

An M-quantile Random Effects Model for hierarchical and repeated measures data

Nikos Tzavidis¹

University of Manchester, e-mail: nikos.tzavidis@manchester.ac.uk

Nicola Salvati

University of Pisa, e-mail: salvati@ec.unipi.it

Abstract: Quantile analysis of hierarchical and repeated measures data has recently attracted some interest. The most recent attempt for extending the quantile regression model into a random effects quantile model is described in Geraci and Bottai (2007). In this paper we propose an alternative approach for accounting for the hierarchical structure of data when modelling the quantiles of $f(y|\mathbf{X})$. In particular, we extend the M-quantile model into an M-quantile random effects model. The proposed model allows for outlier robust estimation of both fixed and random effects. In addition, by modelling M-quantiles, instead of the ordinary quantiles, we gain algorithmic stability.

Keywords: multilevel model; robust estimation; influence function; longitudinal data

1. Introduction

The idea of modelling the quantiles of the conditional distribution of an outcome variable y given a set of covariates $\mathbf{X}$ has a long history in statistics. The seminal paper by Koenker and Bassett (1978) is usually regarded as the first detailed development of this idea. Quantile regression can be viewed as a generalization of median regression. In the same way, expectile regression (Newey and Powell, 1987) is a quantile-like generalization of mean regression. M-quantile regression (Breckling and Chambers, 1988; Chambers and Tzavidis, 2006) integrates these two concepts within a common framework defined by a quantile-like generalization of regression based on influence functions (M-regression). Quantile analysis of hierarchical and repeated measures data has recently attracted some interest. The most recent attempt for extending the quantile regression model into a random effects quantile model is described in Geraci and Bottai (2007). These authors use the Asymmetric Laplace Distribution for estimating fixed effects. The distribution of the random effects, however, is assumed to be Gaussian. In this paper we extend the M-quantile model into an M-quantile random effects model. The proposed model allows for outlier robust estimation of both fixed and random effects. In addition, by modelling M-quantiles, instead of the ordinary quantiles, we gain algorithmic stability. Robust maximum likelihood estimation of fixed and random effects is discussed in detail. The proposed model is evaluated in a series of simulation studies and is contrasted with alternative models.

¹Address of correspondence: Nikos Tzavidis, Social Statistics and Centre for Census and Survey Research, University of Manchester, Manchester, M13 9PL, UK

2. Outlier robust estimation for random effects models

Let us assume that we have data on an outcome variable y and a set of covariates $\mathbf{X}$ for individuals i clustered within groups j . The most popular approach for modelling hierarchically structured data is to use random effects models. In the general case,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon} \quad (1)$$

where $\boldsymbol{\gamma} \sim N(0, \boldsymbol{\Sigma}_u)$, $\boldsymbol{\epsilon} \sim N(0, \boldsymbol{\Sigma}_\epsilon)$ and $\mathbf{Z}$ is a design matrix that defines the random part of (1). Estimates of (1) are obtained by maximum likelihood estimation, assuming that the error components follow a Gaussian distribution. The likelihood function of (1) is a function of the squared residuals. In practice, however, data may contain outliers that result in non Gaussian distributions. If Gaussianity is still assumed, the estimated model parameters will be inefficient. One approach for robustifying the mixed model is to use an alternative loss function that grows at slower rate than the squared loss function. This is the approach followed in work that discusses outlier robust estimation for random effects models (Huggins, 1993; Richardson and Welsh, 1995). Robust maximum likelihood estimation for (1) is performed by maximizing the following log-likelihood function

$$l(\boldsymbol{\beta}_\psi, \sigma_\gamma^2, \sigma_\epsilon^2) = -\frac{K_1}{2} \log|\mathbf{V}| - \rho\{\mathbf{V}^{-1/2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_\psi)\}, \quad (2)$$

where ρ is a loss function, ψ is its derivate, $K_1 = E[\epsilon\psi(\epsilon)^T]$ computed over the standard normal distribution and $\mathbf{V}$ denotes the covariance matrix of $\mathbf{y}$ under (1). For the purposes of this work we assume a Huber-type loss function.

