\theta})$ in (7) with respect to $\boldsymbol{\theta}$ evaluated at $\boldsymbol{\theta}_{\tau,\mathbf{u}}$. Then, following the idea in (2), the RIF is obtained from (9) by adding back the multivariate M-quantile $\boldsymbol{\theta}_{\tau,\mathbf{u}}$:

$$RIF(\mathbf{y};\boldsymbol{\theta}_{\tau,\mathbf{u}}) = \boldsymbol{\theta}_{\tau,\mathbf{u}} + IF(\mathbf{y};\boldsymbol{\theta}_{\tau,\mathbf{u}}). \qquad (11)$$

Two remarks are worth noticing. Firstly, (11) generalizes the RIF of the multivariate quantile when $c = 0$, where the matrix of first order derivatives of $\Psi(\mathbf{r})$ is equal to:

$$\frac{d\Psi(\mathbf{r})}{d\mathbf{r}} = \frac{1}{\|\mathbf{r}\|}\left\{\mathbf{I}_p - \frac{\mathbf{r}\mathbf{r}'}{\|\mathbf{r}\|^2}\right\}, \quad (12)$$

with $\mathbf{I}_p$ being the identity matrix of dimension p, and it coincides with the RIF of the multivariate expectile when $c \to \infty$, which implies $\frac{d\Psi(\mathbf{r})}{d\mathbf{r}} \propto \mathbf{I}_p$. Secondly, when $p = 1$ and $u = 1$, (11) reduces to the univariate RIF of standard M-quantiles which, in turn,

Accepted Manuscript

includes the RIF of the quantile in Firpo et al. (2009) and the RIF of the expectile for c arbitrarily large. Further, using this approach we are able to investigate the correlation structure of multivariate responses at different values of τ. More in detail, to study the association between multiple outcomes we analyze the covariance matrix of the RIF in (11) which, by simple calculations, can be written as:

$$\boldsymbol{\Delta}(\boldsymbol{\theta}_{\tau,\mathbf{u}}) = \mathbb{E}[IF(\mathbf{Y};\boldsymbol{\theta}_{\tau,\mathbf{u}})IF(\mathbf{Y};\boldsymbol{\theta}_{\tau,\mathbf{u}})']. \quad (13)$$

Given $\mathbf{u}$, τ and c, the off-diagonal elements of $\boldsymbol{\Delta}(\boldsymbol{\theta}_{\tau,\mathbf{u}})$ provide a measure of tail correlation between the components of $\mathbf{Y}$.

In a regression framework where covariates $\mathbf{X}$ are available, from (11) we define the unified unconditional regression model as follows:

$$\mathbb{E}[RIF(\mathbf{Y};\boldsymbol{\theta}_{\tau,\mathbf{u}}) \mid \mathbf{X} = \mathbf{x}] = m_{\boldsymbol{\theta}_{\tau,\mathbf{u}}}(\mathbf{x}), \quad (14)$$

where $m(\cdot)$ is an unknown function of explanatory variables $\mathbf{X}$ to be estimated. Our objective is to identify how changes in the distribution of $\mathbf{X}$ affect the multivariate quantile, M-quantile and expectile of the unconditional distribution of $\mathbf{Y}$. Following (4), for a given level τ, direction $\mathbf{u}$, and constant $c \geq 0$, the Unconditional M-Quantile Partial Effect (UMQPE), $\boldsymbol{\alpha}_{\tau,\mathbf{u}}$, is formally defined as:

$$\boldsymbol{\alpha}_{\tau,\mathbf{u}} = \int \frac{d\mathbb{E}[RIF(\mathbf{Y};\boldsymbol{\theta}_{\tau,\mathbf{u}}) \mid \mathbf{X} = \mathbf{x}]}{d\mathbf{x}} dF_{\mathbf{X}}(\mathbf{x}). \quad (15)$$

It is worth noting that the proposed approach has several appealing properties. Firstly, the UMQPE in (15) is easy to compute as it does not depend on the density of $\mathbf{Y}$ which would entail the use of nonparametric density estimation procedures (Kokic

Accepted Manuscript

et al. 2002). Secondly, this methodology allows us to directly control the robustness to outliers and estimation efficiency by means of the tuning constant c, i.e., when $c = 0$ we have the UQPE and when $c \to \infty$ we have the UEPE. In practice, the UMQPE indicates the effect of increasing the years of schooling or income, say, across the consumption and wealth distributions, as we will discuss in the next section.

Before concluding this section, it is relevant to compare the proposed unconditional regression model with the standard conditional regression approach. Suppose that $\mathbf{Y} = h(\mathbf{X}) + \epsilon$, where $h : \mathcal{X} \to \mathbb{R}^p$ is an unknown function with bounded first partial derivatives and ϵ is a p-dimensional random error independent of $\mathbf{X}$. For a given τ, $\mathbf{u}$ and $c \geq 0$, we denote the τ-th multivariate M-quantile of $\mathbf{Y}$ conditional on $\mathbf{X} = \mathbf{x}, \boldsymbol{\theta}_{\tau,\mathbf{u}}(\mathbf{x}) = h(\mathbf{x})$, as the solution of the following estimating equation:

$$\int \eta_\delta(\varphi)\Psi(\mathbf{y} - \boldsymbol{\theta}_{\tau,\mathbf{u}}(\mathbf{x}))dF_{\mathbf{Y}|\mathbf{X}}(\mathbf{y} \mid \mathbf{x}) = \mathbf{0}. \quad (16)$$

Consequently, the effect of a small change in $\mathbf{X}$ on the conditional M-quantile of $\mathbf{Y}, \boldsymbol{\theta}_{\tau,\mathbf{u}}(\mathbf{x})$, which we denote as Conditional M-Quantile Partial Effect (CMQPE), is given by:

$$\boldsymbol{\alpha}_{\tau,\mathbf{u}}(\mathbf{x}) = \frac{dh(\mathbf{x})}{d\mathbf{x}}. \quad (17)$$

In order to clarify the interpretation of (15), following Firpo et al. (2009), we provide a useful representation of the UMQPE in terms of the conditional distribution of $\mathbf{Y}$ given $\mathbf{X}$ and show how it is related to the CMQPE in (17). Let $\mathbf{W}_{\tau,\mathbf{u}} : \mathcal{X} \to \mathbb{R}^{p \times p}$ define the weighting matrix function:

$$\mathbf{W}_{\tau,\mathbf{u}}(\mathbf{x}) = \mathbf{M}(\boldsymbol{\theta}_{\tau,\mathbf{u}})^{-1}\mathbb{E}[\nabla\left(\eta_\delta(\varphi)\Psi(\mathbf{Y} - \boldsymbol{\theta}_{\tau,\mathbf{u}})\right) \mid \mathbf{X} = \mathbf{x}] \quad (18)$$

Accepted Manuscript

and let $s_{\tau,\mathbf{u}}$ be an auxiliary function $s_{\tau,\mathbf{u}}:\mathcal{X}\to(0,1)$ required to establish the link between the UMQPE and the CMQPE. The mapping $s_{\tau,\mathbf{u}}$ can be thought as a "matching" function indicating where the unconditional multivariate M-quantile $\boldsymbol{\theta}_{\tau,\mathbf{u}}$ falls in the conditional distribution of $\mathbf{Y}$ given covariates, i.e.:

$$s_{\tau,\mathbf{u}}(\mathbf{x})=\{\tilde{\tau}:\boldsymbol{\theta}_{\tilde{\tau},\mathbf{u}}(\mathbf{x})=\boldsymbol{\theta}_{\tau,\mathbf{u}}\}. \quad (19)$$

In this work, we require that, under the condition that $F_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}\,|\,\mathbf{x})$ is continuous and strictly monotonic (Breckling and Chambers 1988), $s_{\tau,\mathbf{u}}(\mathbf{x})$ is a singleton, that is, there is only one $\tilde{\tau}$ that satisfies the equality in (19). Generally, three situations may occur: $s_{\tau,\mathbf{u}}(\mathbf{x})$ is unique, it is interval-valued or it is empty for some $\mathbf{x}$ (Alejo et al. 2021). From a practical point of view, for fixed $\mathbf{x}$, if $s_{\tau,\mathbf{u}}(\mathbf{x})$ is interval-valued, then a possible solution would be to take the average of all $\tilde{\tau}$'s inside that particular interval. If, on the other hand, $s_{\tau,\mathbf{u}}(\mathbf{x})$ is empty, one could impute $\tilde{\tau}$ by taking the average of the nearest neighbors to the covariate value $\mathbf{x}$. Refer to the Supplementary Materials for a discussion and an estimator of the matching function.

