Estimation and Testing in M-quantile Regression with application to small area estimation

Annamaria Bianchi
DSAEMQ, Università di Bergamo
and
Enrico Fabrizi
DISES,
Università Cattolica del S. Cuore
and
Nicola Salvati
DEM, Università di Pisa
and
Nikos Tzavidis
University of Southampton

August 4, 2015

Abstract

In recent years M-quantile regression has been applied to small area estimation with the aim to obtain reliable and outlier robust estimators without recourse to strong parametric assumptions. Nonetheless goodness-of-fit measures and testing related to model selection received so far little attention. In this paper we formally cover several topics related to estimation, model assessment and hypothesis testing

for M-quantile regression. In particular, a pseudo-R^2 goodness of fit measure is proposed, along with likelihood ratio and Wald type tests for linear hypotheses on the M-quantile regression parameters. A new estimator for the residuals scale based on a parametric representation of the M-quantile regression estimation is also proposed. The proposed parametric representation, which generalizes the Asymmetric Laplace distribution used in quantile regression, motivates the selection of data-driven tuning parameters associated with the loss function. Finally, a test for the presence of clustering in the data is also proposed. The properties of the tests are theoretically studied and their finite sample properties empirically assessed in Monte-Carlo simulations. The use of the proposed methods is illustrated in a well-known real data application in small area estimation field.

Keywords: Generalized Asymmetric Least Informative distribution; goodness-of-fit; likelihood ratio type test; loss function; robust regression

1 Introduction

Quantile regression (Koenker and Bassett, 1978; Koenker, 2005) represents a useful generalization of median regression whenever the interest is not limited to the estimation of a location parameter at the centre of the conditional distribution of the target variable y given a set of predictors $\mathbf{x}$ but extends to location parameters (quantiles) at other parts of this conditional distribution. Similarly, expectile regression (Newey and Powell, 1987) generalizes least squares regression at the centre of a distribution to estimation of location parameters at other parts of the target conditional distribution namely, expectiles. Breckling and Chambers (1988) introduce M-quantile regression that extends the ideas of M-estimation (Huber, 1964; Huber and Ronchetti, 2009) to a different set of location parameters of the target conditional distribution that lie between quantiles and expectiles. M-quantiles aim at combining the robustness properties of quantiles with the efficiency properties of expectiles.

Given a random variable y with cdf $F(y)$ and a (a.e.) continuously differentiable convex loss function $\rho(u)$, $u \in \mathcal{R}$, we define the tilted version of the loss function as

$$\rho_\tau(u) = |\tau - I(u < 0)|\rho(u), \tag{1}$$

with $\tau \in (0, 1)$. The τ-th M-quantile θ_τ is obtained as the minimizer of,

$$\int \rho_\tau(y - \theta_\tau) F(dy). \tag{2}$$

Depending on the choice of the loss function, M-quantiles may reduce to ordinary quantiles ($\rho(u) = |u|$) and expectiles ($\rho(u) = u^2$) while other choices are also possible (Dodge and Jureckova, 2000). However, as it is well known, quantiles and expectiles should be treated separately due to different properties of the corresponding influence functions. In regression

the argument in the loss functions is defined by standardized residuals $u = \sigma_\tau^{-1}(y - \mathbf{x}^T\boldsymbol{\beta}_\tau)$, where σ_τ is a scale parameter for the residuals' distribution.

Early applications of M-quantile regression include Breckling and Chambers (1988) and Kocic et al. (1997). Chambers and Tzavidis (2006) apply M-quantile regression in small domain prediction. The distinguishing features of their approach include the protection that a careful choice of $\rho(u)$ offers against the effect of outliers and the characterization of domain heterogeneity in terms of domain-specific M-quantiles. These can be viewed as an alternative to the random effects approach for measuring cluster-specific unobserved heterogeneity. A number of papers on M-quantile regression that focus on theoretical developments (Tzavidis et al., 2010; Bianchi and Salvati, 2015; Chambers et al., 2014a; Fabrizi et al., 2014a), extensions to non-linear models (Chambers et al., 2014b; Dreassi et al., 2014; Tzavidis et al., 2015) and various applications (Tzavidis et al., 2012; Fabrizi et al., 2014b) has been published in recent years. Nonetheless, in all these papers little attention has been paid so far to goodness-of-fit statistics and hypothesis testing related to model selection.

The main objective of this paper is to fill this gap in the literature. In particular, we propose a pseudo-R^2 goodness-of-fit measure and likelihood ratio and Wald type tests for linear hypotheses testing for the M-quantile regression parameters following the line of work proposed by Koenker and Machado (1999) for quantile regression. Although we assume that the loss functions belong to the large class of (a.e.) continuously differentiable convex functions, a special attention will be devoted to the tilted version of the popular Huber loss function,

$$\rho_\tau(u) = 2\begin{cases} (c|u| - c^2/2)|\tau - I(u \le 0)| & |u| > c \\ u^2/2|\tau - I(u \le 0)| & |u| \le c. \end{cases} \tag{3}$$

We note that if we set $\tau = 0.5$, a well-defined distribution, the so-called Least Informa-

tive (LI) distribution, is associated to this function (Huber, 1981, Section 4.5) in the same way as the normal distribution is associated with quadratic loss function. We consider the parametric distribution associated to a general loss $\rho_\tau(u)$, that we will call Generalized Asymmetric Least Informative (GALI) distribution. This distribution plays a role similar to that of the Asymmetric Laplace (AL) distribution in quantile regression (Yu and Moyeed, 2001). We use this parametric representation for proposing an estimator of the scale parameter σ_τ. With reference to the special case (3), we use the distribution associated to this loss function to propose an estimator for the tuning constant c, using a method than can be generalized to other loss functions involving tuning constants.

We further propose a test for the presence of clustering in the data. Clustering is measured by cluster-specific M-quantile coefficients (Chambers and Tzavidis, 2006). The proposed test offers an alternative approach to more conventional hypothesis tests for the significance of the between cluster variance component or for the intra-cluster correlation coefficient. The toolkit we propose in the paper can be applied in small area estimation framework to validate the use of M-quantile models for prediction; in fact a common criticism of M-quantile regression in this field is the relative lack of diagnostics and model testing.

The paper is organized as follows. In Section 2 we review M-quantile regression and introduce a new estimator for the scale parameter based on the GALI distribution. In Section 3 we introduce the pseudo-R^2 goodness-of-fit measure and likelihood ratio and Wald type tests for linear hypotheses on the M-quantile regression parameters. Section 4 reviews the use of M-quantile regression for measuring cluster heterogeneity, its application in small area estimation and presents the test for the presence of clustering. In Section 5 we present simulation studies aimed at assessing the finite sample properties of the proposed methods and in Section 6 we present the application of the methods to real data. Finally,

Section 7 concludes the paper with some final remarks.

2 M-quantile regression

Let y be a random variable and $\mathbf{x}$ a p-dimensional random vector with first component $x_1 = 1$. The observed data $\{(\mathbf{x}_i, y_i), i = 1, \ldots, n\}$ is assumed to be a random sample of size n drawn from the population; thus $(\mathbf{x}_i, y_i)$ are independent and identically distributed random variables. Assuming a linear model, for any $\tau \in (0, 1)$, the M-quantile (hereafter, MQ) of order τ of y_i given $\mathbf{x}_i$ is defined by

$$MQ_\tau(y_i|\mathbf{x}_i) = \mathbf{x}_i^T \boldsymbol{\beta}_\tau, \tag{4}$$

where $\boldsymbol{\beta}_\tau \in \Theta \subset \mathcal{R}^p$ is the solution to

$$\min_{\boldsymbol{\beta} \in \Theta} E\left[\rho_\tau\left(\frac{y_i - \mathbf{x}_i^T \boldsymbol{\beta}}{\sigma_\tau}\right)\right], \tag{5}$$

and σ_τ is a scale parameter that characterizes the distribution of $\varepsilon_{\tau i} = y_i - \mathbf{x}_i^T \boldsymbol{\beta}_\tau$. The linear specification in (4) can be alternatively written as

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta}_\tau + \varepsilon_{\tau i},$$

where $\{\varepsilon_{\tau i}\}$ is a sequence of independent and identically distributed errors with unknown distribution function F_τ satisfying, by definition, $MQ_\tau(\varepsilon_{\tau i}|\mathbf{x}_i) = 0$. The estimator of the MQ regression coefficients (Breckling and Chambers, 1988) is defined as

$$\hat{\boldsymbol{\beta}}_\tau = \text{argmin} \sum_{i=1}^{n} \rho_\tau\left(\frac{y_i - \mathbf{x}_i^T \boldsymbol{\beta}}{\hat{\sigma}_\tau}\right), \tag{6}$$

where $\hat{\sigma}_\tau$ is a consistent estimator of σ_τ. Since ρ is (a.e.) continuously differentiable and convex, the vector $\hat{\boldsymbol{\beta}}_\tau$ can equivalently be obtained as the solution of the following system

of equations

$$\sum_{i=1}^{n} \psi_\tau \left(\frac{y_i - \mathbf{x}_i^T \boldsymbol{\beta}}{\hat{\sigma}_\tau} \right) \mathbf{x}_i = \mathbf{0}, \tag{7}$$

where $\psi_\tau(u) = d\rho_\tau(u)/du = |\tau - I(u < 0)|\psi(u)$, with $\psi(u) = d\rho(u)/du$. An iterative method is needed here to obtain a solution, like an iteratively re-weighted least squares algorithm or the Newton-Raphson algorithm.

Regarding the scale parameter σ_τ, it may generally be defined by an implicit relation of the form

$$E\Big[\chi\left(\frac{\varepsilon_{\tau i}}{\sigma_\tau}\right)\Big] = 0, \tag{8}$$

where the expectation is taken with respect to the distribution of $\varepsilon_{\tau i}$. In MQ regression, a typical choice for χ is $\chi(u) = sgn(|u - Med(u)| - 1)$, which leads to the scaled population median absolute deviation $\sigma_\tau = \frac{Med\{|\boldsymbol{\varepsilon}_\tau - \xi_{1/2,\tau}|\}}{q}$, $\xi_{1/2,\tau} = Med(F_\tau(\boldsymbol{\varepsilon}_\tau))$, $q = \Phi^{-1}(3/4) = 0.6745$, with Φ denoting the distribution function of the standard normal distribution. The corresponding estimator is the scaled sample median absolute deviation (MAD)

$$\hat{\sigma}_\tau = \frac{Med\{|\hat{\boldsymbol{\varepsilon}}_\tau - Med(\hat{\boldsymbol{\varepsilon}}_\tau)|\}}{q}, \tag{9}$$

where $\hat{\boldsymbol{\varepsilon}}_\tau = (\hat{\varepsilon}_{\tau 1}, \ldots, \hat{\varepsilon}_{\tau n})$, $\hat{\varepsilon}_{\tau i} = y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_\tau$.

