Quantile and Restricted Maximum Likelihood Approach for Robust Regression of Clustered Data

May Ann S. Estoy
Visayas State University

Joseph Ryan G. Lansangan
School of Statistics, University of the Philippines Diliman

Quantile regression and restricted maximum likelihood are incorporated into a backfitting approach to estimate a linear mixed model for clustered data. Simulation studies covering a wide variety of scenarios relating to clustering, presence of outliers, and model specification error are conducted to assess the performance of the proposed methods. The methods yield biased estimates yet high predictive ability compared to ordinary least squares and ordinary quantile regression.

Keywords: linear mixed models; quantile regression; restricted maximum likelihood; backfitting; bootstrap; clustered data

1. Introduction

There is an increasing awareness on the negative effect of outliers to the classical statistical methods. However, despite this awareness, robust methods remain mostly unused or even unknown by most data analysts (Maronna et al., 2006). Even when robust methods provide good fit to the data with little to no influence coming from outliers and without necessarily identifying which observations deviates from the rest, classical methods are commonly preferred due to ease of computation and interpretability (Maronna et al., 2006).

Linear regression modeling via ordinary least squares (OLS) requires certain model assumptions (e.g., linearity, uncorrelatedness, homoscedasticity) and data quality (e.g., outlier- and/or influencer-free, acceptable collinearity) to derive the best regression estimators. Such model assumptions and data quality however, are not always satisfied. Some data from fields including biostatics, econometrics, and social sciences for example, frequently fail the assumption of homoscedasticity (Hao & Naiman, 2007; Abdullahi & Yahaya, 2015). In some circumstances, extreme and unpredictable data points are observed, or an incorrect model is specified. In these cases, problem in the estimation of the statistical model arise. Quantile regression, which is known to have robustness property, can be considered as an alternative to model such data.

Quantile regression as presented by Koenker & Bassett (1978) has several advantages over the OLS. It is less sensitive to extreme outliers, and it captures not only the measure of the central tendency but also the information about the tails of a distribution. Even with heteroscedasticity issues, the estimates produced are considered optimal. Quantile regression for this reason presents a better and more complete picture of the linear relationship (Koenker & Bassett, 1978; Koenker & Hallock, 2001; Maronna et al., 2006).

In this paper, an approach for robust estimation in a simple linear model is presented. The hypothesized model is postulated as a linear mixed model (LMM) with the intercept assumed as random cluster effect. To estimate random (intercept) and fixed (slope coefficient) effects in the linear model, the approach incorporates quantile regression and restricted/residual maximum likelihood (REML) under a backfitting framework and then utilizing bootstrapping. Specifically, the effects/parameters of the model are estimated via the quantile regression and REML in backfitting (QRB), and the bootstrapped quantile regression and REML in backfitting (BootQRB). To assess the performance of the methods in the presence of clustering, model misspecification and outliers, the methods are compared to the OLS regression and ordinary quantile regression (OQR) in terms of percentage bias (PBias) and mean absolute percentage error (MAPE).

2. Robust Estimation and Quantile Regression

Robustness

Wolters and Kateman (1989) defined robustness as the insensitivity to small deviations from the assumptions. Huber (1981) claims that a robust procedure should possess the following features: it should have a reasonably good efficiency for the assumed model; small deviations from the model assumptions should impair the performance of the procedure only slightly; and larger deviations from the model should not cause a catastrophe.

One important role of robust statistics is in the identification and proper handling of outliers (Filzmoser and Rousseeuw, 2002). An outlier must not be disregarded as it might be a legitimate and most important observation of the sample (Hampel, 2001). One possible source for outlier is a wrong choice/formulation of a statistical model (Hampel et al., 2005). Sources of model misspecification might be through omission of a relevant independent variable, inclusion of an irrelevant coefficient, and misidentification of an independent variable (e.g. treating a homoscedastic normal variable into a heteroscedastic normal variable, and vice versa) in the model (Light, 2010; Lo, 2011). This may result to misleading inferences about the parameters (Lee, 2014). Robust methods can address this issue by effectively retaining useful factors in the model and eliminate or minimize the effect of irrelevant ones (Gospodinov et al., 2014). A robust estimation method can give reasonable results in the presence of outliers or model misspecification, while still maintaining efficiency in case there is no contamination (Koller, 2013).