The next Theorem establishes a link between the UMQPE and the CMQPE.

Theorem 1. *Assume that* $\mathbf{Y}=h(\mathbf{X})+\epsilon$ *where* $h(\cdot)$ *is an unknown function with bounded first partial derivatives and* ϵ *is an error term independent of* $\mathbf{X}$*. For a given* $\tau\in(0,\frac{1}{2}],\mathbf{u}\in\mathcal{S}^{p-1}$ *and* $c\geq 0$*, the UMQPE* $\boldsymbol{\alpha}_{\tau,\mathbf{u}}$ *can be written as:*

$$\boldsymbol{\alpha}_{\tau,\mathbf{u}}=\mathbb{E}[\mathbf{W}_{\tau,\mathbf{u}}(\mathbf{X})\boldsymbol{\alpha}_{s_{\tau,\mathbf{u}}(\mathbf{X}),\mathbf{u}}(\mathbf{X})], \quad (20)$$

where the expectation is taken over the distribution of $\mathbf{X}$*.*

Accepted Manuscript

Proof. See Proof of Theorem 1 in the Supplementary Materials. □

From Theorem 1 there follow several interesting considerations. Firstly, it formally shows that, unlike conditional means which average up to the unconditional mean thanks to the law of iterated expectations, conditional multivariate M-quantiles do not average up to their unconditional counterparts. On the contrary, the UMQPE is equal to a weighted average, over the distribution of the covariates, of the CMQPE at the $s_{\tau,\mathbf{u}}(\mathbf{X})$-th conditional M-quantile level corresponding to the τ-th unconditional M-quantile of $\mathbf{Y}$ in the direction $\mathbf{u}$. Secondly, it generalizes the result in Firpo et al. (2009) to the multivariate setting which also holds in the univariate case when $p = 1$ and $u = 1$. Furthermore, Theorem 1 is useful for interpreting the parameters of the proposed regression method. For instance, in a linear model where $h(\mathbf{X}) = \boldsymbol{\beta}'\mathbf{X}$, the UMQPE and CMQPE are both equal to the matrix of regression coefficients $\boldsymbol{\beta}$ for any choice of τ and $\mathbf{u}$. More generally, the UMQPE and CMQPE will be different depending on the structural form of $h(\cdot)$ and the distribution of $\mathbf{X}$ as described in Theorem 1. Lastly, in the presence of heteroskedastic errors, one can consider the linear heteroskedastic model $\mathbf{Y} = \boldsymbol{\beta}'\mathbf{X} + \text{diag}(\boldsymbol{\gamma}'\mathbf{X})\epsilon$ which is often used in economics, with $\boldsymbol{\gamma}$ being a matrix of coefficients. In this case, to exploit Theorem 1 a practical solution would be to adopt a flexible form of $h(\cdot)$ and estimate the conditional expectation of the RIF using nonparametric regression approaches. Alternatively, a possible way to alleviate heteroskedasticity is to apply a transformation to the original response variable.

3.1 Estimation

In this section, we discuss the estimation of the UQPE, UMQPE and UEPE using the RIF regression approach. Following Firpo et al. (2009), we suggest two methods for modeling the conditional expectation of the RIF. For a given τ, direction $\mathbf{u}$ and $c \geq 0$, the first one assumes that the RIF in (14) is linear in the covariates, $m_{\theta_{\tau,\mathbf{u}}}(\mathbf{X}) = \boldsymbol{\beta}'(\boldsymbol{\theta}_{\tau,\mathbf{u}})\mathbf{X}$ and estimate $\boldsymbol{\alpha}_{\tau,\mathbf{u}}$ in (15) via an OLS regression

Accepted Manuscript

of the $RIF(\mathbf{Y};\boldsymbol{\theta}_{\tau,\mathbf{u}})$ as a dependent variable onto the covariates $\mathbf{X}$ by using a two-step procedure. Specifically, an estimate $\boldsymbol{\theta}_{\tau,\mathbf{u}}$ of $\boldsymbol{\theta}_{\tau,\mathbf{u}}$ is obtained by solving (7) via IRLS, substitute $\boldsymbol{\theta}_{\tau,\mathbf{u}}$ in (11) and then estimate $\boldsymbol{\alpha}_{\tau,\mathbf{u}}$ by regressing the $RIF(\mathbf{Y};\boldsymbol{\theta}_{\tau,\mathbf{u}})$ on $\mathbf{X}$. Let $(\mathbf{Y}_i, \mathbf{X}_i), i = 1,\ldots,n$, denote a random sample of size *n*, the estimator of the UMQPE in (15), $\boldsymbol{\alpha}_{\tau,\mathbf{u}}$, is defined as follows:

$$\boldsymbol{\alpha}_{\tau,\mathbf{u}} = \hat{\Omega}_{\mathbf{X}}^{-1} \frac{1}{n} \sum_{i=1}^{n} \{\mathbf{X}_i RIF'(\mathbf{Y}_i;\boldsymbol{\theta}_{\tau,\mathbf{u}})\}. \quad (21)$$

where $\hat{\Omega}_{\mathbf{X}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i \mathbf{X}'_i$.

The second method estimates non-parametrically $\mathbb{E}[RIF(\mathbf{Y};\boldsymbol{\theta}_{\tau,\mathbf{u}}) \mid \mathbf{X}]$ using regression B-splines to account for nonlinear effects of $\mathbf{X}$ on the RIF. Once we have regressed $RIF(\mathbf{Y};\boldsymbol{\theta}_{\tau,\mathbf{u}})$ on the basis functions of the original covariates, as the object of interest is the average of $\frac{d\mathbb{E}[RIF(\mathbf{Y};\boldsymbol{\theta}_{\tau,\mathbf{u}}) \mid \mathbf{X} = \mathbf{x}]}{d\mathbf{x}}$, to obtain the UMQPE we simply take derivative of B-spline basis functions and average them with respect to $\mathbf{X}$. Note also that other nonparametric approaches can be adopted such as power series estimators (Newey 1994) or orthogonal polynomials.

Finally, the estimators of the UQPE and UEPE related to (15) can be obtained following the same procedure by setting $c = 0$ and c large enough such that $\| \mathbf{Y}_i - \boldsymbol{\theta}_{\tau,\mathbf{u}} \| < c, \forall i = 1,\ldots,n$, respectively. If the researcher is interested in modeling multivariate quantiles or expectiles, one can simply set $c = 0$ or c to a relatively large value. In all other cases, the UMQPE estimator in (21) requires choosing a reasonable value for the tuning constant, based on the data structure and the aim of the analysis.

Accepted Manuscript

The choice of an appropriate value for c is not straightforward. Ideally, it should be data-driven and account for possible outliers in the data. In the literature on univariate M-estimation, c can be either fixed a-priori or defined by the data analyst to achieve a specified asymptotic efficiency under normality (Huber and Ronchetti 2009), maximize the asymptotic efficiency (Wang et al. 2007) or it can be estimated in a likelihood framework as illustrated by Bianchi et al. (2018). In our multivariate context, we propose to select the tuning constant via K-fold cross-validation which allows us to consider c as a data-driven parameter. In particular, for fixed τ and $\mathbf{u}$, we construct a uniform grid of values from c_{min} = 0.1 to $c_{max} = \max_{i=1,\ldots,n} \| \mathbf{Y}_i \|$. Then, for each value of $c \in [c_{min}, \ldots, c_{max}]$, we fit the proposed model and determine the optimal value, denoted with $c^\star$, in the sense that it minimizes the estimated prediction error across the K folds, that is:

$$\mathrm{CV}_K(c) = \frac{1}{n} \sum_{k=1}^{K} \sum_{i \in I_k} (\mathbf{Y}_i - \boldsymbol{\theta}^{(k)}_{\tau,\mathbf{u}}(c))' (\mathbf{Y}_i - \boldsymbol{\theta}^{(k)}_{\tau,\mathbf{u}}(c)), \qquad (22)$$

where $I_1, \ldots, I_K$ is a random partition of the n observations into K folds and $\boldsymbol{\theta}^{(k)}_{\tau,\mathbf{u}}(c)$ is the estimate of $\boldsymbol{\theta}_{\tau,\mathbf{u}}$ obtained using the entire sample except data points in the k-th fold for a given value of c. Finally, $c^\star$ is estimated by:

$$c^\star = \underset{c \in [c_{min}, \ldots, c_{max}]}{\arg\min} \ \mathrm{CV}_K(c). \qquad (23)$$