The asymptotic theory for MQ regression with i.i.d. errors and fixed regressors can be derived from the results in Huber (1973), as pointed out in Breckling and Chambers (1988). Bianchi and Salvati (2015) show the consistency of the estimator of $\boldsymbol{\beta}_\tau$ and its asymptotic variance,

$$\widehat{Var}(\hat{\boldsymbol{\beta}}_\tau) = (n-p)^{-1} n \hat{\mathbf{W}}_\tau^{-1} \hat{\mathbf{G}}_\tau \hat{\mathbf{W}}_\tau^{-1} \tag{10}$$

where

$$\hat{\mathbf{W}}_\tau = (n\hat{\sigma}_\tau)^{-1} \sum_{i=1}^{n} \hat{\psi}'_{\tau i} \mathbf{x}_i \mathbf{x}_i^T,$$

$$\hat{\mathbf{G}}_\tau = n^{-1} \sum_{i=1}^{n} \hat{\psi}^2_{\tau i} \mathbf{x}_i \mathbf{x}_i^T,$$

with $\hat{\psi}'_{\tau i} := \psi'_\tau(\hat{\varepsilon}_{i\tau}/\hat{\sigma}_\tau)$, $\hat{\psi}_{\tau i} = \psi_\tau(\hat{\varepsilon}_{i\tau}/\hat{\sigma}_\tau)$ in case of stochastic regressors and in the presence of heteroskedasticity.

2.1 A likelihood perspective for M-quantiles: the Generalized Asymmetric Least Informative distribution

Yu and Moyeed (2001) show the relationship between the loss function for quantile regression and the maximization of a likelihood function formed by combining independently distributed Asymmetric Laplace densities. In this Section we show a similar relationship for MQ regression models.

Given a loss function ρ_τ, we can define the GALI random variable with density function

$$f_\tau(y;\mu_\tau,\sigma_\tau) = \frac{1}{\sigma_\tau B_\tau} exp\left\{-\rho_\tau\left(\frac{y-\mu_\tau}{\sigma_\tau}\right)\right\}, \quad -\infty < y < +\infty. \tag{11}$$

where $B_\tau = \int_{-\infty}^{+\infty} \frac{1}{\sigma_\tau} exp\left\{-\rho_\tau\left(\frac{y-\mu_\tau}{\sigma_\tau}\right)\right\} dy < +\infty$ and μ_τ and σ_τ are location and scale parameters. We note that μ_τ coincides with the τ^{th} MQ of the distribution; in fact μ_τ can be obtained as the solution of

$$\int_{-\infty}^{+\infty} \psi_\tau\Big(\frac{y-\mu_\tau}{\sigma_\tau}\Big) f_\tau(y;\mu_\tau,\sigma_\tau)dy = 0,$$

that defines the MQ of the distribution.

For linear MQ regression, that is when $\mu_\tau = \mu_{\tau i} = \mathbf{x}_i^T\boldsymbol{\beta}_\tau$, the estimators of the unknown regression parameters $\boldsymbol{\beta}_\tau$ and the scale σ_τ may be obtained by maximizing the log-likelihood function:

$$l_\tau(y) = -n\log\sigma_\tau - n\log B_\tau - \sum_{i=1}^{n} \rho_\tau\Big(\frac{y_i - \mathbf{x}_i^T\boldsymbol{\beta}_\tau}{\sigma_\tau}\Big). \tag{12}$$

The estimating equations for the regression coefficients $\boldsymbol{\beta}_\tau$ are the same as those of equation (7). The estimating equation for σ_τ is

$$-\frac{n}{\sigma_\tau} + \frac{1}{\sigma_\tau^2}\sum_{i=1}^{n} \psi_\tau\Big(\frac{y_i - \mathbf{x}_i^T\boldsymbol{\beta}_\tau}{\sigma_\tau}\Big)(y_i - \mathbf{x}_i^T\boldsymbol{\beta}_\tau) = 0, \tag{13}$$

and its solution defines a new estimator for σ_τ alternative to (9). With respect to (8) in this case $\chi(u) = -u\psi_\tau(u) - 1$ and the parameter is defined as the solution of

$$E\left[-\varepsilon_{\tau i}\psi_\tau\left(\frac{\varepsilon_{\tau i}}{\sigma_\tau}\right)\right] = \sigma_\tau.$$

This choice is in line with what Koenker and Machado (1999) and Yu and Zhang (2005) propose for quantile regression, considering the maximum likelihood estimator under the asymmetric Laplace distribution.

Solving equations (7) and (13) requires an iterative algorithm. The steps of this algorithm are as follows:

1. For specified τ define initial estimates $\hat{\boldsymbol{\beta}}_\tau^{(0)}$ and $\hat{\sigma}_\tau^{(0)}$.
2. At each iteration t calculate $w_{\tau i}^{(t-1)} = \psi_\tau(u_{\tau i}^{(t-1)})/u_{\tau i}^{(t-1)}$ with $u_{\tau i}^{(t-1)} = (y_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}}_\tau^{(t-1)})/\hat{\sigma}_\tau^{(t-1)}$.
3. Compute the new weighted least squares estimates from

$$\hat{\boldsymbol{\beta}}_\tau^{(t)} = \Big\{\sum_{i=1}^{n}(w_{\tau i}^{(t-1)}\mathbf{x}_i\mathbf{x}_i^T)\Big\}^{-1}\Big\{\sum_{i=1}^{n}(y_i w_{\tau i}^{(t-1)}\mathbf{x}_i)\Big\}. \tag{14}$$

4. Compute the new estimate of $\hat{\sigma}_\tau$ by

$$\hat{\sigma}_\tau^{(t)} = \Big\{n^{-1}\sum_{i=1}^{n} w_{\tau i}^{(t-1)}(y_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}}_\tau^{(t-1)})^2\Big\}^{1/2}. \tag{15}$$

5. Repeat steps 2-4 until convergence. Convergence is achieved when the difference between the estimated model parameters obtained from two successive iterations is less than a small pre-specified value.

The consistency of the scale estimators (MAD and MLE) can be proved by standard theory of M-estimators (Wooldridge, 2010), assuming that (8) has a unique solution.

If $\rho_\tau(\cdot)$ is the Huber loss function defined in (3) the normalizing constant is given by

$$B_\tau = \sqrt{\frac{\pi}{\tau}}\Big[\Phi(c\sqrt{2\tau}) - 1/2\Big] + \sqrt{\frac{\pi}{1-\tau}}\Big[\Phi(c\sqrt{2(1-\tau)}) - 1/2\Big]$$
$$+\frac{1}{2c\tau}exp\{-c^2\tau\} + \frac{1}{2c(1-\tau)}exp\{-c^2(1-\tau)\}, \tag{16}$$

where $\tau \in (0,1)$ and Φ is the cumulative distribution function of the standard Normal distribution. In this case we call (11) the Asymmetric Least Informative (ALI) distribution. This distribution is essentially a modified standard normal distribution with heavier tails (when $y > c$). For $\tau = 0.5$, this distribution was derived by Huber (1981, Section 4.5) as the one minimizing the Fisher information in the ε-contaminated neighborhood of the normal distribution. Formulae for the cumulative distribution function and moments of the ALI distribution ($\tau \in (0,1)$) are in the Appendix.

The ALI distribution depends on the tuning constant c. In M-regression, the tuning constant is defined by the data analyst such that the M-estimate has a specified asymptotic efficiency (generally 95%) under normality (Huber, 1981). Alternatively, Wang et al. (2007) propose a data-driven method, based on efficiency arguments.

In this paper, we propose to interpret c as a parameter of the density f_τ and estimate $\boldsymbol{\beta}_\tau$, σ_τ and c by maximizing the log-likelihood function (12). For estimating the tuning constant there is no closed form. In this case the compass search algorithm or the Nelder-Mead (Griva et al., 2008) can be used. The final estimating procedure works by adding to the proposed iterative algorithm the new step 4′ below:

4′ Given $\hat{\boldsymbol{\beta}}_\tau^{(t)}$ and $\hat{\sigma}_\tau^{(t)}$ maximize the log-likelihood function (12) with respect to c using the compass search algorithm (Bottai et al., 2015) or the Nelder-Mead algorithm.

An `R` function that implements an iterative algorithm for estimating the parameters is available from the authors.

The idea of estimating the tuning constant using likelihood equations can be applied to other loss functions as well whenever they include an additional parameter or tuning constant.

3 Goodness-of-fit and likelihood ratio type tests in M-quantile regression

In this section we present a pseudo-R^2 goodness-of-fit statistic for MQ regression and likelihood ratio and Wald type tests for linear hypotheses on the regression parameters. For a given quantile, the introduction of the pseudo-R^2 is motivated by the need for a measure analogous to the ordinary R^2 used in least squares regression. Since this goodness-of-fit statistic will be quantile-dependent, it is also useful to study its variation across quantiles.

3.1 A goodness-of-fit measure

We start by partitioning MQ regression as follows,

$$MQ_\tau(y_i|\mathbf{x}_i) = \mathbf{x}_{i1}^T\boldsymbol{\beta}_{1\tau} + \mathbf{x}_{i2}^T\boldsymbol{\beta}_{2\tau}, \tag{17}$$

where $\boldsymbol{\beta}_\tau = (\boldsymbol{\beta}_{1\tau}^T, \boldsymbol{\beta}_{2\tau}^T)^T$, $\boldsymbol{\beta}_{1\tau}$ is a $(p-k)\times 1$ vector and $\boldsymbol{\beta}_{2\tau}$ is a $k\times 1$ $(0<k<p)$ vector. We are interested in testing the null hypothesis:

$$H_0 : \boldsymbol{\beta}_{2\tau} = \mathbf{0}. \tag{18}$$

Let $\hat{\boldsymbol{\beta}}_\tau$ denote the MQ estimator of the full model and let $\tilde{\boldsymbol{\beta}}_\tau = (\tilde{\boldsymbol{\beta}}_{1\tau}^T, \mathbf{0}^T)^T$ denote the MQ estimator under the null hypothesis specified in (18).

A relative goodness-of-fit measure comparing the full to the reduced MQ regression model is defined as

$$R^2_\rho(\tau) = 1 - \frac{\sum_{i=1}^n \rho_\tau\left(\frac{y_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}}_\tau}{\hat{\sigma}_\tau}\right)}{\sum_{i=1}^n \rho_\tau\left(\frac{y_i - \mathbf{x}_i^T\tilde{\boldsymbol{\beta}}_\tau}{\hat{\sigma}_\tau}\right)}. \tag{19}$$

When the reduced model includes only the intercept, this measure is the natural analog of the usual R^2 goodness-of-fit measure used in mean regression. It varies between 0 and 1 and it represents a measure of goodness-of-fit for a specified τ.