Quantile regression

As a robust method in linear modeling, quantile regression is a statistical technique intended to estimate and conduct inference about the conditional quantile functions. Quantile regression offers a mechanism for estimating models for conditional median function, as well as the full range of other conditional quantile functions. The quantiles are the results of minimizing a sum of asymmetrically weighted absolute residuals. The weights or penalties are given to the positive and negative residuals by different complementing values (Koenker & Hallock, 2000).

In a simple linear model, the quantile regression estimate is derived from

$$\hat{\beta}_\tau = argmin_{\beta\in\mathbb{R}^p} \sum_{i=1}^{n} \rho_\tau (y_i - x_i\beta) \tag{1}$$

where y_i is the response and x_i the covariate (or independent variable) for the i^{th} observation and $\hat{\beta}_\tau$ is the parameter estimate of the τ^{th} quantile. Also, $\rho_\tau(\cdot)$ is the penalized residual, i.e. if we let $z = y_i - x_i\beta$, then $\rho_\tau(z) = z(\tau - I(z<0)), 0<\tau<1$, and $I(\cdot)$ denotes the indicator function. This can be viewed as a natural extension of ordinary least squares (Koenker & Hallock, 2000).

Quantile regression overcomes various OLS shortcomings. Even without outliers, the effects of the covariates on the response variable may vary for different subsections of the sample. Quantile regression can suggest which of the subsections/conditional distributions differ and thus gives a better and more complete view of the relationship among random variables (Abdullahi & Yahaya, 2015).

3. Linear Mixed Models and REML

Linear mixed models

Linear mixed models (LMMs) are extensions of linear models that include both fixed and random effects. The general form of an LMM with a single independent variable is

$$y_i = x_i\,\beta + z_i\,\delta + \varepsilon_i, \tag{2}$$

where y_i is the response variable, x_i is the independent or predictor variable with fixed effect regression coefficient β, z_i is the design variable (i.e., a random complement to x_i) with random effects δ, and ε_i is the stochastic error. The random effects δ and the errors ε_i are both assumed to follow a normal distribution and are independent of each other. The residual of the i^{th} observation $resid_i = y_i - x_i\hat{\beta} - z_i\hat{\delta}$ is a mixture of both the observation level errors and the random effects. Thus, an objective function which accounts for the observation level residuals and random effects separately is suitable, so the effect of model misspecification is separated to the random effects and residual errors (Koller, 2013).

Restricted maximum likelihood

Restricted/residual maximum likelihood (REML) is an estimation procedure for LMMs. It is an improved method of the maximum likelihood (ML) estimation of variance components. In LMMs, REML maximizes the likelihood that only depends on the variance components by conditioning on the fixed effects (Bates, 2014). The REML estimators of the variance components take into account the loss in the degrees of freedom needed to estimate the fixed effects. The variance component estimator is approximately unbiased in the REML procedure, whereas the ML approach is negatively biased. Thus, REML approach is more preferred to ML estimation method (Harville, 1977).

There are many iterative algorithms that can be considered for computing the REML estimates. There is no single iterative numerical algorithm suitable for every application. An algorithm with fast convergence in one setting may converge slowly or even fail to converge in another. Thus, in choosing which of the available algorithms for REML is best to use, one should consider the computational requirements and other properties as applied to the given conditions of the model (for details see Harville, 1977).

4. Backfitting and Bootstrapping

The backfitting algorithm

A generalization of the usual linear regression model is an additive model expressed as the sum of functions. One estimation procedure for an additive model is the backfitting algorithm. As defined by Hastie and Tibshirani (1990), it is a general algorithm that enables one to fit an additive model using any regression-type fitting mechanisms.

As stated by Hardle et al. (2004), the key idea of backfitting is to regress the additive components separately on partial residuals. Consider the parametric model

$$E(y_i \mid x_{ij}, z_{ij}) = x_i\beta + z_i\delta \tag{3}$$

which consists of only two additive components. Let $\rho(\cdot)$ be any regression procedure fit for the separate additive models. Backfitting then iteratively solves for $x_i\beta = \rho\ (y_i - z_i\delta)$ and $z_i\delta = \rho\ (y_i - z_i\beta)$.