3.2 Asymptotic properties

This section presents the asymptotic properties of the estimator $\boldsymbol{\alpha}_{\tau,\mathbf{u}}$ in (21) where the $\mathbb{E}[RIF(\mathbf{Y}; \nu) \mid \mathbf{X} = \mathbf{x}]$ is modeled as a linear function of $\mathbf{X}$. Specifically, we derive the Bahadur-type (Bahadur 1966) representation, consistency and asymptotic normality for

Accepted Manuscript

fixed τ, direction $\mathbf{u}$ and c. To prove the following results, we follow Firpo et al. (2009) where they consider the IF and not its recentered version. Either using the IF or the RIF, all regression coefficients are the same, the only exception being the intercept.

Consider the following assumptions:

(A1) The distribution of the random vector $\mathbf{Y}$ is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^p$, with a density bounded on every compact subset of $\mathbb{R}^p$.

(A2) The observations $(\mathbf{Y}_i, \mathbf{X}_i), i = 1, \ldots, n$ are an i.i.d. sample from $(\mathbf{Y}, \mathbf{X})$.

(A3) $\mathbb{E}[|| \eta_\delta(\varphi)\Psi(\mathbf{Y} - \boldsymbol{\theta}) ||] < \infty, \forall \boldsymbol{\theta} \in \mathbb{R}^p$.

(A4) $\mathbb{E}[|| \eta_\delta(\varphi)\Psi(\mathbf{Y} - \boldsymbol{\theta}) ||^2] < \infty$ for each $\boldsymbol{\theta}$ in a neighborhood of $\boldsymbol{\theta}_{\tau,\mathbf{u}}$.

(A5) The $p \times p$ matrix $\mathbf{M}(\boldsymbol{\theta}_{\tau,\mathbf{u}})$ in (10) is positive definite.

(A6) $\hat{\Omega}_{\mathbf{X}}$ is nonsingular almost surely for n sufficiently large and converges to $\Omega_{\mathbf{X}} = \mathbb{E}[\mathbf{X}\mathbf{X}']$.

It is worth noticing that assumptions A1-A6 are quite mild and are standard in robust estimation theory. For instance, assumptions A1-A5 are needed for the Bahadur (Bahadur 1966) representation and ensure the invertibility of $\mathbf{M}(\boldsymbol{\theta}_{\tau,\mathbf{u}})$.

In order to present the asymptotic properties $\boldsymbol{\alpha}_{\tau,\mathbf{u}}$, we first need to establish the Bahadur-type representation for $\boldsymbol{\theta}_{\tau,\mathbf{u}}$ and its limiting distribution.

Accepted Manuscript

Theorem 2. *Let assumptions A1-A5 hold. Then, for any* $\tau \in (0, \frac{1}{2}]$ *and* $\mathbf{u} \in \mathcal{S}^{p-1}$*, the following asymptotic linear representation holds:*

$$\sqrt{n}(\boldsymbol{\theta}_{\tau,\mathbf{u}} - \boldsymbol{\theta}_{\tau,\mathbf{u}}) = \mathbf{M}(\boldsymbol{\theta}_{\tau,\mathbf{u}})^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \eta_{\delta}(\varphi_i) \Psi(\mathbf{Y}_i - \boldsymbol{\theta}_{\tau,\mathbf{u}}) + o_p(1) \quad (24)$$

and

$$\sqrt{n}(\boldsymbol{\theta}_{\tau,\mathbf{u}} - \boldsymbol{\theta}_{\tau,\mathbf{u}}) \xrightarrow{p} \mathcal{N}(\mathbf{0}, \mathbf{M}(\boldsymbol{\theta}_{\tau,\mathbf{u}})^{-1} \mathbf{D}(\boldsymbol{\theta}_{\tau,\mathbf{u}}) \mathbf{M}(\boldsymbol{\theta}_{\tau,\mathbf{u}})^{-1}) \quad \text{as} \quad n \to \infty, \quad (25)$$

where $\mathbf{D}(\boldsymbol{\theta}_{\tau,\mathbf{u}})$ *defines a p × p matrix:*

$$\mathbf{D}(\boldsymbol{\theta}_{\tau,\mathbf{u}}) = \mathbb{E}[\eta_{\delta}^2(\varphi) \Psi(\mathbf{Y} - \boldsymbol{\theta}_{\tau,\mathbf{u}}) \Psi'(\mathbf{Y} - \boldsymbol{\theta}_{\tau,\mathbf{u}})]. \quad (26)$$

Proof. See Proof of Theorem 2 in the Supplementary Materials. □

To prove consistency and asymptotic normality of $\boldsymbol{\alpha}_{\tau,\mathbf{u}}$, we exploit Theorem 2 and define the *k* × *p* matrix of OLS regression coefficients of $IF(\mathbf{Y}; \boldsymbol{\theta}_{\tau,\mathbf{u}})$ on $\mathbf{X}$:

$$\boldsymbol{\beta}(\boldsymbol{\theta}_{\tau,\mathbf{u}}) = \hat{\Omega}_{\mathbf{X}}^{-1} \frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i IF'(\mathbf{Y}_i; \boldsymbol{\theta}_{\tau,\mathbf{u}}) \quad (27)$$

whose population counterpart is:

Accepted Manuscript

$$\boldsymbol{\beta}(\boldsymbol{\theta}_{\tau,\mathbf{u}}) = \Omega_{\mathbf{X}}^{-1}\mathbb{E}[\mathbf{X}IF'(\mathbf{Y};\boldsymbol{\theta}_{\tau,\mathbf{u}})]. \qquad (28)$$

Also, let us denote for $i = 1,\ldots,n$,

$$\boldsymbol{\gamma}_i^{\star}(\boldsymbol{\theta}_{\tau,\mathbf{u}}) = \mathrm{vec}\left(\Omega_{\mathbf{X}}^{-1}\mathbf{X}_i\mathbf{z}'_i(\boldsymbol{\theta}_{\tau,\mathbf{u}})\right) \quad \text{and} \quad \mathbf{z}_i(\boldsymbol{\theta}_{\tau,\mathbf{u}}) = IF(\mathbf{Y}_i;\boldsymbol{\theta}_{\tau,\mathbf{u}}) - \boldsymbol{\beta}'(\boldsymbol{\theta}_{\tau,\mathbf{u}})\mathbf{X}_i. \qquad (29)$$

where the $\mathrm{vec}(\cdot)$ operator converts a matrix into a column vector by stacking its columns on top of one another. Finally, we define:

$$\overset{\star}{\boldsymbol{\alpha}}_{\tau,\mathbf{u}} = \mathrm{vec}(\boldsymbol{\alpha}_{\tau,\mathbf{u}}) \quad \text{and} \quad \boldsymbol{\alpha}^{\star}_{\tau,\mathbf{u}} = \mathrm{vec}(\boldsymbol{\alpha}_{\tau,\mathbf{u}}). \qquad (30)$$