3.2 Hypothesis testing

For testing the null hypothesis (18), the following theorem presents the distribution of the likelihood ratio statistic when the residuals follow a general distribution. This leads to a likelihood ratio type test. The theorem is proved by using the following regularity conditions (Bianchi and Salvati, 2015):

(C1) ρ is convex continuously differentiable a.e. and ψ is bounded with one bounded derivative a.e., not identically zero;

(C2) $E|\mathbf{x}_i|^4 < +\infty$, $E|y_i|^4 < +\infty$;

(C3) $E\left[\mathbf{x}_i\mathbf{x}_i^T\right]$ is nonsingular;

(C4) the errors $\varepsilon_{\tau i}$ are independent of $\mathbf{x}_i$.

Further, let

$$\hat{V}(\tau) = \sum_{i=1}^n \rho_\tau\left(\frac{y_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}}_\tau}{\sigma_\tau}\right), \qquad \tilde{V}(\tau) = \sum_{i=1}^n \rho_\tau\left(\frac{y_i - \mathbf{x}_i^T\tilde{\boldsymbol{\beta}}_\tau}{\sigma_\tau}\right).$$

Theorem 1. *Provided conditions (C1)-(C4) are satisfied under the null hypothesis* H_0

$$-2\frac{E\psi'_{\tau i}}{E\psi^2_{\tau i}}(\hat{V}(\tau)-\tilde{V}(\tau)) \xrightarrow{d} \chi^2_k, \tag{20}$$

where $\psi'_{\tau i}=\psi'_\tau(\varepsilon_{\tau i}/\sigma_\tau)$, $\psi_{\tau i}=\psi_\tau(\varepsilon_{\tau i}/\sigma_\tau)$.

Proof. Using a second order Taylor expansion

$$2[\tilde{V}(\tau)-\hat{V}(\tau)]=\sqrt{n}(\tilde{\boldsymbol{\beta}}_\tau-\hat{\boldsymbol{\beta}}_\tau)^T(\boldsymbol{\Psi}_\tau/\sigma_\tau)\sqrt{n}(\tilde{\boldsymbol{\beta}}_\tau-\hat{\boldsymbol{\beta}}_\tau)+o_p(1), \tag{21}$$

where, by using (C4), $\boldsymbol{\Psi}_\tau=\sigma_\tau^{-1}E(\psi'_{\tau i})E(\mathbf{x}_i\mathbf{x}_i^T)$. Theorem 1 in Bianchi and Salvati (2015) ensures that

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_\tau-\boldsymbol{\beta}_\tau)=\boldsymbol{\Psi}_\tau^{-1}n^{-1/2}\sum_{i=1}^n\psi_{\tau i}\mathbf{x}_i+o_p(1). \tag{22}$$

Similarly, a standard mean value expansion (under H_0) gives

$$n^{-1/2}\sum_{i=1}^n\tilde{\psi}_{\tau i}\mathbf{x}_i=n^{-1/2}\sum_{i=1}^n\psi_{\tau i}\mathbf{x}_i-\boldsymbol{\Psi}_\tau\sqrt{n}(\tilde{\boldsymbol{\beta}}_\tau-\boldsymbol{\beta}_\tau)+o_p(1),$$

where $\tilde{\psi}_{\tau i}=\psi_\tau(\tilde{\varepsilon}_{\tau i}/\sigma_\tau)$, $\tilde{\varepsilon}_{\tau i}=y_i-\mathbf{x}_i^T\tilde{\boldsymbol{\beta}}_\tau$. Hence,

$$\sqrt{n}(\tilde{\boldsymbol{\beta}}_\tau-\boldsymbol{\beta}_\tau)=\boldsymbol{\Psi}_\tau^{-1}n^{-1/2}\left[-\sum_{i=1}^n\tilde{\psi}_{\tau i}\mathbf{x}_i+\sum_{i=1}^n\psi_{\tau i}\mathbf{x}_i\right]+o_p(1). \tag{23}$$

Substituting (22) and (23) into (21), we obtain

$$2[\tilde{V}(\tau)-\hat{V}(\tau)]=\left(n^{-1/2}\sum_{i=1}^n\tilde{\psi}_{\tau i}\mathbf{x}_i\right)^T\left(E(\psi'_{\tau i})E(\mathbf{x}_i\mathbf{x}_i^T)\right)^{-1}\left(n^{-1/2}\sum_{i=1}^n\tilde{\psi}_{\tau i}\mathbf{x}_i\right)+o_p(1).$$

Following Wooldridge (2010), we introduce the $k\times p$ full rank matrix $\mathbf{R}=[\mathbf{0}:\mathbf{I}_k]$ and write H_0 as $\mathbf{R}\boldsymbol{\beta}_\tau=\mathbf{0}$. Since $\mathbf{R}\sqrt{n}(\tilde{\boldsymbol{\beta}}_\tau-\boldsymbol{\beta}_\tau)=\mathbf{0}$, it can be proved (multiplying equation (23) by $\mathbf{R}\boldsymbol{\Psi}_\tau^{-1}$) that

$$\mathbf{R}\boldsymbol{\Psi}_\tau^{-1}n^{-1/2}\sum_{i=1}^n\tilde{\psi}_{\tau i}\mathbf{x}_i\xrightarrow{d}N(\mathbf{0},\mathbf{R}\boldsymbol{\Sigma}_\tau\mathbf{R}^T),$$

where

$$\boldsymbol{\Sigma}_\tau = \sigma_\tau^2 \frac{E\psi_{\tau i}^2}{E\psi'_{\tau i}} E\left[\mathbf{x}_i \mathbf{x}_i^T\right]^{-1}, \tag{24}$$

so that

$$\left(n^{-1/2}\sum_{i=1}^n \tilde{\psi}_{\tau i}\mathbf{x}_i\right)^T \boldsymbol{\Psi}_\tau^{-1}\mathbf{R}^T\left(\mathbf{R}\boldsymbol{\Sigma}_\tau\mathbf{R}^T\right)^{-1}\mathbf{R}\boldsymbol{\Psi}_\tau^{-1}\left(n^{-1/2}\sum_{i=1}^n \tilde{\psi}_{\tau i}\mathbf{x}_i\right) \xrightarrow{d} \chi_k^2.$$

The previous expression can be simplified to

$$\left(n^{-1/2}\sum_{i=1}^n \tilde{\psi}_{\tau i}\mathbf{x}_i\right)^T \left(E(\psi_{\tau i}^2)E(\mathbf{x}_i\mathbf{x}_i^T)\right)^{-1}\left(n^{-1/2}\sum_{i=1}^n \tilde{\psi}_{\tau i}\mathbf{x}_i\right) \xrightarrow{d} \chi_k^2$$

and therefore we have that

$$\begin{aligned} &2\frac{E\psi'_{\tau i}}{E\psi_{\tau i}^2}[\tilde{V}(\tau) - \hat{V}(\tau)] \\ &\quad = \left(n^{-1/2}\sum_{i=1}^n \tilde{\psi}_{\tau i}\mathbf{x}_i\right)^T \left(E(\psi_{\tau i}^2)E(\mathbf{x}_i\mathbf{x}_i^T)\right)^{-1}\left(n^{-1/2}\sum_{i=1}^n \tilde{\psi}_{\tau i}\mathbf{x}_i\right) + o_p(1) \xrightarrow{d} \chi_k^2. \end{aligned} \tag{25}$$

□

A hypothesis test for H_0 is obtained by substituting the unknown quantities in (20) with consistent estimators leading to,

$$-2\frac{(n-p)^{-1}\sum_{i=1}^n \hat{\psi}'_{\tau i}}{n^{-1}\sum_{i=1}^n \hat{\psi}_{\tau i}^2}\left[\sum_{i=1}^n \rho_\tau\left(\frac{y_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}}_\tau}{\hat{\sigma}_\tau}\right) - \sum_{i=1}^n \rho_\tau\left(\frac{y_i - \mathbf{x}_i^T\tilde{\boldsymbol{\beta}}_\tau}{\hat{\sigma}_\tau}\right)\right],$$

where $\hat{\psi}'_{\tau i}$ and $\hat{\psi}_{\tau i}$ have been previously defined and the nuisance parameter σ_τ is estimated under the full model. This is to ensure that the test statistic is nonnegative. This test is more commonly known as likelihood ratio (LR) type test since the density of the $\varepsilon_{\tau i}$ does not have to correspond to the loss function. Notice also that the proposed test can be easily extended to test more general linear hypotheses for example, $H_0 : \mathbf{R}\boldsymbol{\beta}_\tau = \mathbf{r}$, where $\mathbf{R}$ is a $k \times p$ full rank matrix and $\mathbf{r}$ is a $k \times 1$ vector. Similar results for M-regression estimators

are provided by Schrader and Hettmansperger (1980) in the case of fixed regressors, and for quantile regression with fixed regressors by Koenker and Machado (1999).

An alternative to the LR-type test is to use a Wald type test. The test statistic is derived by using Theorem 1 in Bianchi and Salvati (2015). Let $\mathbf{R} = [\mathbf{0} : \mathbf{I}_k]$. It follows that under H_0

$$n(\mathbf{R}\hat{\boldsymbol{\beta}}_\tau)^T[\mathbf{R}\boldsymbol{\Sigma}_\tau\mathbf{R}]^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}}_\tau) \xrightarrow{d} \chi^2_k,$$

where $\boldsymbol{\Sigma}_\tau$ is defined in (24). Replacing $\boldsymbol{\Sigma}_\tau$ with a consistent estimator

$$\hat{\boldsymbol{\Sigma}}_\tau = \hat{\sigma}^2_\tau \frac{(n-p)^{-1}\sum_{i=1}^n \hat{\psi}^2_{\tau i}}{n^{-1}\sum_{i=1}^n \hat{\psi}'_{\tau i}} \left[\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^T\right]^{-1},$$

the statistic

$$W \equiv n(\mathbf{R}\hat{\boldsymbol{\beta}}_\tau)^T[\mathbf{R}\hat{\boldsymbol{\Sigma}}_\tau\mathbf{R}]^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}}_\tau)$$

follows asymptotically a χ^2 distribution. A major difference between the LR-type test and the Wald type test is that the latter can be made robust to the presence of heteroskedasticity by using a robust estimator of the covariance matrix instead of $\hat{\boldsymbol{\Sigma}}_\tau$.

4 A Likelihood Ratio-type test for the presence of clustering

In this section we present a LR-type test for the of presence unobserved heterogeneity (clustering). The proposed test has a similar aim to that of a hypothesis test for the strict positiveness of variance components in the case of a linear mixed (random) effects model. Testing for the presence of significant clustering is a well known problem in literature (Greven et al., 2008; Crainiceanu and Ruppert, 2004; Datta et al., 2011). Clustering can exist either because of the design used to collect the data (i.e. use of a multi-stage cluster

design) or because of natural structures that exist in the population (i.e. pupils nested within schools or individuals nested within households). The discussion in this section will pay special attention to the existence of area-effects in small area estimation.