With the model of interest being a linear mixed model, the modified backfitting algorithm of Hardle et al. (2004) which was previously established by Hastie and Tibshirani (1990) is a better option (among other backfitting algorithms). The idea is to start each iteration step with a parametric linear least squares regression on the explanatory variable/s. Then smooth the fit by using a smoother to the partial residuals. This in turn yields updated estimates of the additive function. Hastie and Tibshirani (1990) (also in Hardle et al., 2004) claims that such algorithm makes it easier to characterize a unique solution and it eliminates the dependence of the final results on the initial function specification (in backfitting).

Bootstrap methods

The bootstrap approach is a method for assigning measures of accuracy to statistical estimates (Efron and Tibshirani, 1994). The original aim in developing this resampling procedure is to better understand the jackknife (an earlier resampling method) through deriving properties of its bootstrap estimates. It turns out that bootstrap is even better as a resampling procedure than the jackknife in many circumstances (Chernick & LaBudde, 2011). In the simplest independent but not identically distributed settings, the standard approach is the so-called (x,y)-pair nonparametric bootstrap. Pairs (x_i, y_i) $i = 1,\ldots, n$ are drawn at random from the original observations with replacement. For each resampling the estimator is recomputed. Repeating this procedure R times yields a sample of R p-vectors whose sample estimate is a valid estimator of the original one (Koenker and Hallock, 2000).

There are other bootstrap procedures available such as parametric bootstrap, double bootstrap, m-out-of-n bootstrap, wild bootstrap, and block bootstrap, to name a few (Chernick & LaBudde, 2011). And in recent studies, Gospodinov et al. (2014) and Lee (2014) suggest that a Monte Carlo simulation or bootstrap method is robust to model misspecification.

5. Postulated Model and Estimation Procedures

The postulated model

A linear mixed model with clustered data is postulated as

$$y_{ij} = \beta x_{ij} + \delta_j + \varepsilon_{ij},\ \ i = 1,\ldots, m,\ \ j = 1,\ldots, k. \tag{4}$$

where y_{ij} is a continuous response variable and x_{ij} is a predictor variable for the i^{th} observation in the j^{th} cluster; β is an unknown parameter; δ_j is a random cluster effect; and ε_{ij}'s are unobserved random errors, which are independently and identically distributed as $N(0,\sigma^2)$. Observations belong to a cluster $j = 1,\ldots, k$ and indexed by $i = 1,\ldots, m$ within their clusters, so that k is the number of clusters, m is the number of observations per cluster, and $n = km$ is the total number of observations. Equivalently, the τ^{th} quantile linear mixed model with clustered data is postulated as

$$y_{ij} = \beta^{(\tau)} x_{ij} + \delta_j^{(\tau)} + \varepsilon_{ij}, \;\; i = 1, ..., m, \;\; j = 1, ..., k. \tag{5}$$

Note that the OLS and OQR fit the models $yi = \alpha + \beta x_i + \varepsilon_i$ and $y_i = \alpha^{(\tau)} + \beta^{(\tau)} x_i + \varepsilon_i$, respectively, where α, β, $\alpha^{(\tau)}$ and $\beta^{(\tau)}$ are the unknown parameters. However, for both models, α and $\alpha^{(\tau)}$ are treated as fixed and may represent (conservatively) the random cluster effect/parameter δ_j in equations (4) or (5).

The quantile regression and REML in backfitting (QRB) approach

A parametric backfitting algorithm is proposed to estimate the coefficients of a quantile regression mixed model in equation/model (5). The properties of model (5) are the same with model (4) except with the coefficients being dependent on the quantile τ. To deal with the presence of mixed effects, quantile regression and REML were fused to estimate the fixed effect $\beta^{(\tau)}$ and the random effect $\delta_j^{(\tau)}$, respectively, within the backfitting algorithm. The proposed estimation procedure is shown in Algorithm 5.1.