Theorem 3. *Let assumptions A1-A6 hold. Then, for any* $\tau \in (0,\frac{1}{2}]$ *and* $\mathbf{u} \in \mathcal{S}^{p-1}$*, the following asymptotic linear representation holds:*

$$\sqrt{n}(\overset{\star}{\boldsymbol{\alpha}}_{\tau,\mathbf{u}} - \boldsymbol{\alpha}^{\star}_{\tau,\mathbf{u}}) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{S}_i(\boldsymbol{\theta}_{\tau,\mathbf{u}}) + o_p(1) \qquad (31)$$

and

$$\sqrt{n}(\overset{\star}{\boldsymbol{\alpha}}_{\tau,\mathbf{u}} - \boldsymbol{\alpha}^{\star}_{\tau,\mathbf{u}}) \xrightarrow{p} \mathcal{N}\left(\mathbf{0}, \mathbb{E}\left[\mathbf{S}(\boldsymbol{\theta}_{\tau,\mathbf{u}})\mathbf{S}'(\boldsymbol{\theta}_{\tau,\mathbf{u}})\right]\right) \quad \text{as} \quad n \to \infty, \qquad (32)$$

where $\mathbf{S}_i(\boldsymbol{\theta}_{\tau,\mathbf{u}})$ *is a kp-dimensional vector:*

$$\mathbf{S}_i(\boldsymbol{\theta}_{\tau,\mathbf{u}}) = \nabla_{\theta_{\tau,\mathbf{u}}}\boldsymbol{\beta}^{\star}(\boldsymbol{\theta})\mathbf{M}(\boldsymbol{\theta}_{\tau,\mathbf{u}})^{-1}\eta_{\delta}(\varphi_i)\Psi(\mathbf{Y}_i - \boldsymbol{\theta}_{\tau,\mathbf{u}}) + \boldsymbol{\gamma}_i^{\star}(\boldsymbol{\theta}_{\tau,\mathbf{u}}), \quad (33)$$

Accepted Manuscript

and $\nabla_{\theta_{\tau,\mathbf{u}}}\boldsymbol{\beta}^{\star}(\boldsymbol{\theta})$ *is the derivative of* $\boldsymbol{\beta}^{\star}(\boldsymbol{\theta})$ *with respect to* $\boldsymbol{\theta}$ *evaluated at* $\boldsymbol{\theta}_{\tau,\mathbf{u}}$.

Proof. See Proof of Theorem 3 in the Supplementary Materials. □

In addition, when multiple directions are of interest, we may estimate the proposed model at different directions simultaneously to gain better estimation accuracy by incorporating the associations among the considered directions. Hence, by proceeding as in the proof of Theorem 3 and applying the multivariate Central Limit Theorem to the Bahadur representation in (31), we derive the asymptotic distribution of the UMQPE estimator when multiple directions are considered jointly, as shown in the following remark.

Remark 1. Let $\{\mathbf{u}_1,\ldots,\mathbf{u}_J\}\subset\mathcal{S}^{p-1}$, with J being a fixed positive integer. Suppose assumptions A1-A6 hold, then the joint asymptotic distribution of $\sqrt{n}\left(\boldsymbol{\alpha}^{\star}_{\tau,\mathbf{u}_1}-\boldsymbol{\alpha}^{\star}_{\tau,\mathbf{u}_1},\ldots,\boldsymbol{\alpha}^{\star}_{\tau,\mathbf{u}_J}-\boldsymbol{\alpha}^{\star}_{\tau,\mathbf{u}_J}\right)$ is Gaussian with zero mean and the asymptotic covariance matrix between $\sqrt{n}(\boldsymbol{\alpha}^{\star}_{\tau,\mathbf{u}_r}-\boldsymbol{\alpha}^{\star}_{\tau,\mathbf{u}_r})$ and $\sqrt{n}(\boldsymbol{\alpha}^{\star}_{\tau,\mathbf{u}_s}-\boldsymbol{\alpha}^{\star}_{\tau,\mathbf{u}_s})$, where $1\le r,s\le J$, will be given by $\mathbb{E}\left[\mathbf{S}(\boldsymbol{\theta}_{\tau,\mathbf{u}_r})\mathbf{S}'(\boldsymbol{\theta}_{\tau,\mathbf{u}_s})\right]$.

In order to use Theorem 3 to build confidence intervals and hypothesis tests, in what follows we provide a consistent estimator of the asymptotic covariance matrix of $\boldsymbol{\alpha}^{\star}_{\tau,\mathbf{u}}$ in (32). The analytical form of the asymptotic covariance matrix suggests the following estimator:

$$\mathbf{V}(\boldsymbol{\theta}_{\tau,\mathbf{u}})=\frac{1}{n}\sum_{i=1}^{n}\mathbf{S}_i(\boldsymbol{\theta}_{\tau,\mathbf{u}})\mathbf{S}'_i(\boldsymbol{\theta}_{\tau,\mathbf{u}}) \qquad (34)$$

where $\mathbf{S}_i(\boldsymbol{\theta}_{\tau,\mathbf{u}})$ is the *kp*-dimensional vector:

Accepted Manuscript

$$\mathbf{S}_i(\boldsymbol{\theta}_{\tau,\mathbf{u}}) = \left(\nabla_{\boldsymbol{\theta}_{\tau,\mathbf{u}}} \boldsymbol{\beta}^{\star}(\boldsymbol{\theta})\mathbf{M}(\boldsymbol{\theta}_{\tau,\mathbf{u}})^{-1}\eta_{\delta}(\hat{\varphi}_i)\Psi(\mathbf{Y}_i - \boldsymbol{\theta}_{\tau,\mathbf{u}}) + \hat{\boldsymbol{\gamma}}_i^{\star}(\boldsymbol{\theta}_{\tau,\mathbf{u}})\right) \qquad (35)$$

and $\nabla_{\boldsymbol{\theta}_{\tau,\mathbf{u}}} \boldsymbol{\beta}^{\star}(\boldsymbol{\theta})$ can be obtained via numerical differentiation.

In order to establish consistency of the estimator in (34) we impose the following additional assumption.

(A7) $\mathbf{S}(\boldsymbol{\theta}_{\tau,\mathbf{u}})$ is continuous at $\boldsymbol{\theta}_{\tau,\mathbf{u}}$ and there exists a neighborhood of $\boldsymbol{\theta}_{\tau,\mathbf{u}}$, $\mathcal{I}$, such that for any $\boldsymbol{\theta}$ in this region, it holds that $\mathbb{E}[\sup_{\boldsymbol{\theta}\in\mathcal{I}} \| \mathbf{S}(\boldsymbol{\theta})\mathbf{S}'(\boldsymbol{\theta}) \|] < \infty$.

Then, the next Theorem proves consistency of $\mathbf{V}(\boldsymbol{\theta}_{\tau,\mathbf{u}})$.

Theorem 4. *Let assumptions A1-A7 hold,*

$$\mathbf{V}(\boldsymbol{\theta}_{\tau,\mathbf{u}}) - \mathbb{E}\left[\mathbf{S}(\boldsymbol{\theta}_{\tau,\mathbf{u}})\mathbf{S}'(\boldsymbol{\theta}_{\tau,\mathbf{u}})\right] \xrightarrow{p} 0, \qquad (36)$$

where the notation is understood to indicate convergence of the matrices element by element.

Proof. See Proof of Theorem 4 in the Supplementary Materials. □

4 Application

Accepted Manuscript

In this section, we consider data from the Survey on Household Income and Wealth (SHIW) 2016 conducted by the Bank of Italy to show the relevance of our methodology. We analyze the impact of economic and socio-demographic factors on households wealth and consumption levels, accounting both for the presence of outliers and the correlation structure between the two outcomes. We are interested in evaluating whether these effects are more pronounced on more disadvantaged families than on richer ones. In this setting, the limitation of using a conditional (M-)quantile model is that the effect of the covariates at different quantile levels may be masked by the set of conditioning variables, i.e., the characteristics of the family. Once we have conditioned on the explanatory variables, for instance, the 10-th (M-)quantile of the unconditional distribution of the responses may potentially be very different from the 10-th (M-)quantile of the conditional distribution, so the coefficients of conditional (M-)quantile regression cannot be interpreted as unconditional effects. On the other hand, by using our unconditional method, the UMQPE provides an estimate of the impact of covariates across the entire population and not merely among population subgroups, consisting of families who share the same values of the included covariates. In what follows, we fit the proposed regression method at different points of the unconditional distributions of family wealth and consumption, and illustrate the difference between the conditional and unconditional approaches.