Let us start with the simplest, two-level, hierarchical structure. We introduce a second subscript in our notation for indicating the hierarchical nature of the data, $\{(\mathbf{x}_{ij}, y_{ij}), i = 1, \ldots, n_j; j = 1, \ldots, d\}$, where d is the number of the primary units (e.g. small areas) drawn from the population of size N. Let us suppose that a population is divided into d non-overlapping primary units of size N_j, $j = 1, ..., d$ and $n_j > 0$ is the number of secondary units drawn from each primary unit.

The papers by Chambers and Tzavidis (2006) and Aragon et al. (2005) were among the first to introduce the idea of measuring heterogeneity in the data via M-quantiles. In particular, Chambers and Tzavidis (2006) characterize the variability across the population of interest by introducing the concept of MQ-coefficients. At the population level the MQ-coefficient for a secondary-level unit within a primary-level unit is defined as the value τ_{ij} such that $MQ_{\tau_{ij}}(y_{ij}|\mathbf{x}_{ij}) = y_{ij}$. If a hierarchical structure does explain part of the variability, after accounting for the effect of covariates, units within primary units (in short, groups) are expected to have similar MQ-coefficients. Chambers and Tzavidis (2006) propose to characterize each group j by the average of the MQ-coefficients of the units that belong to that group. The group-specific MQ-coefficient, denoted by τ_j, identifies the most characteristic MQ regression line for that group. We can think of this in the context of linear mixed models as the group-specific regression line that is distinguished from population-average line by the random effect.

In small area estimation we assume that a sample s is drawn from the population and that area-specific samples s_j of size $n_j \geq 0$ are available for each area. Note that non-sample areas have $n_j = 0$, in which case s_j is the empty set. The set r_j contains the $N_j - n_j$ indices

of the non-sampled units in small area j. The aim is to use this data to predict various area specific quantities, including (but not only) the area j mean m_j of y. When (4) holds, and $\boldsymbol{\beta}_\tau$ is a sufficiently smooth function of τ, Chambers and Tzavidis (2006) suggest a predictor of m_j of the form:

$$\hat{m}_j^{MQ} = N_j^{-1}\Big\{\sum_{i\epsilon s_j} y_{ij} + \sum_{i\epsilon r_j} \mathbf{x}_{ij}^T \hat{\boldsymbol{\beta}}_{\hat{\tau}_j}\Big\}, \tag{26}$$

where $\hat{\tau}_j$ is an estimate of the average value of the MQ-coefficients of the units in area j. When there is no sample in area or there is no area effect, a synthetic M-quantile predictor can be formed by setting $\hat{\tau}_j = 0.5$ ($\hat{m}_j^{MQ/SYN}$).

Our aim is to test for the presence of significant area/cluster effects by proposing a testing procedure for the cluster-specific M-quantile coefficients τ_j. This procedure is presented in the next section.

4.1 Testing procedure

Differently from Chambers and Tzavidis (2006), in the present work we define the MQ-coefficients $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_d)^T$ by adopting an approach that is explicitly based on the loss function. Within group j, τ_j is defined to be the one that uniquely solves

$$\min_{\tau} E\left[\rho\left(\frac{y_{ij} - \mathbf{x}_{ij}^T \boldsymbol{\beta}_\tau}{\sigma}\right)|j\right].$$

Intuitively, τ_j is defined as the MQ for which the regression plane identified by $\boldsymbol{\beta}_{\tau_j}$ is closer to observations from group j, according to the metrics of $\rho(\cdot)$. Note that $\rho(\cdot)$ is the untilted loss function, i.e. $\rho_{0.5}(\cdot)$, so the scale σ coincides with $\sigma_{0.5}$. The use of the untilted loss function is motivated by the search of the regression plane that best fits the units in a specific sub-group of the population. Testing for the presence of clustering is equivalent to

testing whether the group-specific MQ-coefficients are all equal, that is,

$$H_0 : \tau_j = 0.5 \ \forall j = 1, \ldots, d$$

$$H_A : \tau_j \neq 0.5 \text{ for at least one } j.$$

By means of a Taylor expansion centered at $\tau = 0.5$ (which is the global minimizer), we see that the τ_j's satisfy a linear constraint as their mean is 0.5. Hence the null hypothesis H_0 corresponds to $d-1$ restrictions or, equivalently, the hypothesis may be expressed in the form $\mathbf{R}\boldsymbol{\tau} = \mathbf{r}$, where $\mathbf{R}$ is a matrix of rank $d-1$.

A natural estimator $\hat{\tau}_j$ for τ_j is obtained by solving

$$\min_{\tau} \sum_{i=1}^{n_j} \rho\left(\frac{y_{ij} - \mathbf{x}_{ij}^T \hat{\boldsymbol{\beta}}_\tau}{\hat{\sigma}}\right),$$

where $\hat{\sigma}$ is an estimator of σ such as the one obtained solving (13) for $\tau = 0.5$. Since ρ is a positive function, the problem may be rewritten as follows. The vector of estimated MQ-coefficients $\hat{\boldsymbol{\tau}} = (\hat{\tau}_1, \ldots, \hat{\tau}_d)^T$ is obtained by the solution of

$$\min_{(\tau_1,\ldots,\tau_d)} \sum_{j=1}^{d} \sum_{i=1}^{n_j} \rho\left(\frac{y_{ij} - \mathbf{x}_{ij}^T \hat{\boldsymbol{\beta}}_{\tau_j}}{\hat{\sigma}}\right). \tag{27}$$

The test statistic to be used for testing H_0 is provided by the following theorem.

Theorem 2. *Assuming conditions (C1)-(C4) are satisfied and that $\boldsymbol{\beta}_\tau$ is differentiable in τ with $\partial^2 \boldsymbol{\beta}_\tau / \partial \tau^2 = 0$ (i.e. $\boldsymbol{\beta}_\tau$ linear in τ), under H_0*

$$-2\frac{E\psi'_{ij}}{E\psi_{ij}^2}\left[\sum_{j=1}^{d}\sum_{i=1}^{n_j} \rho\left(\frac{y_{ij} - \mathbf{x}_{ij}^T \boldsymbol{\beta}_{\hat{\tau}_j}}{\sigma}\right) - \sum_{j=1}^{d}\sum_{i=1}^{n_j} \rho\left(\frac{y_{ij} - \mathbf{x}_{ij}^T \boldsymbol{\beta}_{0.5}}{\sigma}\right)\right] \xrightarrow{d} \chi^2_{d-1}$$

where $\psi'_{ij} = \psi'(\varepsilon_{0.5ij}/\sigma)$, $\psi_{ij} = \psi(\varepsilon_{0.5ij}/\sigma)$ and $\varepsilon_{0.5ij} = (y_{ij} - \mathbf{x}_{ij}^T \boldsymbol{\beta}_{0.5})$.

Proof. Under the assumptions of the theorem, convergence of $\hat{\tau}_j$ to τ_j is verified by using standard Taylor linearization techniques. For the asymptotic distribution of the test statistic, let $Q(\boldsymbol{\tau}) = \sum_{j=1}^{d}\sum_{i=1}^{n_j}\rho\left(\frac{y_{ij}-\mathbf{x}_{ij}^T\boldsymbol{\beta}_{\tau_j}}{\sigma}\right)$, $\mathbf{s}(\boldsymbol{\tau}) = \left\{\frac{1}{\sqrt{n_j}}\sum_{i=1}^{n_j}\frac{\partial\rho}{\partial\tau_j}\left(\frac{y_{ij}-\mathbf{x}_{ij}^T\boldsymbol{\beta}_{\tau_j}}{\sigma}\right)\right\}_{j=1}^{d}$, $\mathbf{H}(\boldsymbol{\tau}) = diag\left\{\frac{1}{n_j}\sum_{i=1}^{n_j}\frac{\partial^2\rho}{\partial\tau_j^2}\left(\frac{y_{ij}-\mathbf{x}_{ij}^T\boldsymbol{\beta}_{\tau_j}}{\sigma}\right)\right\}$, and $\mathbf{n} = (n_1,\ldots,n_d)^T$. Let $\mathbf{A}_0 = diag\{a_j\}$ and $\mathbf{B}_0 = diag\{b_j\}$ with

$$\begin{aligned} a_j &= E\left[\frac{\partial^2\rho}{\partial\tau_j^2}\left(\frac{y_{ij}-\mathbf{x}_{ij}^T\boldsymbol{\beta}_{\tau_j}}{\sigma}\right)\Big|_{0.5}\right] = \sigma^{-2}E\psi'_{ij}E\left(\mathbf{x}_{ij}^T\frac{\partial\boldsymbol{\beta}_{\tau_j}}{\partial\tau_j}\Big|_{0.5}\right)^2 \\ b_j &= E\left[\frac{\partial\rho}{\partial\tau_j}\left(\frac{y_{ij}-\mathbf{x}_{ij}^T\boldsymbol{\beta}_{\tau_j}}{\sigma}\right)\Big|_{0.5}\right]^2 = \sigma^{-2}E\psi_{ij}^2E\left(\mathbf{x}_{ij}^T\frac{\partial\boldsymbol{\beta}_{\tau_j}}{\partial\tau_j}\Big|_{0.5}\right)^2. \end{aligned}$$

Under H_0, a mean value expansion yields

$$\mathbf{0} = \mathbf{s}(\hat{\boldsymbol{\tau}}) = \mathbf{s}(\mathbf{0.5}) + \sqrt{\mathbf{n}}\cdot(\hat{\boldsymbol{\tau}}-\mathbf{0.5}) + o_p(1),$$

implying $\sqrt{\mathbf{n}}\cdot(\hat{\boldsymbol{\tau}}-\mathbf{0.5}) \xrightarrow{d} N(\mathbf{0}, \mathbf{A}_0^{-1}\mathbf{B}_0\mathbf{A}_0^{-1})$, as $n_j \to +\infty$, $j = 1,\ldots,d$, where $\cdot$ denotes the Hadamard product.