Algorithm 5.1. The QRB algorithm

1. Initialization:
   (a) Fit the quantile regression model $y_{ij} = \delta^{(\tau)} + \beta^{(\tau)} x_{ij} + \varepsilon_{ij}$ to obtain the initial estimates $\hat{\delta}_{(0)}^{(\tau)}$ and $\hat{\beta}_{(0)}^{(\tau)}$. Note that $\delta_j^{(\tau)}$ may be considered fixed, that is, all clusters have the same effects.
   (b) Get $\hat{y}_{ij}^{(0)} = \hat{\delta}_{(0)}^{(\tau)} + \hat{\beta}_{(0)}^{(\tau)} x_{ij}$.
2. Iteration: For $l = 1,2,3,\ldots, L$, where L is the final iteration
   (a) Fit the quantile regression model $\hat{y}_{ij}^{(l-1)} = \beta^{(\tau)} x_{ij} + \varepsilon_{ij}$, ignoring $\delta_j^{(\tau)}$ to obtain the estimate $\hat{\beta}_{(l)}^{(\tau)}$. Get residuals as $\hat{\varepsilon}_{ij}^{(l)} = \hat{y}_{ij}^{(l-1)} - \hat{\hat{y}}_{ij}^{(l-1)}$ where $\hat{\hat{y}}_{ij}^{(l-1)} = \hat{\beta}_{(l)}^{(\tau)} x_{ij}$.
   (b) Fit the model $\hat{\varepsilon}_{ij}^{(l)} = \delta_j^{(\tau)} + \varepsilon_{ij}$ using REML to obtain the estimate $\hat{\delta}_{j(l)}^{(\tau)}$.
   (c) Get $\hat{y}_{ij}^{(l)} = \hat{\beta}_{(l)}^{(\tau)} x_{ij} + \hat{\delta}_{j(l)}^{(\tau)}$.
3. Stopping Criterion:
   (a) Continue Step 2 replacing $\hat{y}_{ij}^{(l-1)}$ with $\hat{y}_{ij}^{(l)}$ until the estimate of y_{ij} does not change or until $\sum \left| \hat{y}_{ij}^{(l)} - \hat{y}_{ij}^{(l-1)} \right|$ <0.0001.

Bootstrapped quantile regression and REML in backfitting (BootQRB)

The bootstrap algorithm works by drawing many independent bootstrap samples, evaluating the corresponding bootstrap replications, and estimating the parameters. The bootstrap procedure is integrated with the QRB algorithm, as presented in Algorithm 5.2.

Algorithm 5.2. The BootQRB algorithm

1. Using the realizations $\{(y_{ij},x_{ij})\}$, where cluster j = 1,…, k and indexed by i = 1,…, m within their cluster, separately draw for each cluster a simple random sample of size m with replacement. The bootstrap sample is then the merged outcomes from all clusters.
2. Evaluate the bootstrap sample in (1) using Algorithm 5.1. This gives the value of the parameter estimates.
3. Repeat (1) and (2) R times, yielding R statistics $\hat{\beta}^{(\tau)}$ and $\hat{\delta}_j^{(\tau)}$.
4. Obtain bootstrap estimates by computing the mean of the R statistics.

6. Simulation Studies

The proposed estimation procedures are assessed through simulation studies. A summary of the different simulation settings is shown in Table 6.1. Sets of simulations are carried out to examine the behavior of the estimation procedures under effects of clustering, model misspecification, and presence of outliers.

Table 6.1 Parameters for Simulation Settings

1. Distribution of δ_j	$\delta_j \sim N(d_j, \sigma_{d_j}$, $d_j = 0$, increases by 2 for the succeeding clusters; and σ_{d_j}=0.5, the same for all clusters
2. No. of clusters, k	small: 3 clusters; large: 10 clusters
3. Cluster size, m	small size: 8 observations; large size: 30 observations
4. Distribution of x_{ij}	$x_{ij} \sim N(30,3^2)$
5. Value of β in βx_{ij}	$\beta = 1$ for 3 clusters; $\beta = 3$ for 10 clusters
6. Distribution of ε_{ij}	$\varepsilon_{ij} \sim N\,(0,1)$
7. Misspecification w in the model	$w = 1$ without misspecification; $w = 5$ with misspecification (w as multiplier of ε_{ij})
8. Outliers in the data	a. frequency of outliers – small: 3 outliers for $k=3$; 10 outliers for $k=10$; large: 9 outliers for $k=3$; and 30 outliers for $k=10$ b. value of outlier: $\bar{y}$ is multiplied by v with i. same values: v=0.5, relatively small-valued outlier, and v=0.25, extremely small-valued outlier; $v=1.5$, relatively large-valued outlier, and v=2, extremely large-valued outlier ii. varying values: $v=\{0.25,0.5,0.75\}$ and $v=\{1.5,1.75,2\}$ for 3 clusters; $v=\{0.2,0.35,0.5,0.65,0.8\}$ and $v=\{1.25,1.5,1.75,2,2.25\}$ for 10 clusters c. scope of outlier relative to the clusters i. wide: spread to all clusters ii. narrow: spread to 66% out of the 3 clusters; and spread to 50% out of the 10 clusters iii.narrower: spread to 33% out of the 3 clusters; and spread to 30% out of the 10 clusters