4.1 Data description

The SHIW (https://www.bancaditalia.it) is an annual survey conducted by the Bank of Italy whose aims are to provide information on the economic and financial behaviours of Italian households and collect reliable, comparable and representative data of the population resident in Italy. This survey is widely regarded as the basis of the most reliable estimates for macroeconomics studies. The sample is drawn in two stages, the primary and secondary sampling units are municipalities and households, respectively. Before the primary units are selected, they are stratified by region and population size. Data are collected mainly via an electronic

Accepted Manuscript

questionnaire using the Computer Assisted Personal Interviewing program while the remaining interviews are conducted using the Paper And Pencil Personal Interviewing program.

In this work, following established custom we transform the dependent variables, i.e., household consumption (LCON) and net wealth (LWEA) to natural logarithm. In particular, LCON is defined as the sum of household's expenditure on durables and non-durable goods while net wealth is obtained as the algebraic sum of real assets, financial assets and financial liabilities. The set of considered predictors includes the log of net disposable income (LINC), defined as the sum of payroll income, pensions, net transfers, net self-employment income and property income sources, and relevant information on the household's head such as age (Age) and age squared (Age2) measured in years. Also, gender (male (baseline)), marital (married (baseline), never married, separated, widowed), employment status (employee (baseline), self-employed, not-employed) and educational level (elementary (baseline), middle, vocational, high school, university or higher) are included as dummy variables. Finally, a categorical variable is included to investigate the presence of regional divergences in wealth and consumption levels depending on the region of residence (north (baseline), centre, south and islands). Table S7 in the Supplementary Materials summarizes the descriptive statistics of the included variables. The considered sample contains 6802 households.

As a preliminary step, we study the unconditional distributions of households wealth and consumption. The histograms of LWEA and LCON unconditional distributions in Figure S5 of the Supplementary Materials reveal that, while normality seems tenable for LCON, there are potentially influential observations in the distribution of LWEA and indicate a departure from the Gaussian assumption, having fat tails and pronounced asymmetries. Furthermore, the empirical correlation between LCON and LWEA equals to 0.473. As expected, consumption and wealth are positively correlated, justifying the need for a multivariate approach that considers these two dimensions together. The discreteness of LWEA that shows up especially at low values of wealth is due to the rounding effect of the interviews towards round figures (Groves et al. 2011). Consequently, the presented unconditional regression

Accepted Manuscript

model is appropriate to account for outlying observations and investigate how the relationship between responses and explanatory variables can vary across the unconditional distribution of family wealth and consumption.

4.2 Modeling household wealth and consumption

We analyze the SHIW 2016 data to jointly model households' log-wealth, LWEA, and log-consumption, LCON, as a function of the predictors in Table S7. We fit the proposed approach at levels $\tau = 0.10, \tau = 0.50$ and $\tau = 0.90$, which can estimated by simply noting that for $\tau \in (0, \frac{1}{2}], \boldsymbol{\theta}_{1-\tau,\mathbf{u}} = \boldsymbol{\theta}_{\tau,-\mathbf{u}}$ (see Kokic et al. 2002). As explained in Section 3, a meaningful direction $\mathbf{u}$ shall be determined for ordering multivariate observations, taking into account the problem under consideration. A possible way to do this is to observe that ratios of the type $Z = \log\left(\frac{Y_1^{u_1}}{Y_2^{-u_2}}\right) = \mathbf{u}'(\log Y_1, \log Y_2)$ bears great importance in several fields like finance, economics and growth models, with Y_1 and Y_2 representing wealth and consumption at the household level, respectively. In order to explicitly account for the existing positive correlation between wealth and consumption, we use the direction $\mathbf{u} = (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})'$ throughout the rest of the section (the analysis with $\mathbf{u} = (\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}})'$ is shown in the Supplementary Materials). To give a graphical intuition on the use of such direction, Figure S6 in the Supplementary Materials shows the scatter plots of family wealth and consumption where $\mathbf{u}$ is represented by the green arrow, while the blue points denote the multivariate quantile $\boldsymbol{\theta}_{\tau,\mathbf{u}}$ with $c = 0$ at $\tau = 0.10$ (left) and $\tau = 0.90$ (left) as an example. The black dashed lines correspond to the hyperplanes orthogonal to $\mathbf{u}$ and passing through $\boldsymbol{\theta}_{\tau,\mathbf{u}}$ where datapoints in the upper half-plane are shown in red while those in the lower half-plane are shown in gray. In doing so, the sample space is divided into two regions by the hyperplane orthogonal to the selected direction and any observation in the region in

Accepted Manuscript

direction $\mathbf{u}$ from $\boldsymbol{\theta}_{\tau,\mathbf{u}}$ receives a certain weight, while observations in the opposite direction receive a different weight depending both on their length and the angle they form with $\mathbf{u}$, according to the weighting function $\eta_{\delta}(\varphi)$ below (7). In particular, at $\tau = 0.10$, households below the half-plane generally have low or moderate levels of consumption and wealth, and a certain weight will be assigned to them compared to wealthier and high spender families located in the other region. When $\tau = 0.90$, the multivariate quantile shifts upwards to the right and thus also the separating hyperplane so that it distinguishes families with jointly high consumption and wealth patterns who will be weighted differently than all those below as shown in gray.

The estimation of the optimal tuning constant $c^{\star}$ is obtained using a 5-fold cross-validation over a sequence of 200 possible values as described in Section 3.1. The results are reported in Table 1 ($c = c^{\star}$) where we compare the proposed unconditional regression with the standard conditional regression approach. In particular, the left-hand side columns (labeled as Unconditional Regression) refer to the estimates of the UMQPE, $\boldsymbol{\alpha}_{\tau,\mathbf{u}}$, obtained as described in (21). Conversely, the right-hand columns (labeled as Conditional Regression) refer to the estimates of the coefficients $\boldsymbol{\beta}_{\tau,\mathbf{u}}$ of a multivariate linear conditional M-quantile regression, i.e., $\boldsymbol{\theta}_{\tau,\mathbf{u}}(\mathbf{x}) = \boldsymbol{\beta}'_{\tau,\mathbf{u}}\mathbf{x}$, obtained using the IRLS algorithm mentioned in Section 3. To display the sampling variation, asymptotic standard errors obtained using the results in Section 3.2 are presented in parentheses. Parameter estimates are displayed in boldface when significant at the standard 5% level.

The main findings can be summarised as follows. The selected values of *c* are 1.373, 10.966 and 10.378 at level 0.10, 0.50 and 0.90, respectively. This implies that the estimates are close to the quantile case at $\tau = 0.10$ as more than 77% of residuals are down-weighted (Huberised) meanwhile, at $\tau = 0.50$ and $\tau = 0.90$ correspond to expectile estimation as no observations are Huberised. The estimated values for *c* reflect the negatively skewed distribution of wealth and support the exploratory analysis in Figure S5.

Accepted Manuscript

Because the selected values of c lead towards expectile estimation, especially above $\tau = 0.50$, we also consider the case $c = 0$ (see Table S8 of the Supplementary Materials) which allows us to estimate the impact of the covariates on the unconditional multivariate quantiles of the response. Comparing Tables 1 and S8, slight differences in terms of estimation can be found. This is attributable to the choice of c as the two models allow to target different population parameters by selecting different values for the tuning constant of the Huber function. The above points demonstrate the flexibility of the methodology proposed to extend the classical OLS regression for assessing the effect of covariates, not only at the center, but also at different parts of the unconditional distribution of interest.