Then

$$\begin{aligned} Q(\mathbf{0.5}) - Q(\hat{\boldsymbol{\tau}}) &= \frac{1}{2}(\hat{\boldsymbol{\tau}}-\mathbf{0.5})^T H(\dot{\boldsymbol{\tau}})(\hat{\boldsymbol{\tau}}-\mathbf{0.5}) \\ &= \frac{1}{2}[\sqrt{\mathbf{n}}\cdot(\hat{\boldsymbol{\tau}}-\mathbf{0.5})]^T\mathbf{A}_0[\sqrt{\mathbf{n}}\cdot(\hat{\boldsymbol{\tau}}-\mathbf{0.5})] + o_p(1), \end{aligned}$$

where $\dot{\boldsymbol{\tau}}$ is a value between $\hat{\boldsymbol{\tau}}$ and $\mathbf{0.5}$. Hence

$$2[Q(\mathbf{0.5}) - Q(\hat{\boldsymbol{\tau}})]\frac{E\psi'_{ij}}{E\psi_{ij}^2} = [\sqrt{\mathbf{n}}\cdot(\hat{\boldsymbol{\tau}}-\mathbf{0.5})]^T[\mathbf{A}_0^{-1}\mathbf{B}_0\mathbf{A}_0^{-1}]^{-1}[\sqrt{\mathbf{n}}\cdot(\hat{\boldsymbol{\tau}}-\mathbf{0.5})] + o_p(1).$$

Due to the linear relationship existing among parameters and hence estimators $\hat{\tau}_j$'s, the previous expression may be reparametrized leading to a χ^2_{d-1} asymptotic distribution. □

A test may hence be based on

$$-2\frac{(n-p)^{-1}\sum_{ij}\hat{\psi}'_{ij}}{n^{-1}\sum_{ij}\hat{\psi}_{ij}^2}\left[\sum_{j=1}^{d}\sum_{i=1}^{n_j}\rho\left(\frac{y_{ij}-\mathbf{x}_{ij}^T\hat{\boldsymbol{\beta}}_{\hat{\tau}_j}}{\hat{\sigma}}\right) - \sum_{j=1}^{d}\sum_{i=1}^{n_j}\rho\left(\frac{y_{ij}-\mathbf{x}_{ij}^T\hat{\boldsymbol{\beta}}_{0.5}}{\hat{\sigma}}\right)\right],$$

where $\hat{\psi}'_{ij} = \psi'(\hat{\varepsilon}_{0.5ij}/\hat{\sigma})$, $\hat{\psi}_{ij} = \psi(\hat{\varepsilon}_{0.5ij}/\hat{\sigma})$ $\hat{\varepsilon}_{0.5ij} = (y_{ij} - \mathbf{x}_{ij}^T\hat{\boldsymbol{\beta}}_{0.5})$, and $\boldsymbol{\beta}_{\hat{\tau}_j}$ and $\boldsymbol{\beta}_{0.5}$ are replaced by the corresponding consistent estimators.

The proposed test can assist the decision to include or not cluster effects in the model. We note that the asymptotic result holds if $n_j \to +\infty$ for each $j = 1, \ldots, d$. Even though the test is asymptotically valid when the sample size within each group tends to infinity, we empirically show in Section 5 that it provides reasonable results in the small area estimation context as well. In Section 5 we explore the validity of this asymptotic result for different scenarios of the group-specific sample sizes.

The test we propose has a different aim to that of specification tests such as that recently proposed by Parente and Silva (2013) as we are not testing the assumptions needed for the estimation of $\boldsymbol{\beta}_\tau$ but whether units belonging to the same cluster are characterized by similar quantile coefficients, which is useful in prediction.

5 Simulation study

In this section we present results from three simulation studies used to investigate the finite sample properties of the tests we proposed in Section 3, the method for selecting the tuning constant c we proposed in Section 2.1 and the test statistic used for testing the presence of clustering in Section 4. Since these tests can be useful in small area estimation we generate data under linear mixed (random) effects models that incorporate area specific variation. The results for the Wald type test are not reported because they are very similar to the likelihood ratio type test. However, they are available to the prospective reader from the authors.

5.1 Likelihood Ratio type test

For evaluating the LR and Wald type tests for linear hypotheses on the MQ regression parameters, data is generated under the following mixed (random) effects model,

$$y_{ij} = \beta_0 + \beta_1 x_{ij1} + \beta_2 x_{ij2} + \beta_3 x_{ij3} + u_i + \varepsilon_{ij}, \quad i = 1, \ldots, n_j, \ j = 1, \ldots, d, \tag{28}$$

where j indexes the areas (clusters) and i units within areas. The regression coefficients are set as follows: $\beta_0 = 0$, $\beta_1 = 0.5$ and β_2, β_3 vary pairwise from 0 to 1, i.e. $(\beta_2, \beta_3) = (0, 0)$, $(\beta_2, \beta_3) = (0.25, 0.25)$, $(\beta_2, \beta_3) = (0.5, 0.5)$ and $(\beta_2, \beta_3) = (1, 1)$. The values of x_1, x_2 and x_3 are drawn from a Normal distribution with mean 5, 3 and 2, respectively and variance equal to 1. The number of small areas is set equal to $d = 20,\ 100$ and sample size in each small area $n_j = 5$, so we consider two different overall sample sizes: $n = 100,\ 500$. The error terms of the mixed model, u_i and ε_{ij}, are generated by using different parametric assumptions. Three settings for generating ε_i are considered,

1. Gaussian with mean 0, variance 1;

2. t-student distribution with 3 degrees of freedom (t_3);

3. Chi-squared errors with 2 degrees of freedom ($\chi^2(2)$).

T-students and Chi-squared random variables are re-scaled so to have variance equal to 1; in the case of chi-squared we substract the mean to generate zero-meaned residuals. The random effects are generated from a Normal distribution with mean 0 and $\sigma_u^2 = 0.43$. This entails that for all the scenarios the value of intraclass correlation is approximately equal to 0.3. These choices define a $4 \times 3 \times 2$ design of simulations. Each scenario is independently simulated $T = 10000$ times. MQ regression is fitted at $\tau = 0.5,\ 0.75,\ 0.90$ by using the Huber influence function with $c = 1.345$ for t-student and Chi-squared errors, $c = 100$ for

Gaussian errors and the maximum likelihood estimator (15) based on ALI as the estimator of σ_τ. Setting c equal to 1.345 gives reasonably high efficiency under normality and protects against outliers when the Gaussian assumption is violated (Huber, 1981). For the Gaussian scenario the resistance against outliers is not necessary and a large value for the tuning constant is preferred.

The results for the LR-type test for the null hypothesis

$$H_0 : \beta_{2\tau} = \beta_{3\tau} = 0$$

at the significance level $\alpha = 0.10,\ 0.05,\ 0.01$ are presented in Table 1. In all cases when $\beta_2 = \beta_3 = 0$ and the null hypothesis is true, the Type I error is very close to the nominal α, with small deviations in the case of $\tau = 0.9$ in the t_3 and $\chi^2(2)$ scenarios with $d = 20$ ($n = 100$) where the test turns out to be slightly conservative. For the Gaussian scenario, the power of the test tends to 1 as soon as the values of β_2 and β_3 increase, i.e. the null hypothesis is rejected for both sample sizes. In case of departures from normality, for example under the t_3 scenario, the value of the power of the test tends to 1 at $\tau = 0.5$ and 0.75 once the $\beta_2, \beta_3 = 0.25$ especially for $d = 100$ ($n = 500$). At $\tau = 0.9$ the likelihood ratio type test performs well as regression coefficients increase (as soon as $\beta_2, \beta_3 = 0.5$). Under the Chi-squared setting the test at $\tau = 0.75,\ 0.90$ appears to have lower power in rejecting the null hypothesis especially for the scenario with $d = 20$. Results for this scenario improve as the number of groups, d, and the values of the regression parameters (β_2, β_3) increase.

5.2 Choosing the tuning constant

In this Section we present results from a simulation study that is used to evaluate the estimation of the tuning constant c under the ALI distribution as proposed in Section 2.1. At each iteration of the algorithm the equations for $\boldsymbol{\beta}_\tau$, σ_τ, c are re-evaluated until

Table 1: Type I error and power of the proposed likelihood ratio type test under Gaussian, t_3 and $\chi^2(2)$ distributions at $\tau = 0.50,\ 0.75,\ 0.90$ with β_2, β_3 varying pairwise from 0 to 1, $\alpha = 0.10,\ 0.05,\ 0.01$ and $d = 20,\ 100$ with $n_j = 5$.

d	α	Gaussian, $c = 100$			t_3, $c = 1.345$			$\chi^2(2)$, $c = 1.345$		
		$\tau = 0.50$	$\tau = 0.75$	$\tau = 0.90$	$\tau = 0.50$	$\tau = 0.75$	$\tau = 0.90$	$\tau = 0.50$	$\tau = 0.75$	$\tau = 0.90$
		$(\beta_2, \beta_3) = (0, 0)$								
	0.10	0.110	0.114	0.133	0.103	0.114	0.147	0.109	0.120	0.181
20	0.05	0.059	0.062	0.075	0.050	0.063	0.089	0.057	0.064	0.112
	0.01	0.012	0.015	0.021	0.012	0.016	0.030	0.012	0.016	0.049
	0.10	0.101	0.105	0.109	0.102	0.108	0.122	0.103	0.106	0.126
100	0.05	0.052	0.058	0.058	0.052	0.053	0.063	0.050	0.055	0.069
	0.01	0.010	0.011	0.012	0.013	0.012	0.017	0.010	0.011	0.018
		$(\beta_2, \beta_3) = (0.25, 0.25)$								
	0.10	0.574	0.547	0.481	0.681	0.605	0.457	0.497	0.313	0.273
20	0.05	0.453	0.430	0.371	0.566	0.488	0.357	0.375	0.215	0.191
	0.01	0.245	0.225	0.192	0.337	0.267	0.191	0.184	0.088	0.082
	0.10	1.000	0.999	0.964	1.000	0.996	0.909	0.984	0.823	0.395
100	0.05	0.991	0.998	0.934	0.998	0.991	0.846	0.967	0.728	0.282
	0.01	0.962	0.989	0.914	0.991	0.968	0.671	0.903	0.498	0.128
		$(\beta_2, \beta_3) = (0.50, 0.50)$								
	0.10	0.978	0.962	0.920	0.993	0.982	0.852	0.944	0.729	0.449
20	0.05	0.960	0.941	0.873	0.987	0.961	0.784	0.905	0.619	0.352
	0.01	0.883	0.841	0.729	0.953	0.890	0.619	0.774	0.400	0.196
	0.10	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.872
100	0.05	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.795
	0.01	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.995	0.590
		$(\beta_2, \beta_3) = (1, 1)$								
	0.10	1.000	1.000	1.000	1.000	1.000	0.998	1.000	0.996	0.841
20	0.05	1.000	1.000	1.000	1.000	1.000	0.996	1.000	0.990	0.767
	0.01	1.000	1.000	1.000	1.000	1.000	0.985	1.000	0.965	0.604
	0.10	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
100	0.05	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	0.01	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

convergence. The data is generated under the following mixed (random) effects model,

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + u_i + \varepsilon_{ij}, \quad i = 1, \ldots, n_j, \;\; j = 1, \ldots, d, \tag{29}$$

where $\beta_0 = 1$, $\beta_1 = 2$, x follows a Uniform distribution $(0, 5)$, $d = 100$, $n_j = 5$ ($n = 500$). The error terms of the mixed model, u_i and ε_{ij}, are generated by using different parametric assumptions; the random effects u_i are generated from a Normal distribution with mean 0 and $\sigma_u^2 = 1$ and ε are drawn from different error distributions,