3. An M-quantile random effects model

Traditionally, with regression models we model the expectation of the conditional distribution $f(y|\mathbf{X})$. An approach to outlier robust regression analysis is M-regression that is based on the use influence functions. M-regression controls the effect of outliers by treating a residual whose magnitude is greater than a given cutoff, c , as if its magnitude equals c . With M-regression we model the median of the conditional distribution $f(y|\mathbf{X})$. A more complete picture is offered by modeling not only measures of central tendency of $f(y|\mathbf{X})$ but also other quantiles, q . The M-quantile model (Breckling and Chambers, 1988) for the q th quantile of $f(y|\mathbf{X})$ is

$$Q_q(y|\mathbf{X}) = \mathbf{X}\boldsymbol{\beta}_\psi(q). \quad (3)$$

In this paper we propose extending (3) into an M-quantile random effects model for taking into account the hierarchical structure of the data. The M-quantile random effects model (MQLMM) is defined as follows

$$Q_q(y|\mathbf{X}, \boldsymbol{\gamma}_\psi(q)) = \mathbf{X}\boldsymbol{\beta}_\psi(q) + \mathbf{Z}\boldsymbol{\gamma}_\psi(q). \quad (4)$$

Maximum Likelihood Estimation for the model parameters of (4) is performed by maximising

$$l(\boldsymbol{\beta}_\psi(q), \sigma_{\gamma,q}^2, \sigma_{\epsilon,q}^2) = -\frac{K_1}{2} \log|\mathbf{V}_q| - \rho_q\{\mathbf{V}_q^{-1/2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_\psi(q))\}. \quad (5)$$

It is easy to see that the log-likelihood function (5) is an extension of the log-likelihood function (2). The difference now is that the outlier robust random effects model is extended to model other than the median quantiles. When $q = 0.5$ positive and negative residuals are equally weighted. In the case of (4) this is achieved by extending the idea of asymmetric weighting of the residuals used in fitting (3), to asymmetric weighting for both the level 1 and level 2 residuals. The estimation algorithm is as follows:

- 1 Start by assuming that $(\sigma_{\gamma,q}^2, \sigma_{\epsilon,q}^2)$ are known.
- 2 Given these variance components, form the covariance matrix V_q , and estimate $\beta_\psi(q)$ using Iterative Weighted Least Squares

$$\hat{\beta}_\psi(q) = \{(\mathbf{V}_q^{-1/2} \mathbf{X})^T \mathbf{W} \mathbf{V}_q^{-1/2} \mathbf{X}\}^{-1} (\mathbf{V}_q^{-1/2} \mathbf{X})^T \mathbf{W} \mathbf{V}_q^{-1/2} \mathbf{y}.$$

where $\mathbf{W}$ denotes a square matrix of weights that depend on the quantile we are modeling and the influence function.

- 3 After the IWLS algorithm has converged, use the estimates of $\beta_\psi(q)$ to obtain estimates of the variance components by numerically maximising (5).
- 4 Iterate steps 2 and 3 until convergence.

Estimates of the variances of the model parameters are obtained by a Taylor series linearization of the score functions. Denoting by $H_q(\hat{\beta}_\psi(q), \hat{\sigma}_{\gamma,q}^2, \hat{\sigma}_{\epsilon,q}^2)$ the vector of score functions,

$$\begin{aligned} Var_q(\hat{\beta}_\psi(q), \hat{\sigma}_{\gamma,q}^2, \hat{\sigma}_{\epsilon,q}^2) &\approx \left\{ E_{0q}(\partial \beta_{\psi(q), \sigma_{\gamma,q}^2, \sigma_{\epsilon,q}^2} \mathbf{H}_{0q}) \right\}^{-1} \\ Var_{0q} \left\{ \mathbf{H}_q(\beta_{\psi(q), \sigma_{\gamma,q}^2, \sigma_{\epsilon,q}^2}) \right\} &\left[\left\{ E_{0q}(\partial \beta_{\psi(q), \sigma_{\gamma,q}^2, \sigma_{\epsilon,q}^2} \mathbf{H}_{0q}) \right\}^{-1} \right]^T. \end{aligned}$$

4. A Simulation Study

For evaluating the MQLMM model we designed the following simulation study. Data are generated under the following random intercepts model

$$y_{ij} = 100 + 2x_{ij} + \gamma_j + \epsilon_{ij}, \quad i = 1, \dots, 6, \quad j = 1, \dots, 100$$

where $x_{ij} \sim U(0, 40)$ generated in each simulation, and γ_j, ϵ_{ij} are generated under four scenarios for the contamination mechanism

$$\epsilon_{ij} = (1 - \delta)N(0, 6) + \delta N(0, 33), \quad \delta = 0 \text{ or } \delta = 0.1,$$

$$\gamma_j = (1 - \delta)N(0, 4) + \delta N(0, 22), \quad \delta = 0 \text{ or } \delta = 0.1.$$