Nevertheless, there are still similar results between Tables 1 and S8. Point estimates generally increase in magnitude when moving outward from the bulk of the data. Income elasticity is positively associated with both wealth and consumption for all investigated τ levels. One can see that there are small differences in consumption expenditure among males and females. Moreover, education, marital and employment status are important determinants of family's consumption and wealth levels, with an increasing trend as $\tau \to 0$. There also appears to be significant regional disparities across the distributions of the responses as southern regions and islands generally have lower levels of wealth and consumption. It is important to note that the effect of education and marital status between the conditional and unconditional models is very different, especially at $c = 0$. As shown in Theorem 1, this corresponds to the case where large differences exist between the UMQPE and the CMQPE. This may be due to the fact that the matching, $s_{\tau,\mathbf{u}}(\mathbf{X})$, and weighting, $\mathbf{W}_{\tau,\mathbf{u}}(\mathbf{X})$, functions vary across the values of $\mathbf{X}$, which means that the conditional effects do not average up to their respective unconditional effects. By contrast, the conditional and unconditional models provide similar estimates for LINC, age and employment status at $c = c^{\star}$ in particular, suggesting that $s_{\tau,\mathbf{u}}(\mathbf{X}) \approx \tau$ does not vary very much for all values of $\mathbf{X}$. In this case, we have that $\boldsymbol{\alpha}_{\tau,\mathbf{u}} = \mathbb{E}[\mathbf{W}_{\tau,\mathbf{u}}(\mathbf{X})\boldsymbol{\alpha}_{s_{\tau,\mathbf{u}}(\mathbf{X}),\mathbf{u}}(\mathbf{X})] \approx \mathbb{E}[\mathbf{W}_{\tau,\mathbf{u}}(\mathbf{X})\boldsymbol{\alpha}_{\tau,\mathbf{u}}(\mathbf{X})]$. Under the considered linear model targeting the conditional multivariate M-quantile $\boldsymbol{\theta}_{\tau,\mathbf{u}}(\mathbf{x})$, the CMQPE, $\boldsymbol{\alpha}_{\tau,\mathbf{u}}(\mathbf{x}) = \boldsymbol{\beta}_{\tau,\mathbf{u}}$, which implies that $\boldsymbol{\alpha}_{\tau,\mathbf{u}} \approx \boldsymbol{\beta}_{\tau,\mathbf{u}}$.

Accepted Manuscript

As a means of further comparison, we analyze the log-levels of wealth and consumption with the UQR of Firpo et al. (2009) by fitting two univariate models independently. Table S9 of the Supplementary Materials reports the corresponding parameter estimates and standard errors at the examined τ levels. We can observe that the results are generally in line with the estimates from the proposed multivariate unconditional quantile regression in Table S8. However, some differences can be identified at $\tau = 0.10$ due to the selected direction $\mathbf{u}$ and the fact that the univariate UQR completely disregards the dependence between wealth and consumption. By contrast, the proposed model allows to study the direction and magnitude of such correlation at different levels τ. In particular, using (13) we represent in Tables 1 and S8 the estimated correlation coefficient, r_{12}, which indicates that consumption and wealth are strongly correlated with each other and this association slightly decreases for households at the upper end of the responses distribution. At $\tau = 0.10$, the estimated coefficients (0.324 and 0.308) suggest that low-consumption households are likely to be accompanied by low wealth, with an increasing pattern as we move towards the $\tau = 0.50$ and $\tau = 0.90$ levels.

We conclude the analysis by using the nonparametric estimator of the UMQPE described in Section 3.1 to take into account possible nonlinear effects of the covariates. That is, we model the $\mathbb{E}[RIF(\mathbf{Y};\boldsymbol{\theta}_{\tau,\mathbf{u}}) \mid \mathbf{X}]$ including cubic B-splines specified for income and age. To test the linearity assumption between the RIF and the covariates, we conduct a goodness-of-fit test based on the Pillai' s Trace test statistic (Rencher and Christensen 2012, see Table S11 in the Supplementary Materials). The results provides evidence that the model with B-splines fits better than the linear specification for the RIF regression. In Table 2 we thus show the estimated UMQPEs using the nonparametric method at $\tau = 0.10, \tau = 0.50$ and $\tau = 0.90$ for the optimal tuning constant $c^{\star}$, where 95% confidence intervals (in parentheses) are obtained via nonparametric bootstrap (Efron and Tibshirani 1994). By looking at the results, the estimated UMQPEs are similar, in terms of both sign and magnitude, to the estimates in Table 1 where the RIF is modeled as a linear function of the covariates.

Accepted Manuscript

Finally, in Section 6 of the Supplementary Materials we provide a graphical representation of how well the considered approach can approximate the effect of more general changes in distribution of the covariates such as those contemplated in the policy effect (Firpo et al. 2007).

5 Conclusions

Extending the univariate work of Firpo et al. (2009), this paper proposes a unified approach to model the entire unconditional distribution of a multivariate response variable in a regression setting. We make several contributions to the literature. First, by employing the multidimensional Huber's function in Hampel et al. (2011) we are able to build a comprehensive modeling framework to estimate multivariate unconditional quantiles, M-quantiles and expectiles, choosing the tuning constant in an appropriate manner. Second, in contrast to univariate methods, our multivariate model accounts for the, potentially asymmetric, association structure between the outcome variables. Third, the proposed methodology is easy to implement through an OLS regression of the RIF on the explanatory variables. From a theoretical standpoint, we show that the introduced estimators are consistent, asymptotically normal and can be written as a weighted average of conditional effects. In addition, we propose a data-driven procedure based on cross-validation to select the optimal tuning constant for estimating the UMQPE and the policy effect. Finally, we contribute to the empirical literature by analyzing log-levels of wealth and consumption of Italian households collected in the SHIW 2016 data.

Possible future developments of this work are as follows. First, the independence assumption in Theorem 1 can be relaxed by introducing in the RIF regression a control function constructed using instrumental variables to account for endogenous covariates. Second, the extension of this result to the case of heteroskedastic errors would be an interesting topic for future research. Lastly, it would also be interesting the study of the theoretical behavior of the multivariate M-quantile cross-validated estimator when selecting the tuning constant of the Huber's influence function.

Accepted Manuscript

SUPPLEMENTARY MATERIALS

Simulations, additional results and proofs: Simulation studies, additional results and technical derivations that are used to support the results in the manuscript. (PDF file)

FUNDING

Competing interests: The authors report there are no competing interests to declare.

References

Alejo, J., F. Favata, G. Montes-Rojas, and M. Trombetta (2021). Conditional vs unconditional quantile regression models: A guide to practitioners. *Economía 44* (88), 76–93.

Bahadur, R. R. (1966). A note on quantiles in large samples. *The Annals of Mathematical Statistics 37* (3), 577–580.

Bianchi, A., E. Fabrizi, N. Salvati, and N. Tzavidis (2018). Estimation and testing in M-quantile regression with applications to small area estimation. *International Statistical Review 86* (3), 541–570.

Accepted Manuscript

Borah, B. J. and A. Basu (2013). Highlighting differences between conditional and unconditional quantile regression approaches through an application to assess medication adherence. *Health Economics 22* (9), 1052–1070.

Breckling, J. and R. Chambers (1988). M-quantiles. *Biometrika 75* (4), 761–771.

Breckling, J., P. Kokic, and O. Lübke (2001). A note on multivariate M-quantiles. *Statistics & Probability Letters 55* (1), 39–44.

Chaudhuri, P. (1996). On a geometric notion of quantiles for multivariate data. *Journal of the American Statistical Association 91* (434), 862–872.

Efron, B. and R. J. Tibshirani (1994). *An introduction to the bootstrap*. CRC press.

Farcomeni, A., M. Geraci, and C. Viroli (2022). Directional quantile classifiers. *Journal of Computational and Graphical Statistics* , 1–10.

Firpo, S., N. M. Fortin, and T. Lemieux (2007, July). Unconditional Quantile Regressions. NBER Technical Working Papers 0339, National Bureau of Economic Research, Inc.

Firpo, S., N. M. Fortin, and T. Lemieux (2009). Unconditional quantile regressions. *Econometrica: Journal of the Econometric Society 77* (3), 953–973.

Firpo, S. P., N. M. Fortin, and T. Lemieux (2018). Decomposing wage distributions using recentered influence function regressions. *Econometrics 6* (2), 28.

Accepted Manuscript

Fraiman, R. and B. Pateiro-López (2012). Quantiles for finite and infinite dimensional data. *Journal of Multivariate Analysis 108*, 1–14.

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

Geraci, M., N. S. Boghossian, A. Farcomeni, and J. D. Horbar (2020). Quantile contours and allometric modelling for risk classification of abnormal ratios with an application to asymmetric growth-restriction in preterm infants. *Statistical methods in medical research 29* (7), 1769–1786.

Girard, S. and G. Stupfler (2017). Intriguing properties of extreme geometric quantiles. *REVSTAT-Statistical Journal 15* (1), 107–139.

Groves, R. M., F. J. Fowler Jr, M. P. Couper, J. M. Lepkowski, E. Singer, and R. Tourangeau (2011). *Survey methodology*. John Wiley & Sons.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association 69* (346), 383–393.