1. Gaussian with mean 0, variance 1;

2. t-student with 3 degrees of freedom (t_3);

3. Contaminated Normal with $\varepsilon \sim (1-\gamma)N(0,1)+\gamma N(0,25)$ where γ is an independently generated Bernoulli random variable with $Pr(\gamma = 1) = 0.1$, i.e. the individual errors are independent draws from a mixture of two normal distributions, with 90% on average drawn from a well-behaved $N(0,1)$ distribution and 10% on average drawn from an outlier $N(0,25)$ distribution;

4. Cauchy with location 0 and scale 1.

As in the previous section, the residuals are rescaled so their variance is equal to 1 and the value of intraclass correlation under different scenarios is always approximately equal to 0.3. Figure 1 shows the distribution, over 10000 Monte-Carlo samples of the estimated tuning constants for the four scenarios at $\tau = 0.25,\ 0.5,\ 0.75$. The horizontal dashed line represents the usual choice of $c = 1.345$. Under the Gaussian setting, the values of the tuning constants are clearly larger than the value 1.345 (the conventional value used in MQ regression) at each τ. The estimated value of the tuning constant suggests that using a robust estimator in this case is not justified as one would expect under the

assumptions we made in scenario 1. In contrast, the values of the estimated tuning constant are smaller than 1.345 in the contaminated and Cauchy scenarios. For instance, in the case of the contaminated scenario, the median value of the estimated tuning constant at $\tau = 0.5$ is 0.794. In the case of the Cauchy scenario the median value of the estimated tuning constant, at each quantile, degenerates to 0 because the Cauchy distribution has heavier tails than the exponential distributions and it should be truncated as the level of influential units becomes higher. For the t-student scenario the median value of the estimated tuning constant is 1.27 at $\tau = 0.5$ and it becomes higher than 1.345 (about 2.0) at $\tau = 0.25,\ 0.75$. In applications a unique c should be chosen; it can be the optimal one at 0.5 or chosen taking into consideration also optimal values at other quantiles.

5.3 Testing for the presence of clustering

In this section we present an empirical evaluation of the properties of the test used for the hypothesis of the presence of clustering and we show how this test can be useful in small area estimation context. For these simulations, data is generated under model (29). Two scenarios for the number of groups, d, are used, $d = 20$ and $d = 100$ and three scenarios for the within group samples size, $n_j = 5$, $n_j = 20$ and $n_j = 50$. The error terms of the mixed model, u_i and ε_{ij}, are generated by using different parametric assumptions. In particular, the random effects are generated from a Normal distribution with mean 0 and different scenarios for the level 2 variance components $\sigma_u^2 = 0,\ 1,\ 2.5,\ 7.5$. For $\sigma_u^2 = 0$, data is generated under the null hypothesis of no clustering. For the values of σ_u^2 other than 0 we start introducing clustering in the simulated data. Individual effects are generated according to Normal distribution with mean 0 and variance 5. When $\sigma_u^2 = 0$, i.e. under the null hypothesis, we empirically study the Type I error by using the proposed test. For all other scenarios of $\sigma_u^2 \neq 0$ we study the power of the proposed test. Each scenario is

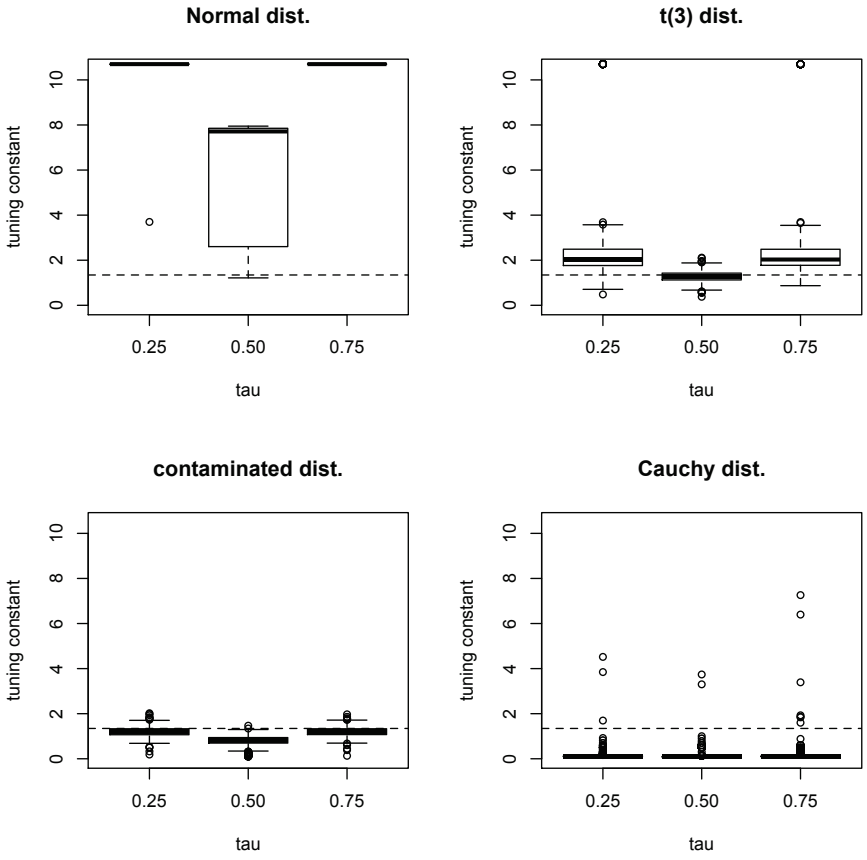

Figure 1: The distribution of the values of the tuning constant over Monte-Carlo samples and different settings for the error distribution at $\tau = 0.25,\ 0.50,\ 0.75$ and $d = 100$. The horizontal dashed line represents the choice of $c = 1.345$.

independently simulated $T = 10000$ times.

In this Monte-Carlo simulation, MQ regression is fitted by using the Huber influence function with $c = 100$ and the maximum likelihood estimator for the scale (15) under the ALI distribution. Table 2 reports the results of the simulation experiment. The Table shows the values of the intraclass correlation, $r = \sigma_u^2/(\sigma_u^2 + \sigma_\varepsilon^2)$, the Type I error and power of the proposed test statistic for $\alpha = 0.01,\ 0.05,\ 0.10$. To start with, we note that under the null hypothesis the Type I error is very close to the nominal value of α. As the value of σ_u^2 increases the power of the test increases too. The power increases more sharply for larger within cluster sample sizes. The number of clusters also seems to impact on the power of the test. The power of the test increases fairly sharply when we have a larger number of clusters even if each cluster consists of a small number of units. Under the null hypothesis we have also computed the empirical expected value and variance of the test statistic. We expect that, under the χ^2_{d-1} asymptotic approximation, the expected value of the test statistic will be equal to $d-1$ and the variance equal to $2\times(d-1)$. Results from the simulation studies confirm that the χ^2_{d-1} is a good approximation to the distribution of this test statistic. Finally, we have run a simulation where the individual effects are generated according to t-student with 3 degrees of freedom and the MQ regression is fitted by using the Huber influence function with $c = 1.345$. Also in this case under the null hypothesis the Type I error is very close to the nominal value of α and power of the test increases as the value of σ_u^2 increases. The detailed results are available to the interested reader from the authors.

The test can be used in small area estimation framework to detect the presence of area effects. If the test rejects H_0 it means that there is unobserved heterogeneity between areas and predictor (26) can be used to estimate the small area mean. Otherwise, if H_0 is not rejected, the synthetic estimator can be used for predicting the small area quantity

because, in the case of absence of unobserved heterogeneity between areas, it guarantees less variability and bias than estimator (26). To evaluate the performance of the synthetic predictor and the MQ predictor (26) the absolute relative bias (ARB) and the relative root mean squared error (RRMSE) of estimates of the mean value in each small area are computed. Table 3 reports the average values over areas of these indices for $n_j = 5,\ 20,\ 50$ and $d = 100$. The results for $d = 20$ are not reported because these are very similar to those for $d = 100$, but are available from from the authors upon request. Table 3 shows that the average ARB and RRMSE of the synthetic predictor increase as the intraclass correlation increases. The average values of ARB and RRMSE for estimator (26) remain constant at different values of r given the sample size. From the results in Table 3 it is apparent that when the assumption of significant between area heterogeneity is not rejected, the synthetic estimator offers the best performance. On the other hand, as soon as the intraclass correlation increases the predictor (26) performs best. Thus the LR-type test for the presence of clustering can drive the choice of the M-quantile predictor in small area estimation. The increase in the RRMSE when incorporating the area effect into prediction unnecessarily has been documented by other authors (see Datta et al., 2011). Our work extends these results to the case of small area estimation based on M-quantile regression.

6 Application

In this Section we use a dataset well-known in the small area estimation literature for illustrating the proposed model fit, selection and diagnostic criteria. Battese et al. (1988) analyse survey and satellite data for corn and soybean production for 12 counties in North Central Iowa. The dataset comes from the June 1978 Enumerative Survey, consists of 37 observations and includes information on the number of segments in each county, the

Table 2: Type I error and power of the proposed test statistic for clustering under Gaussian distribution with r varying between 0 and 0.6, α = 0.10, 0.05, 0.01, d = 20, 100 and n_j = 5, 20, 50.

α	$d = 20$			$d = 100$		
	$n_j = 5$	$n_j = 20$	$n_j = 50$	$n_j = 5$	$n_j = 20$	$n_j = 50$
			$r = 0$			
0.10	0.141	0.104	0.099	0.120	0.089	0.103
0.05	0.075	0.059	0.047	0.060	0.036	0.042
0.01	0.015	0.012	0.008	0.018	0.009	0.009
			$r = 0.16$			
0.10	0.702	0.999	1.000	0.983	1.000	1.000
0.05	0.565	0.998	1.000	0.969	1.000	1.000
0.01	0.325	0.992	1.000	0.906	1.000	1.000
			$r = 0.33$			
0.10	0.954	1.000	1.000	1.000	1.000	1.000
0.05	0.904	1.000	1.000	1.000	1.000	1.000
0.01	0.763	1.000	1.000	1.000	1.000	1.000
			$r = 0.60$			
0.10	0.999	1.000	1.000	1.000	1.000	1.000
0.05	0.998	1.000	1.000	1.000	1.000	1.000
0.01	0.989	1.000	1.000	1.000	1.000	1.000

Table 3: Values of the average ARB and average RRMSE over small areas for synthetic and (26) predictors under Gaussian distribution with r varying between 0 and 0.6, $d = 100$ and $n_j = 5,\ 20,\ 50$. Values are expressed as percentages.