The evaluating measures to verify the robustness of the proposed methods to model misspecification and data contaminated with outliers are the mean absolute percentage error (MAPE), MAPE percentage difference relative to benchmark (*%diff*), and percentage bias (PBias). As an accuracy measure, the MAPE is computed as $MAPE = \frac{\sum\left|\left(y_i - \hat{y}_i\right)/y_i\right|}{n} * 100\%$ where y_i is the actual value and $\hat{y}_i$ is the estimate of the response variable of the i^{th} observation. The percentage difference of MAPE of the proposed method relative to the benchmark method is $\%diff = \frac{MAPE_{(Benchmark)} - MAPE_{(Proposed\ Method)}}{MAPE_{(Benchmark)}} * 100\%$, where a positive value indicates a better predictive ability of the proposed method compared to the benchmark method. PBias measures the average tendency of the estimated values to be larger or smaller than the true one, and is calculated as

$$PBias = \left[\frac{E(\hat{\theta}) - \theta}{\theta} 1\right] * 100\% \ .$$

7. Results and Discussions

Clean data

The generated clean data is free from outliers and model misspecification. Estimates obtained from OLS regression are then considered ideal. Hence, the proposed methods are evaluated and compared to OLS regression. Presented in Table 7.1 are the MAPE of the mean or median estimates together with the percentage differences of the MAPE of OQR, QRB, and BootQRB with that of the OLS. The MAPEs of QRB and BootQRB are around 45% to 78% better than that of OLS. The median estimates produced from QRB and BootQRB are more optimal than the mean and median estimates obtained from OLS and OQR, respectively.

Table 7.1 MAPE for the mean/median estimates and Percentage difference (in parenthesis)

No. of clusters	Cluster size	MAPE			
		OLS	OQR	QRB	BootQRB
3	8	5.1692	5.0380 (2.53%)	2.7329 (47.13%)	2.7324 (47.14%)
	30	5.2111	5.1741 (0.70%)	2.8162 (45.95%)	2.8164 (45.95%)
10	8	5.1149	5.0843 (0.59%)	1.0966 (78.56%)	1.0945 (78.60%)
	30	5.1582	5.1495 (0.16%)	1.1549 (77.61%)	1.1507 (77.69%)

Table 7.2 shows that the biases of the parameter estimates ($\hat{\beta}$) of the median quantile under QRB and BootQRB are higher than the benchmark method. Although the predictions are relatively better, the proposed methods tend to overestimate β by a relatively small amount (note that $\beta = 1$ or $\beta = 3$). It is apparent from Tables 7.1 and 7.2 that OLS and OQR favor better estimation, while QRB and BootQRB favor better prediction.

Table 7.2 PBias of $\hat{\beta}$ for the mean/median estimates

No. of clusters	Cluster size	PBias of $\hat{\beta}$			
		OLS	OQR	QRB	BootQRB
3	8	0.8785	0.1915	6.6160	6.6190
	30	1.0455	1.0345	6.6285	6.6285
10	8	0.2405	0.1275	9.9120	9.8840
	30	0.2040	0.5035	10.0640	10.0005

Misspecified model

The MAPE under the misspecified model scenarios are presented in Table 7.3. With model misspecification, the proposed methods have better predictive measures. Also, as in the clean data scenario, the proposed methods tend to overestimate β (see Table 7.4).