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

Accepted Manuscript

Heckman, J. J. and E. J. Vytlacil (2007). Econometric evaluation of social programs, part i: Causal models, structural models and econometric policy evaluation. *Handbook of Econometrics 6*, 4779–4874.

Herrmann, K., M. Hofert, and M. Mailhot (2018). Multivariate geometric expectiles. *Scandinavian Actuarial Journal 2018* (7), 629–659.

Huber, P. and E. Ronchetti (2009). *Robust Statistics*. Wiley.

Huber, P. J. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics 35* (1), 73–101.

Huffman, M. L., J. King, and M. Reichelt (2017). Equality for whom? Organizational policies and the gender gap across the German earnings distribution. *ILR Review 70* (1), 16–41.

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

Koenker, R. and G. Bassett Jr (1978). Regression quantiles. *Econometrica: Journal of the Econometric Society* , 33–50.

Koenker, R., V. Chernozhukov, X. He, and L. Peng (2017). *Handbook of Quantile Regression*. CRC press.

Kokic, P., J. Breckling, and O. Lübke (2002). A new definition of multivariate M-quantiles. In *Statistical data analysis based on the L_1-norm and related methods*, pp. 15–24. Springer.

Konen, D. and D. Paindaveine (2021). Multivariate rho-quantiles: a spatial approach. *Bernoulli* .

Accepted Manuscript

Kong, L. and I. Mizera (2012). Quantile tomography: using quantiles with multivariate data. *Statistica Sinica* , 1589–1610.

Machado, J. A. and J. Mata (2005). Counterfactual decomposition of changes in wage distributions using quantile regression. *Journal of Applied Econometrics 20* (4), 445–465.

Maclean, J. C., D. A. Webber, and J. Marti (2014). An application of unconditional quantile regression to cigarette taxes. *Journal of Policy Analysis and Management 33* (1), 188–210.

Melly, B. (2005). Decomposition of differences in distribution using quantile regression. *Labour Economics 12* (4), 577–590.

Merlo, L., L. Petrella, N. Salvati, and N. Tzavidis (2022). Marginal M-quantile regression for multivariate dependent data. *Computational Statistics & Data Analysis* , 107500.

Mises, R. v. (1947). On the asymptotic distribution of differentiable statistical functions. *The Annals of Mathematical Statistics 18* (3), 309–348.

Newey, W. K. (1994). The asymptotic variance of semiparametric estimators. *Econometrica: Journal of the Econometric Society* , 1349–1382.

Newey, W. K. and J. L. Powell (1987). Asymmetric least squares estimation and testing. *Econometrica* , 819–847.

Paindaveine, D. and M. Šiman (2011). On directional multiple-output quantile regression. *Journal of Multivariate Analysis 102* (2), 193–212.

Accepted Manuscript

Petrella, L. and V. Raponi (2019). Joint estimation of conditional quantiles in multivariate linear regression models with an application to financial distress. *Journal of Multivariate Analysis 173*, 70–84.

Rencher, A. C. and W. F. Christensen (2012). *Methods of multivariate analysis*. Wiley.

Rios-Avila, F. (2020). Recentered influence functions (RIFs) in Stata: RIF regression and RIF decomposition. *The Stata Journal 20* (1), 51–94.

Serfling, R. (2002). Quantile functions for multivariate analysis: approaches and applications. *Statistica Neerlandica 56* (2), 214–232.

Torres, R., C. De Michele, H. Laniado, and R. E. Lillo (2017). Directional multivariate extremes in environmental phenomena. *Environmetrics 28* (2), e2428.

Tzavidis, N., N. Salvati, M. Geraci, and M. Bottai (2010). M-quantile and expectile random effects regression for multilevel data.

Wang, Y.-G., X. Lin, M. Zhu, and Z. Bai (2007). Robust estimation using the Huber function with a data-dependent tuning constant. *Journal of Computational and Graphical Statistics 16* (2), 468–481.

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

Accepted Manuscript

Table 1 Unconditional and conditional regression coefficient estimates obtained from the linear specification for the RIF at the investigated τ levels and direction $\mathbf{u} = (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})'$, using the optimal tuning constant $c = c^{\star}$. Parameter estimates are displayed in boldface when significant at the standard 5% level.

	Unconditional Regression						Conditional Regression					
τ	0.10		0.50		0.90		0.10		0.50		0.90	
Variable	W	C	W	C	W	C	W	C	W	C	W	C
Intercept	**–29.370**	**5.960**	**–5.701**	**5.472**	–0.514	**4.742**	**–15.840**	**4.348**	**–5.701**	**5.472**	0.003	**6.156**
	(1.243)	(0.159)	(0.390)	(0.095)	(0.302)	(0.188)	(0.529)	(0.121)	(0.389)	(0.094)	(0.297)	(0.122)
LINC	**2.976**	**0.410**	**1.315**	**0.435**	**0.993**	**0.496**	**2.156**	**0.547**	**1.315**	**0.435**	**0.914**	**0.361**
	(0.105)	(0.016)	(0.032)	(0.008)	(0.025)	(0.015)	(0.044)	(0.010)	(0.032)	(0.008)	(0.025)	(0.010)
Sex	0.044	**–0.035**	0.008	**–0.028**	–0.001	–0.011	–0.022	**–0.037**	0.008	**–0.028**	0.016	**–0.030**
	(0.134)	(0.016)	(0.045)	(0.011)	(0.033)	(0.017)	(0.061)	(0.014)	(0.045)	(0.011)	(0.034)	(0.014)
Age	**0.219**	**–0.014**	**0.081**	0.001	**0.055**	**0.012**	**0.096**	–0.003	**0.081**	0.001	**0.058**	**0.007**
	(0.025)	(0.003)	(0.008)	(0.002)	(0.006)	(0.003)	(0.011)	(0.003)	(0.008)	(0.002)	(0.006)	(0.003)
Age2	**–0.002**	**0.000**	**–0.001**	–0.000	**–0.000**	**–0.000**	**–0.001**	0.000	**–0.001**	–0.000	**–0.000**	**–0.000**
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Marital status												

Accepted Manuscript

	Unconditional Regression						Conditional Regression					
never married	**0.523**	**–0.130**	0.094	**–0.152**	0.006	**–0.158**	**0.348**	**–0.101**	0.094	**–0.152**	0.010	**–0.175**
	(0.170)	(0.023)	(0.056)	(0.014)	(0.042)	(0.021)	(0.077)	(0.018)	(0.056)	(0.014)	(0.043)	(0.018)
separated	**–0.845**	**–0.056**	**–0.303**	**–0.104**	**–0.189**	**–0.141**	**–0.250**	–0.041	**–0.303**	**–0.104**	**–0.239**	**–0.156**
	(0.214)	(0.026)	(0.071)	(0.017)	(0.053)	(0.027)	(0.097)	(0.022)	(0.071)	(0.017)	(0.054)	(0.022)
widowed	**0.408**	**–0.098**	0.067	**–0.098**	–0.008	**–0.084**	**0.323**	–0.029	0.067	**–0.098**	–0.032	**–0.135**
	(0.191)	(0.025)	(0.063)	(0.015)	(0.047)	(0.024)	(0.086)	(0.020)	(0.063)	(0.015)	(0.048)	(0.020)
Education level												
middle school	**0.434**	**0.125**	**0.244**	**0.102**	**0.173**	**0.059**	**0.274**	**0.085**	**0.244**	**0.102**	**0.204**	**0.118**
	(0.174)	(0.021)	(0.058)	(0.014)	(0.043)	(0.022)	(0.079)	(0.018)	(0.058)	(0.014)	(0.044)	(0.018)
vocational school	0.398	**0.133**	**0.299**	**0.117**	**0.278**	**0.065**	**0.303**	**0.064**	**0.299**	**0.117**	**0.296**	**0.165**
	(0.242)	(0.029)	(0.081)	(0.020)	(0.060)	(0.031)	(0.110)	(0.025)	(0.081)	(0.020)	(0.062)	(0.025)
high school	**1.006**	**0.156**	**0.611**	**0.189**	**0.510**	**0.191**	**0.550**	**0.128**	**0.611**	**0.189**	**0.548**	**0.246**
	(0.189)	(0.024)	(0.063)	(0.015)	(0.047)	(0.024)	(0.086)	(0.020)	(0.063)	(0.015)	(0.048)	(0.020)
university	**0.518**	**0.126**	**0.656**	**0.276**	**0.849**	**0.440**	**0.429**	**0.169**	**0.656**	**0.276**	**0.670**	**0.379**
	(0.229)	(0.028)	(0.076)	(0.019)	(0.057)	(0.030)	(0.104)	(0.024)	(0.076)	(0.019)	(0.058)	(0.024)
Employment status												
self–employed	**1.791**	**–0.071**	**0.942**	0.032	**0.947**	**0.153**	**0.842**	–0.028	**0.942**	0.032	**0.812**	**0.066**