Predictor	$n_j = 5$		$n_j = 20$		$n_j = 50$	
	ARB	RRMSE	ARB	RRMSE	ARB	RRMSE
			$r = 0$			
$\hat{m}_j^{MQ}$	11.07	13.62	5.66	7.04	3.55	4.45
$\hat{m}_j^{MQ/SYN}$	1.39	1.74	0.99	1.24	0.86	1.08
			$r = 0.16$			
$\hat{m}_j^{MQ}$	10.63	13.25	5.44	6.82	3.45	4.33
$\hat{m}_j^{MQ/SYN}$	11.41	14.29	11.20	14.02	10.84	13.58
			$r = 0.33$			
$\hat{m}_j^{MQ}$	10.54	13.20	5.60	7.10	3.73	4.87
$\hat{m}_j^{MQ/SYN}$	17.96	22.50	17.67	22.13	17.12	21.44
			$r = 0.60$			
$\hat{m}_j^{MQ}$	11.71	15.10	7.17	10.40	5.46	8.91
$\hat{m}_j^{MQ/SYN}$	31.07	38.92	30.59	38.31	29.65	37.13

number of hectares of corn and soybeans for each sample segment, the number of pixels classified by the LANDSAT satellite as corn and soybeans for each sample segment, and the mean number of pixels per segment in each county classified as corn and soybeans. These data were used by Battese et al. (1988) to predict the hectares of corn and soybean by county. We use this dataset to compute the tuning constant c, the R^2 goodness-of-fit measure, the LR-type test for specifying the explanatory variables to be included in MQ regression, and the likelihood ratio type test for the presence of clustering effects. Clusters are defined by the 12 counties in Iowa. Statistical properties of the small area estimators based on M-quantile regression have been discussed in other papers (see for instance Chambers et al., 2014a). County specific random effects were introduced by Battese et al. (1988) to improve prediction. Hence, the test for clustering looks at whether there is significant between county variation in the MQ-coefficients, something that would justify the inclusion of county specific effects.

The response variable y is the number of hectares of corn and soybeans and the model includes two fixed effects, x_1 and x_2 that represent the number of pixels classified by the LANDSAT satellite as corn and soybeans respectively for each sample segment. Battese et al. (1988) use the following two-level linear mixed model where i denotes the counties and j denotes the segments:

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + u_i + e_{ij}.$$

A random effect u_i is specified at the county level. This model will be used for benchmarking our results. Diagnostic for this model is reported in other papers (see for example Sinha and Rao, 2009). Those indicate that for the soybean variable normality of u and e approximately holds. For the corn variable, on the other hand, there is an influential outlier in the Hardin county.

We present results for MQ regression at $\tau = 0.05,\ 0.10,\ 0.25,\ 0.5,\ 0.75,\ 0.90,\ 0.95$. We

further compare our results at $\tau = 0.5$ to model diagnostics from the linear mixed model used by Battese et al. (1988). For the analysis of the corn outcome, the estimate of the tuning constant c using the ALI likelihood at $\tau = 0.5$ is equal to 1.94, a relatively low value, consistent with the presence of the outlier identified in diagnostic analysis. For the soybean variable the tuning constant c estimate at $\tau = 0.5$ is 7.85. This value suggests that there are no issues with contamination. Using $c = 1.345$, that represents a typical choice in the applications of the Huber loss function, or the value we chose for corn, would increase the robustness unnecessarily at the cost of lower efficiency. Similar conclusions hold for other values of τ.

Estimates of the scale parameter σ_τ obtained with the maximum likelihood method are shown in Figure 2. We note that these are sensitive to the M-quantile being considered and exhibit an inverted u-shape: for quantiles far from 0.5 the proportion of residuals for which $|u| > c$ is larger and this reduces their average size. When τ is close to 0.5 the estimates we obtain are close to those obtained by using the MAD estimator (9). On the contrary, MAD estimates are larger for quantiles far from 0.5 compared to those obtained in the central part of the distribution. This can be due to the fact that the scaling constant q in (9) should be quantile-adjusted. Looking at the R^2 model fit criterion we note that for the corn outcome this increases as τ increases (see Figure 2 solid line). For the soybean outcome there appears to be an almost constant high value of R^2 at all values of τ (see Figure 2 dashed line). Overall, for both outcomes there appears to be a moderate to strong linear relationship between the outcome and the explanatory variables at the different values of τ.

The LR-type tests results for the corn outcome are presented in Table 4 and for the soybean outcome in Table 5. When testing jointly the significance of x_1 and x_2, the tests suggest that these covariates are significant for explaining the variability in both

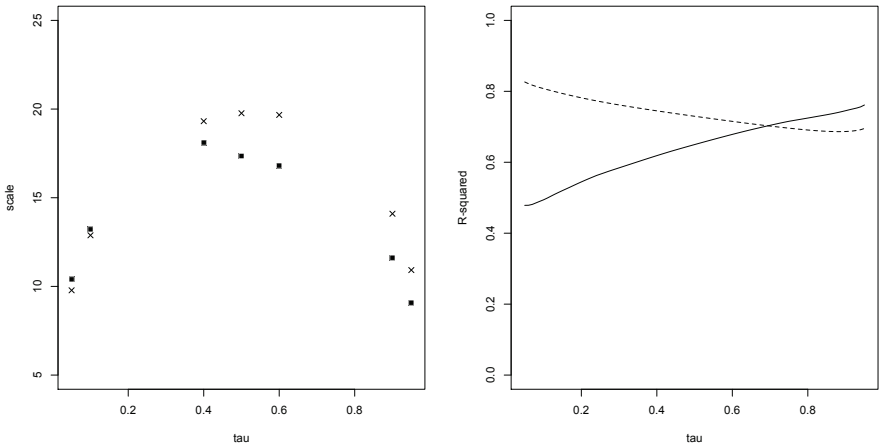

Figure 2: Left plot shows the values of the estimated scale at different value of τ for corn (■) and soybean (×). Right plot presents the R-squared at different value of τ for corn (solid line) and soybean (dashed line).

	$H_0 : (\beta_1, \beta_2) = 0$		$H_0 : \beta_2 = 0$	
τ	LR Test	p-value	LR Test	p-value
0.05	21.4	0.000	1.4	0.4935
0.10	23.8	0.000	0.3	0.8350
0.25	38.4	0.000	0.0	0.9996
0.50	68.3	0.000	0.4	0.7855
0.75	105.1	0.000	0.6	0.7376
0.90	97.1	0.000	0.1	0.9534
0.95	65.8	0.000	0.0	0.9959

Table 4: LR-type test for the model specification of the corn outcome, $H_0 : (\beta_1, \beta_2) = 0$ and $H_0 : \beta_2 = 0$

outcomes. For the corn outcome the tests show that after controlling for the number of pixels classified by the LANDSAT satellite as corn (x_1), the number of pixels classified by the LANDSAT satellite as soybean (x_2) is not significant. Similarly, for the soybean outcome after controlling for the number of pixels classified by the LANDSAT satellite as soybean (x_2), the number of pixels classified by the LANDSAT satellite as corn (x_1) is not significant. Hence, the model specification can be simplified by dropping the non-significant terms. The same conlcusions can be obtained by using the Wald-type test. For validating these results at $\tau = 0.5$, we run the same analysis under the two-level linear mixed model used by Battese et al. (1988). For the corn outcome after controlling for x_1, the p-value for including x_2 is equal to 0.6315 indicating that x_2 can be dropped from the model. For the corn outcome after controlling for x_2, the p-value for including x_1 is equal to 0.6049 indicating that x_1 can be dropped from the model.

We turn our attention to testing the significance of the between county variability. The

two scatter plots in Figure 3 show the relationship between the predicted county random effects computed with the mixed model and the MQ county coefficients computed with the MQ model for the corn outcome (scatter plot (a)) and the soybean outcome (scatter plot (b)). For both outcomes the two measures of county effects are well correlated. For testing the significance of the county MQ coefficients we use the proposed LR-type test. For the corn outcome the value of the test statistic is 17.152 and the corresponding p-value= 0.103. We have also conducted the hypothesis test for the presence of significant between county variation by using the linear mixed model. For testing the null hypothesis of a zero between county variation we compute the conditional-AIC (cAIC) value (Vaida and Blanchard , 2005) and compare this to the AIC value for a linear regression model without random effects. The cAIC for the linear mixed model is 327.5109 and the AIC for the linear regression model is 327.4116. This indicates that the linear model without random effects fits almost as well as the more complex model that includes random effects. Hence, random effects may not be needed in the analysis of the corn outcome.

For the soybean outcome the value of the LR-type test for the presence of clustering is 26.791 and the corresponding p-value= 0.0049. As in the case of the corn outcome, we have also conducted the hypothesis test for the presence of significant between county variation by using the linear mixed model. The cAIC for the linear mixed model is 311.8459 and the AIC for the linear regression model is 333.8107. This indicates that the linear model with county random effects fits better than the simpler model that ignores the random effects.

7 Final remarks

In this paper we have extended the available toolkit for inference in M-quantile regression. For given τ we have proposed a pseudo-R^2 goodness-of-fit measure, a likelihood ratio and

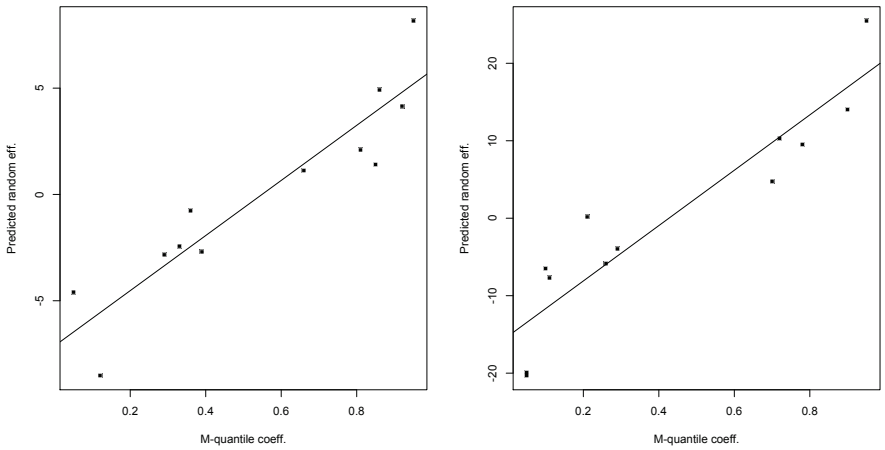

Figure 3: Scatter plots for the relationship between the predicted county random effects (computed with the mixed model) and the MQ county coefficients (computed with the MQ model) for the corn outcome (a) and for the soybean outcome (b).