Table 7.3 MAPE for the mean/median estimates and Percentage difference (in parenthesis)

No. of clusters	Cluster size	MAPE			
		OLS	OQR	QRB	BootQRB
3	8	13.8169	13.4947 (2.33%)	12.9886 (5.99%)	12.8436 (7.04%)
	30	13.9572	13.8684 (0.63%)	13.1608 (5.70%)	13.1484 (5.79%)
10	8	6.3357	6.2925 (0.68%)	3.9157 (38.19%)	3.9088 (38.30%)
	30	6.4035	6.3919 (0.18%)	4.1001 (35.97%)	4.0989 (35.98%)

Table 7.4 PBias of $\hat{\beta}$ for the mean/median estimates

No. of clusters	Cluster size	PBias of $\hat{\beta}$			
		OLS	OQR	QRB	BootQRB
3	8	0.40300	-0.3360	6.4455	6.5045
	30	1.11400	1.2885	6.5260	6.5340
10	8	0.46250	0.7985	9.8220	9.7985
	30	0.10650	0.1550	10.0240	9.9510

Data with outliers

It is known that OQR generally provides better results for the conditional mean/median than OLS in the presence of outliers. Thus, the OQR is considered as benchmark for computing the MAPE percentage differences. Table 7.5 presents the MAPE and percent differences. Results suggest that QRB and BootQRB offer an improvement over OQR in terms of prediction. Such improvement is more noticeable for large number of clusters and/or large cluster sizes.

Table 7.5. MAPE for the mean/median estimates and Percentage difference (in parenthesis)

No. of clusters	Cluster size	MAPE			
		OLS	OQR	QRB	BootQRB
3	8	19.9805	18.0610	18.2181 (-.87%)	18.0494 (0.06%)
	30	13.0194	12.1718	11.2921 (7.23%)	11.2292 (7.74%)
10	8	20.1409	18.2973	17.5927 (3.85%)	17.2731 (5.60%)
	30	13.6236	12.8121	11.2165 (12.45%)	11.1421 (13.03%)

As seen in Table 7.6 below, OLS consistently underestimates the parameter β, while QRB and BootQRB generally overestimates the parameter. OQR on the other hand is clearly is performing in terms of estimation.

Table 7.6. PBias of $\hat{\beta}$ for the mean/median estimates

No. of clusters	Cluster size	PBias of $\hat{\beta}$			
		OLS	OQR	QRB	BootQRB
3	8	-10.8028	0.4523	8.8458	8.6758
	30	-5.3414	1.0314	7.9866	7.9174
10	8	-12.1559	0.2174	12.2847	12.2549
	30	-7.0576	0.3394	11.5808	11.4975

As expected, increasing the frequency of outliers (i.e. from 3 to 9 and from 10 to 30 outliers) leads to larger values of MAPE and more biased (higher magnitude of PBias) estimates of β for the proposed methods (see Table 7.7). With small frequency of outliers, the MAPE of the two proposed methods outperform the OQR, but all three are less-performing under large frequency of outliers.

Table 7.7. Evaluation Measures on the Mean/Median Estimates by Frequency of Outliers

Evaluating Measure	Frequency of Outliers	OLS	OQR	QRB	BootQRB
MAPE	small	9.5791	9.2800	7.4941	7.3848
	large	17.0639	15.7039	15.0145	14.9865
PBias of $\hat{\beta}$	small	-3.2153	0.6750	9.2450	9.2346
	large	-9.1838	0.6958	10.3223	10.1803

Interestingly, having small-valued outliers greatly affects the MAPE across all methods (see Table 7.8). Higher values of MAPE are observed under scenarios with small-valued data outliers compared to those with large-valued outliers. But the percentage bias behaves differently – higher percentage biases for QRB and BootQRB are observed for data with large-valued outliers.

Table 7.8. Evaluation Measures on the Mean/Median Estimates by Magnitude of outliers

Evaluating Measure	Frequency of Outliers	OLS	OQR	QRB	BootQRB
MAPE	small	20.4711	20.5217	17.1203	17.0337
	large	10.6647	8.2537	9.8223	9.6546
PBias of $\hat{\beta}$	small	-8.0177	0.5919	4.7316	4.9902
	large	-7.9013	0.5452	15.3568	14.9299

The performances of QRB and BootQRB are influenced by the presence of either relatively far or extremely far outliers in the data. When the values of the outliers were increased, the MAPE and PBias of the two proposed methods also increased (see Table 7.9 below). Also, there is not much implication in the performance of the proposed methods about having same or varying values and about having a narrow or wide scope of outliers present in the data (see Tables 7.10 and 7.11). Nevertheless, compared to OQR, the proposed methods generally performed well in terms of prediction and relatively poorly in terms of estimation.