Accepted Manuscript

	Unconditional Regression						Conditional Regression					
	(0.206)	(0.025)	(0.068)	(0.017)	(0.051)	(0.027)	(0.093)	(0.021)	(0.068)	(0.017)	(0.052)	(0.021)
not–employed	**1.524**	–0.001	**0.654**	**0.041**	**0.575**	**0.111**	**0.658**	0.021	**0.654**	**0.041**	**0.582**	**0.045**
	(0.180)	(0.023)	(0.060)	(0.015)	(0.045)	(0.023)	(0.082)	(0.019)	(0.060)	(0.015)	(0.046)	(0.019)
Geographical area												
centre	**0.681**	**0.046**	**0.213**	**0.024**	**0.080**	–0.001	**0.325**	**0.056**	**0.213**	**0.024**	**0.107**	0.005
	(0.143)	(0.017)	(0.048)	(0.012)	(0.036)	(0.018)	(0.065)	(0.015)	(0.048)	(0.012)	(0.036)	(0.015)
south and islands	**1.015**	**–0.073**	**0.196**	**–0.102**	0.017	**–0.100**	**0.359**	**–0.055**	**0.196**	**–0.102**	0.055	**–0.154**
	(0.133)	(0.019)	(0.044)	(0.011)	(0.033)	(0.017)	(0.060)	(0.014)	(0.044)	(0.011)	(0.034)	(0.014)
r_{12}	0.324		0.473		0.653							

Accepted Manuscript

Table 2 Unconditional and conditional regression coefficient estimates obtained from the nonparametric method for the RIF at the investigated τ levels and direction $\mathbf{u} = (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})'$, using the optimal tuning constant $c = c^{\star}$. Standard errors are computed via nonparametric bootstrap using 1000 resamples and parameter estimates are displayed in boldface when significant at the standard 5% level.

	Unconditional Regression						Conditional Regression					
τ	0.10		0.50		0.90		0.10		0.50		0.90	
Variable	W	C	W	C	W	C	W	C	W	C	W	C
Intercept	–0.107	**11.293**	**7.275**	**9.798**	**7.230**	**7.708**	**5.450**	**9.264**	**7.275**	**9.798**	**6.342**	**9.943**
	(6.859)	(1.449)	(1.669)	(0.443)	(0.874)	(1.060)	(1.669)	(0.443)	(1.669)	(0.443)	(0.766)	(0.401)
LINC	**3.383**	**0.463**	**1.662**	**0.587**	**1.416**	**0.765**	**2.254**	**0.551**	**1.662**	**0.587**	**1.340**	**0.639**
	(0.343)	(0.076)	(0.083)	(0.034)	(0.182)	(0.140)	(0.083)	(0.034)	(0.083)	(0.034)	(0.089)	(0.047)
Sex	0.114	–0.017	0.028	**–0.023**	–0.006	–0.021	–0.006	**–0.039**	0.028	**–0.023**	0.042	–0.015
	(0.135)	(0.016)	(0.045)	(0.011)	(0.033)	(0.016)	(0.045)	(0.011)	(0.045)	(0.011)	(0.037)	(0.013)
Age	**0.060**	**0.046**	**0.112**	**–0.007**	**0.096**	**–0.006**	**0.223**	**0.043**	**0.112**	**–0.007**	**–0.024**	**–0.045**
	(0.002)	(0.000)	(0.001)	(0.000)	(0.000)	(0.000)	(0.002)	(0.000)	(0.002)	(0.000)	(0.002)	(0.001)
Marital status												
never married	**0.634**	**–0.123**	**0.202**	**–0.102**	**0.152**	**–0.059**	**0.342**	**–0.097**	**0.202**	**–0.102**	**0.119**	**–0.096**

Accepted Manuscript

	Unconditional Regression						Conditional Regression					
	(0.176)	(0.022)	(0.058)	(0.013)	(0.038)	(0.018)	(0.058)	(0.013)	(0.058)	(0.013)	(0.046)	(0.014)
separated	**–0.676**	–0.029	–0.149	**–0.037**	–0.003	–0.027	**–0.193**	**–0.039**	–0.149	**–0.037**	–0.079	**–0.042**
	(0.239)	(0.028)	(0.079)	(0.017)	(0.052)	(0.023)	(0.079)	(0.017)	(0.079)	(0.017)	(0.063)	(0.019)
widowed	**0.520**	**–0.083**	**0.184**	**–0.046**	**0.144**	0.012	**0.359**	–0.022	**0.184**	**–0.046**	0.089	**–0.052**
	(0.191)	(0.023)	(0.061)	(0.015)	(0.044)	(0.021)	(0.061)	(0.015)	(0.061)	(0.015)	(0.047)	(0.017)
Education level												
middle school	0.339	**0.109**	**0.193**	**0.080**	**0.128**	**0.031**	**0.216**	**0.084**	**0.193**	**0.080**	**0.151**	**0.077**
	(0.192)	(0.023)	(0.062)	(0.013)	(0.040)	(0.016)	(0.062)	(0.013)	(0.062)	(0.013)	(0.046)	(0.015)
vocational school	0.268	**0.111**	**0.234**	**0.091**	**0.227**	0.036	**0.238**	**0.067**	**0.234**	**0.091**	**0.227**	**0.120**
	(0.246)	(0.029)	(0.082)	(0.019)	(0.056)	(0.026)	(0.082)	(0.019)	(0.082)	(0.019)	(0.065)	(0.024)
high school	**0.778**	**0.122**	**0.458**	**0.124**	**0.347**	**0.090**	**0.462**	**0.123**	**0.458**	**0.124**	**0.387**	**0.126**
	(0.192)	(0.022)	(0.064)	(0.015)	(0.045)	(0.021)	(0.064)	(0.015)	(0.064)	(0.015)	(0.049)	(0.017)
university	0.350	**0.125**	**0.392**	**0.151**	**0.444**	**0.168**	**0.424**	**0.154**	**0.392**	**0.151**	**0.333**	**0.143**
	(0.212)	(0.027)	(0.075)	(0.019)	(0.061)	(0.032)	(0.075)	(0.019)	(0.075)	(0.019)	(0.061)	(0.023)
Employment status												
self–employed	**1.974**	–0.016	**0.889**	–0.005	**0.748**	0.007	**0.932**	**–0.038**	**0.889**	–0.005	**0.794**	0.038
	(0.178)	(0.024)	(0.058)	(0.018)	(0.056)	(0.030)	(0.058)	(0.018)	(0.058)	(0.018)	(0.051)	(0.020)

Accepted Manuscript

	Unconditional Regression						Conditional Regression					
not–employed	1.950	0.070	0.685	0.043	0.437	0.026	0.676	0.032	0.685	0.043	0.617	0.062
	(0.230)	(0.026)	(0.074)	(0.016)	(0.051)	(0.024)	(0.074)	(0.016)	(0.074)	(0.016)	(0.061)	(0.019)
Geographical area												
centre	0.681	0.043	0.237	0.036	0.125	0.029	0.301	0.059	0.237	0.036	0.148	0.027
	(0.134)	(0.014)	(0.044)	(0.010)	(0.034)	(0.016)	(0.044)	(0.010)	(0.044)	(0.010)	(0.035)	(0.012)
south and islands	1.181	–0.045	0.299	–0.059	0.117	–0.040	0.353	–0.052	0.299	–0.059	0.194	–0.060
	(0.152)	(0.019)	(0.046)	(0.010)	(0.030)	(0.014)	(0.046)	(0.010)	(0.046)	(0.010)	(0.037)	(0.012)

Accepted Manuscript