	$H_0 : (\beta_1, \beta_2) = 0$		$H_0 : \beta_1 = 0$	
τ	LR Test	p-value	LR Test	p-value
0.05	195.7	0.000	2.6	0.2696
0.10	146.6	0.000	1.2	0.5496
0.25	116.0	0.000	0.3	0.8557
0.50	91.8	0.000	0.0	0.9972
0.75	66.7	0.000	0.4	0.8129
0.90	61.9	0.000	1.2	0.5380
0.95	65.3	0.000	01.6	0.4532

Table 5: LR-type test for the model specification of the soybean outcome, $H_0 : (\beta_1, \beta_2) = 0$ and $H_0 : \beta_1 = 0$

Wald type tests for testing linear hypotheses on the M-quantile regression parameters.

The cluster-specific M-quantile coefficients have been used for proposing a test for the presence of clustering in the data. The set of tests we present in the paper can be applied in small area estimation framework to validate the M-quantile models used for prediction. For a large class of continuously differentiable convex functions we showed the relationship between the loss function used in M-quantile regression and the maximization of a likelihood function formed by combining independently distributed GALI densities. Using this parametrization, we further propose an estimator of the scale parameter and a data-driven tuning constant to be used in the loss function. For each test the asymptotic theory has been developed involving recent works on inference by Wooldridge (2010) and Bianchi and Salvati (2015).

The simulation results for studying the finite sample properties of the model-fit criteria and the tests show that the Type I error of the LR-type test and the clustering test is very close to the nominal level α. For both tests, the results also indicate that the power tends to 1 as the values of the regression coefficients and the interclass correlation coefficient increase. In the simulation experiments we have also investigated the behaviour of the method proposed for estimating the tuning constant in the Huber loss function. The tuning constant derived by using the likelihood method is able to reflect different levels of contamination in the data.

References

Aragon, Y., Casanova, S., Chambers, R. and Leconte, E. (2005). Conditional Ordering Using Nonparametric Expectiles. *Journal of Official Statistics*, 21, 617–633.

Battese, G., Harter, R. and Fuller, W. (1988). An error component model for prediction

of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, 83, 28-36.

Bianchi, A. and Salvati, S. (2015), Asymptotic properties and variance estimators of the M-quantile regression coefficients estimators, to appear in *Communications in Statistics - Theory and Methods*, 44, 2016-2429.

Bottai, M., Orsini, N. and Geraci, M. (2015), A gradient search maximization algorithm for the asymmetric Laplace likelihood, *Communications in Statistics - Computation and Simulation*, 85, 1919-1925.

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

Chambers, R. and Tzavidis, N. (2006), M-quantile Models for Small Area Estimation, *Biometrika*, 93, 255-268.

Chambers, R., Chandra, H., Salvati, N. and Tzavidis, N. (2014a), Outlier robust small area estimation, *Journal of the Royal Statistical Society, series B*, 76, 47–69.

Chambers, R., Dreassi, E., Salvati, N. (2014b), Disease mapping via Negative Binomial regression M-quantiles, *Statistics in Medicine*, 33, 4805-4824.

Crainiceanu, C.M. and Ruppert, D. (2004), Likelihood ratio tests in linear mixed models with one variance components, *Journal of the Royal Statistical Society, ser. B* 66, 165–185.

Datta, G.S., Hall, P., Mandal, A. (2011), Model selection by testing for the presence of small-area effects, and application to area level data, *Journal of the American Statistical Association* 106, 362–374.

Dodge, Y. and Jureckova, J. (2000), Adaptive regression, Springer, New York

Dreassi, E., Ranalli, M.G., Salvati, N. (2014), Semiparametric M-quantile regression for count data, *Statitstical Methods in Medical Research*, 23, 591–610.

Fabrizi, E., Salvati, N., Pratesi, M., Tzavidis, N. (2014a), Outlier robust model-assisted small area estimation, *Biometrical Journal*, 56, 157–175.

Fabrizi, E., Salvati, N., Giusti, C., Tzavidis, N. (2014b), Mapping average equivalized income using robust small area methods, *Papers in Regional Science*, 93, 685-702.

Greven, S., Crainiceanu, C.M., Kuechenhoff, H., Peters, A. (2008), Restricted likelihood ratio testing for zero variance components in linear mixed models, *Journal of Computational and Graphical Statistics*, 17, 870–891.

Griva, I., Nash, S.G., Sofer, A. (2008), Linear and Nonlinear Optimization, Second Edition, SIAM.

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

Huber, P. J. (1973), Robust regression: Asymptotics, conjectures and monte carlo, *The Annals of Statistics*, 1, 799–821.

Huber, P. J. (1981), *Robust Statistics*, John Wiley & Sons, New York.

Huber, P. J. and Ronchetti, E.M. (2009), *Robust Statistics*, John Wiley & Sons, New York.

Koenker, R. (2005), *Quantile regression*, Economic Society Monographs, Cambridge University press, New York.

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

Koenker, R. and Machado, J.A.F. (1999), Goodness of Fit and Related Inference Processes for Quantile Regression, *Journal of the American Statistical Association*, 94, 1296–1310.

Kocic, P., Chambers, R., Breckling, J. and Beare, S. (1997) A measure of production performance. *Journal of Business and Economics Statistics*, 15, 445–451.

Newey, W.K. and Powell, J.L. (1987), Asymmetric least squares estimation and testing, *Econometrica*, 55, 819–847.

Parente, P.M.D.C. and Santos Silva, J.M.C. (2013), Quantile regression with clustered data, Economics Department Discussion Papers Series, Paper number 13/05.

Schrader, R.M. and Hettmansperger, T.P. (1980), Robust analysis of variance based upon a likelihood ratio criterion, *Biometrika*, 67, 93-101.

Sinha, S. K. and Rao, J. N. K. (2009), Robust small area estimation, *Canadian Journal of Statistics*, 37, 381-399.

Tzavidis, N., Marchetti, S., Chambers, R. (2010). Robust estimation of small-area means and quantiles. *The Australian and New Zealand Journal of Statistics*, 52, 167–186.

Tzavidis, N., Salvati, N., Chambers, R., Chandra, H. (2012), Small area estimation in practice: an application to agricultural business survey data. *Journal of the Indian Society of Agricultural Statistics*, 66, 213–238.

Tzavidis, N., Ranalli, M.G., Salvati, N., Dreassi, E., Chambers, R. (2015), Robust small area prediction for counts. *Statistical Methods in Medical Research*, 24, 373–395.

Vaida, F., Blanchard, S. (2005), Conditional Akaike information for mixed-effects models. *Biometrika*, 92, 351–370.

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

Wooldridge, J.M. (2010), *Econometric Analysis of Cross Section and Panel Data, 2nd edition*, The MIT Press, Cambridge (Mass.)

Yu, K. and Moyeed, R.A. (2001), Bayesian quantile regression, *Statistics and Probability Letters*, 54, 437–447.

Yu, K. and Zhang, J. (2005), A three-parameters asymmetric Laplace distribtuion and its extension, *Communication in Statistics: Theory and Methods*, 34, 1867–1879.

Appendix: Properties of the ALI

In this appendix we provide some more properties for special case of the GALI distribution when the $\rho(\cdot)$ is given by (3), that is the ALI we introduced in section 2.1. Suppose that U is a random variable with the standard ALI density ($\mu_\tau = 0, \sigma_\tau = 1$), then its cumulative distribution function is written as

$$F(u) = \begin{cases} \frac{1}{2c(1-\tau)B_\tau} exp\{[2cu + c^2](1-\tau)\} & u \leqslant -c \\ \frac{1}{B_\tau}\left\{\frac{1}{2c(1-\tau)}e^{-c^2(1-\tau)} + \sqrt{\frac{\pi}{1-\tau}}\Big[\Phi(u\sqrt{2(1-\tau)}) - \Phi(-c\sqrt{2(1-\tau)})\Big]\right\} & -c < u \leqslant 0 \\ \frac{1}{B_\tau}\left\{\frac{1}{2c(1-\tau)}e^{-c^2(1-\tau)} + \sqrt{\frac{\pi}{1-\tau}}\Big[\Phi(c\sqrt{2(1-\tau)}) - 1/2\Big] + \sqrt{\frac{\pi}{\tau}}\Big[\Phi(u\sqrt{2\tau}) - 1/2\Big]\right\} & 0 < u \leqslant c \\ \frac{1}{B_\tau}\left\{\frac{1}{2c\tau}e^{-c^2\tau} - \frac{1}{2c\tau}exp\{-2\tau cu + c^2\tau\}\right\} & u > c \end{cases}.$$

For obtaining the expected value and the variance of U, the moment generating function is computed and it can be written as:

$$M_\tau(t) = \frac{1}{B_\tau[2c(1-\tau)+t]}exp\{-c^2(1-\tau) - ct\}$$

$$+\frac{exp\{\frac{t^2}{4(1-\tau)}\}}{B_\tau}\sqrt{\frac{\pi}{(1-\tau)}}\Big[\Phi\left(-\frac{t}{\sqrt{2(1-\tau)}}\right) - \Phi\left(\frac{-2c(1-\tau)-t}{\sqrt{2(1-\tau)}}\right)\Big]$$

$$+\frac{exp\{\frac{t^2}{4\tau}\}}{B_\tau}\sqrt{\frac{\pi}{\tau}}\Big[\Phi\left(\frac{2c\tau - t}{\sqrt{2\tau}}\right) - \Phi\left(-\frac{t}{\sqrt{2\tau}}\right)\Big] - \frac{1}{B_\tau(t-2c\tau)}exp\{-c^2\tau + ct\},$$

for $-2c(1-\tau) < t < 2c\tau$.

The first moment then is

$$E(U) = -\frac{1}{4B_\tau c^2(1-\tau)^2}exp\{-c^2(1-\tau)\} + \frac{1}{4B_\tau c^2\tau^2}exp\{-c^2\tau\} + \frac{1-2\tau}{2\tau(1-\tau)B_\tau}$$

and the variance is

$$Var(U) = \frac{1}{B_\tau}\Bigg[e^{-c^2(1-\tau)}\frac{1+2c^2(1-\tau)}{4c^3(1-\tau)^3} + e^{-c^2\tau}\frac{1+2c^2\tau}{4c^3\tau^3} + \frac{1}{2}\frac{\sqrt{\pi}\Big[\Phi(c\sqrt{2\tau}) - 0.5\Big]}{\tau^{3/2}}$$

$$+\frac{1}{2}\frac{\sqrt{\pi}\Big[\Phi(c\sqrt{2(1-\tau)})-0.5\Big]}{(1-\tau)^{3/2}}\Bigg].$$

These formulae may be easily generalized to the location and scale case. They can be used to obtain method of moments estimates of c and σ_τ to be used as initial values when minimizing (12) when $\rho_\tau(\cdot)$ is the Huber loss function, in line with Yu and Zhang (2005). The computations for obtaining the moment generating function, the expected value and the variance of U are not reported in the paper, but they are available from the authors upon request.