Table 7.9. Evaluation Measures on the Mean/Median Estimates by Extent of Outliers

Evaluating Measure	Extent of outlier	OLS	OQR	QRB	BootQRB
MAPE	relatively far	11.02921	10.39845	9.14789	9.03245
	extremely far	21.61581	19.87526	19.14996	19.01762
PBias of $\hat{\beta}$	relatively far	-7.93515	0.60365	9.25488	9.27426
	extremely far	-7.92594	0.60640	10.38033	10.31378

Table 7.10 Evaluation measures on the Mean/Median Estimates by Type of Outliers

Evaluating Measure	Type of outlier	OLS	OQR	QRB	BootQRB
MAPE	same	16.3225	15.1368	14.1489	14.0250
	varying	14.0586	12.8893	12.1161	11.9825
PBias of $\hat{\beta}$	same	-7.9305	0.6050	9.8176	9.7940
	varying	-8.0173	0.4956	10.4974	10.2922

Table 7.11 Evaluation Measures on the Mean/Median Estimates by Scope of Outliers

Evaluating Measure	Scope of outliers	OLS	OQR	QRB	BootQRB
MAPE	wide	15.3602	14.1950	14.2746	14.1310
	narrow	15.6987	14.5446	13.8837	13.7441
	narrower	15.6447	14.4234	12.2556	12.1575
PBias of $\hat{\beta}$	wide	-7.8423	0.5490	9.2154	9.1002
	narrow	-8.1229	0.4940	9.5927	9.7371
	narrower	-7.9133	0.6626	11.3245	11.0430

Other quantiles

Under the median quantile, results show that the QRB and BootQRB methods are relatively off by about 6% to 15% (depending on which case) in the estimation of β and thus seem inferior to OQR. Looking at other quantiles however, the advantages of QRB and BootQRB become evident (see Tables 7.12 to 7.14). The percentage biases of QRB and BootQRB are way better than those of OQR, suggesting that the new methods are better in terms of characterizing the tails (lower or upper quantiles) of the linear relationship.

Table 7.12 PBias of $\hat{\beta}$ for the 5th percentile estimates under Misspecification scenario

No. of clusters	Cluster size	PBias of $\hat{\beta}$		
		OQR	QRB	BootQRB
3	8	3.4800	2.0920	0.2265
	30	-3.7520	0.8645	0.4655
10	8	-0.7505	0.6195	0.0320
	30	-0.2185	0.2690	0.1280

Table 7.13 PBias of $\hat{\beta}$ for the 5th percentile estimates under Small Outliers scenario

No. of clusters	Cluster size	PBias of $\hat{\beta}$		
		OQR	QRB	BootQRB
3	8	-77.0782	-10.2686	-12.1581
	30	-46.8922	-6.7119	-6.8666
10	8	-93.7786	-9.8200	-16.7268
	30	-57.5828	-7.3321	-9.3052

Table 7.14 PBias of $\hat{\beta}$ for the 95th percentile estimates under Large Outliers scenario

No. of clusters	Cluster size	PBias of $\hat{\beta}$		
		OQR	QRB	BootQRB
3	8	-75.2942	19.6658	23.0943
	30	-47.6772	15.5628	16.6269
10	8	-93.1634	25.4682	34.021
	30	-61.9104	22.0421	25.3657

8. Conclusions

The proposed estimation procedures – quantile regression and REML in backfitting (QRB) and bootstrapped quantile regression and REML in backfitting (BootQRB) – provide better predictive ability and reasonable parameter estimates compared to ordinary quantile regression (OQR) in the presence of either clustered data, misspecification, or presence of outliers. In general, the prediction ability of the methods improves as the number of observations (i.e., number of clusters, or size of clusters) increases. Simulation results also suggest that prediction errors of QRB and BootQRB under misspecification in the model are smaller than those of OQR, while at the same time maintaining a relatively acceptable bias in the parameters (at 6% to 10%). Both QRB and BootQRB yield high predictive ability at different degrees, scopes or types of outliers, and both have biases within tolerable ranges (maximum at 15%). The methods are also found to perform relatively better than OQR on estimation of lower and higher quantiles of the linear relationship under possible clustering, misspecification or presence of outliers. In summary, the proposed methods may be good alternatives in the presence of misspecification and outliers.