We run 250 Monte Carlo simulations in total for each scenario and we compare the following models: (a) the proposed M-quantile random effects model (MQLMM $c = 1.345$), (b) the M-quantile single level model (MQ $c = 1.345$), (c) the quantile single level model (QR) and (d) the quantile random effects model (QRRE). All models are fitted at $q = 0.25$ and $q = 0.5$. Increasing the number of simulations will reduce the Monte-Carlo error but will not change the direction of our conclusions. Estimates of the model parameters θ_q are compared using the following measure of efficiency

$$\widehat{eff}_{model}(\hat{\theta}_q) = \frac{S_{model}^2(\hat{\theta}_q)}{S_{MQLMM}^2(\hat{\theta}_q)}, \quad S^2(\hat{\theta}_q) = \frac{1}{250} \sum_{r=1}^{250} \{\hat{\theta}_{q,r} - \bar{\theta}_q\}^2, \quad \bar{\theta}_q = \frac{1}{250} \sum_{r=1}^{250} \hat{\theta}_{q,r}.$$

In the economy of space, we present results only for three contamination mechanisms.

Table 1: *Results from the simulation study*

Scenario	Model	$q = 0.25$				$q = 0.5$			
		β_0	$\widehat{eff}$	β_1	$\widehat{eff}$	β_0	$\widehat{eff}$	β_1	$\widehat{eff}$
uncontaminated data	MQLMM	94.61	1.00	2.00	1.00	100.04	1.00	2.00	1.00
	MQ	96.72	0.91	2.00	1.35	100.07	1.14	2.00	1.31
	QR	95.22	1.30	2.00	1.96	100.06	1.55	2.00	1.78
	QRRE	94.91	1.44	2.00	1.61	100.20	1.72	2.00	1.56
contamination in level 1	MQLMM	93.55	1.00	2.00	1.00	100.08	1.00	2.00	1.00
	MQ	96.18	1.05	2.00	1.18	100.08	1.12	2.00	1.18
	QR	94.74	1.32	2.00	1.53	100.09	1.47	2.00	1.53
	QRRE	94.38	1.46	2.00	1.25	100.18	1.50	2.00	1.37
contamination in level 1 and in level 2	MQLMM	93.01	1.00	2.00	1.00	100.12	1.00	2.00	1.00
	MQ	95.66	1.06	2.00	1.44	100.12	1.00	2.00	1.31
	QR	94.33	1.11	1.99	1.78	100.12	1.19	2.00	1.78
	QRRE	93.82	1.51	1.99	1.47	100.22	1.21	2.00	1.46

5. Discussion

The results indicate that the MQLMM model is more efficient than the single level MQ model. This demonstrates that the proposed model can be used for modelling hierarchical data. It also appears that the MQLMM model is more efficient than the QRRE model. However, the MQLMM and QRRE models are not directly comparable. With the MQLMM we model M-quantiles instead of ordinary quantiles. The MQLMM model is, nevertheless, a hybrid model. By changing the tuning constant c one can bring the M-quantile model either closer to the quantile model ($c \rightarrow 0$) or closer to a mean model ($c \rightarrow \infty$). Results, not reported here, show that the MQLMM model with $c = 0.5$ retains its higher efficiency over the quantile random effects model. The MQLMM model has the additional advantage of allowing for robust inference for random and fixed effects. Further results from an application to medical data are available upon request.

References

- 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.
- Geraci M. and Bottai M. (2007) Quantile regression for longitudinal data using the asymmetric laplace distribution, *Biostatistics*, 8, 140–154.
- Huggins R. (1993) A robust approach to the analysis of repeated measures, *Biometrics*, 49, 715–720.
- Koenker R. and Bassett G. (1978) Regression quantiles, *Econometrica*, 46, 33–50.
- Newey W. and Powell J. (1987) Asymmetric least squares estimation and testing, *Econometrica*, 55, 819–847.
- Richardson A. and Welsh A. (1995) Robust estimation in the mixed linear model, *Biometrics*, 51, 1429–1439.