REFERENCES

ABDULLAHI, I., and YAHAYA, A., 2015, Analysis of quantile regression as alternative to ordinary least squares. IJASP, 3, 2), 138. http://dx.doi.org/10.14419/ijasp.v3i2.4686

BATES, D., 2014, Computational methods for mixed models, Package lme4 vignette, R package version 1.1-8. [Tech-Rep] Madison, WI: Dept. of Statistics, University of Wisconsin. Retrieved from http://cran.rproject.org/web/packages/lme4/vignettes/Theory.pdf

CHERNICK, M.R. and LABUDDE, R.A., 2011, *An Introduction to Bootstrap Methods with Applications to R*, First Edition. John Wiley & Sons, Inc., Hoboken, New Jersey.

EFRON, B., and TIBSHIRANI, R., 1994, *An introduction to the bootstrap* (Chapters 5, 13, and 14). New York: Chapman & Hall.

FILZMOSER, P. and ROUSSEEUW, P.J., 2002, *Robust statistics. The Encyclopedia of Life Support Systems*. EOLSS Publishers, Oxford, UK. http://www.eolss.net

GOSPODINOV, N., KAN, R. and ROBOTTI, C., 2014, Misspecification-robust inference in linear asset-pricing models with irrelevant risk factors, *Review of Financial Studies*, 27, 2139–2170.

HAMPEL, F., 2001, Robust statistics: A brief introduction and overview. Research Report nº 94, Seminar für Statistik, Eidgenossische Technische Hochschule (ETH), Zürich. http://e-collection.library.ethz.ch/eserv/eth:24068/eth-24068-01.pdf

HAMPEL, F., RONCHETTI, E., ROUSSEEUW, P., and STAHEL, W., 2005, *Robust Statistics,* Canada: John Wiley & Sons, Inc.

HA˙RDLE, W., MU˙LLER, M., SPERLICH, S. and WERWATZ, A., 2004, *Nonparametric and Semiparametric Models*, Springer Series in Statistics. Springer, Berlin. http://www.xplore-stat.de/ebooks/ebooks.html.

HARVILLE, D.A., 1977, Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems, *Journal of the American Statistical Association*, 72:358, 320-338. http://dx.doi.org/10.1080/01621459.1977.10480998

HASTIE, T., and TIBSHIRANI, R., 1990, *Generalized Additive Models* (Chapters 4 and 6, London: Chapman and Hall.

KOENKER, R., and BASSETT, G., 1978, Regression Quantiles, *Econometrica*, 46, 1, 33-50. http://dx.doi.org/10.2307/1913643

KOENKER, R. and HALLOCK, K. , 2000, *Quantile Regression: An Introduction*. Available at (http://www.econ.uiuc.edu/-roger/research/intro/intro.html,

KOENKER, R. and HALLOCK, K., 2001, Quantile Regression. *Journal of Economic Perspectives*, 15(4, 143-156. http://dx.doi.org/10.1257/jep.15.4.143

KOLLER, M., 2013, Robust Estimation of Linear Mixed Models. PhD thesis, ETH Zurich.

LEE, S., 2014, Asymptotic refinements of a misspecification-robust bootstrap for generalized method of moments estimators, *Journal Of Econometrics*, 178, 398-413. http://dx.doi.org/10.1016/j.jeconom.2013.05.008

LIGHT, G.L., 2010, Regression, model misspecification and causation, with pedagogical demonstration, *Applied Mathematical Sciences* 4:225–236.

LO, Y., 2011, Bias from misspecification of the component variances in a normal mixture, *Computational Statistics & Data Analysis*, 55(9, 2739-2747. http://dx.doi.org/10.1016/j.csda.2011.04.007

MARONNA, R., MARTIN, R., and YOHAI, V., 2006, *Robust Statistics*, Chichester, England: J. Wiley.

WOLTERS, R. and KATEMAN, G., 1989, The performance of least squares and robust regression in the calibration of analytical methods under non-normal noise distributions. J. Chemometrics, 3, 2), 329-342. http://dx.doi.org/10.1002/cem.1180030203