M-quantile regression for multivariate longitudinal data

Regressione M-quantile per dati longitudinali multivariati

M. Alfó and M.F. Marino and M.G. Ranalli and N. Salvati

Abstract We propose an M-quantile regression model for the analysis of multivariate continuous longitudinal data. M-quantile regression represents an appealing alternative to standard regression models, as it combines the robustness of quantile and the efficiency of expectile regression, detailing a picture of the response variable distribution. Discrete individual-specific random parameters are considered to account for both dependence within longitudinal profiles and association between multiple responses from the same sample unit. An extended version of the standard EM algorithm for mixed models is proposed to derive model parameter estimates.

Abstract *Si propone un modello di regressione M-quantile per l'analisi di dati (continui) multivariati misurati ripetutamente nel corso del tempo. La regressione M-quantile rappresenta una valida alternativa ai tradizionali modelli di regressione, permettendo di combinare la robustezza della regressione quantile con l'efficienza della regressione expectile. Il risultato è una visione complessiva e dettagliata della distribuzione della variabile risposta. Il modello proposto è caratterizzato dalla presenza di effetti casuali discreti che permettono di modellare congiuntamente sia la dipendenza "entro" i profili longitudinali individuali sia quella esistente tra le risposte rilevate ad ogni istante temporale su una medesima unità. Si propone inoltre l'impiego di un algoritmo EM, opportunamente esteso, per derivare le stime dei parametri del modello.*

Marco Alfó
Dipartimento di Scienze Statistiche, Sapienza Universitá di Roma e-mail: marco.alfo@uniroma1.it

Maria Francesca Marino
Dipartimento di Economia, Universitá di Perugia, e-mail: mariafrancesca.marino@unipg.it

M. Giovanna Ranalli
Dipartimento di Scienze Politiche, Universitá of Perugia e-mail: giovanna.ranalli@unipg.it

Nicola Salvati
Dipartimento di Economia e Management, Universitá di Pisa e-mail: nicola.salvati@unipi.it

Key words: Robust regression, nonparametric maximum likelihood, random effects, multivariate responses, influence function

1 Introduction

Longitudinal data analysis represents an interesting field of research as it allows to obtain deep information about the evolution of phenomena over time. However, when dealing with repeated measurements, dependence between observations coming from the same individual must be taken into consideration to avoid misleading inferential conclusions. M-quantile regression, introduced by [5], extends the ideas of standard M-estimation, see e.g. [7], allowing to characterise the conditional response distribution in terms of different location parameters (the M-quantiles). It can be considered as a quantile-like generalisation of the standard mean regression offering both robustness and efficiency. Individual-specific random parameters are considered in the model to capture dependence within individual profiles and between multiple responses recorded, at each occasion, from the same individuals. For this purpose, a finite mixture approach is adopted [1, 2], and parameter estimates are derived by means of an EM algorithm, [6], based on the assumption of (conditional) Asymmetric Least Informative distributed responses [4].

2 Multivariate M-quantile regression

Let Y_{ith} denote the h-th, $h = 1, \ldots, H$, continuous longitudinal response recorded on $i = 1, \ldots, n$ individuals at times $t = 1, \ldots, T_i$. Also, Let $\mathbf{x}_{ith}$ denote the corresponding p-dimensional vector of covariates. We are interested in analysing how covariates influence the distribution of the observed outcomes. As it is frequent in the longitudinal data literature, the association between observations from the same individual is described via individual-specific random parameters that capture potential sources of unobserved heterogeneity between individuals under observation.

Extending the proposal by [3], the M-quantile of order q for the (conditional) density of Y_{ijh} is modelled according to

$$MQ_q(y_{ith} \mid \mathbf{x}_{ith}, \mathbf{b}_{ih,q}; \psi) = \mathbf{x}'_{ith}\beta_{h,q} + \mathbf{w}'_{it}\mathbf{b}_{ih,q}, \tag{1}$$

where $\beta_{h,q}$ is a p-dimensional vector including fixed effects of the observed covariates on the response distribution. On the other hand, $\mathbf{b}_{ih,q}$ is a (profile-specific) r-dimensional random parameter vector with joint density function $f_{b,q}(\cdot)$.

As it is typically done in the (non Gaussian) mixed model framework, conditional on the random parameters $\mathbf{b}_{ih,q}$, observations from the same individual are assumed to be independent (local independence assumption). The joint conditional distribution for a generic individual $i = 1, \ldots, n$ is obtained as

$$f_q(\mathbf{y}_i \mid \beta_{h,q}, \mathbf{b}_{ih,q}, \sigma) = \prod_{h=1}^{H} \prod_{t=1}^{T_i} f_q\left(y_{ith} \mid \beta_{h,q}, \mathbf{b}_{ih,q}, \sigma\right).$$

3 ML estimation

To derive inference in a ML perspective, we adopt the approach suggested by [3] based on the Asymmetric Least Informative distribution (ALID) assumption for the conditional distribution of the responses [4]. For a given $q \in (0,1)$, the following observed individual likelihood can be derived

$$L_i(\cdot) = \frac{1}{B_q(\sigma)} \int \left\{ \prod_{h=1}^{H} \prod_{t=1}^{T_i} \exp\left\{-\rho_q\left[y_{ith} - MQ_q(y_{ith} \mid \mathbf{x}_{ith}, \mathbf{b}_{ih,q}; \psi)\right]\right\} \right\} \times \\ \times f_{b,q}(\mathbf{b}_{i1,q}, \ldots, \mathbf{b}_{iH,q}) \mathrm{d}\mathbf{b}_{i1,q} \ldots \mathrm{d}\mathbf{b}_{iH,q}, \tag{2}$$

where $\rho_q(\cdot)$ is the Huber loss function and $B_q(\sigma)$ is a normalising constant that ensures the density f_q integrates to one. To improve model flexibility and avoid unverifiable parametric assumptions on the random parameter distribution $f_{b,q}(\cdot)$, we adopt a finite mixture approach based on a discrete distribution defined over the (outcome-specific) support set $\{\zeta_{1h,q}, \ldots, \zeta_{Kh,q}\}$ with masses $\pi_{k,q} \geq 0, \forall k = 1, \ldots, K$, subject to $\sum_k \pi_{k,q} = 1$. That is, outcome-specific locations $\{\zeta_{1h,q}, \ldots, \zeta_{Kh,q}\}$ are linked through the common structure of the prior distribution. While this approach is quite straightforward and easy to implement, it implies that sources of dependence within and across outcomes are not separated; to be more specific, we have both $\mathbf{b}_{ih,q} \sim \sum_k \pi_k \delta(\zeta_{kh,q})$ and $\mathbf{b}_{ih,q} \sim \sum_k \pi_k \delta(\zeta_{k1,q}, \ldots, \zeta_{kH,q})$, where $\delta(\theta)$ puts unit mass on θ. This may be linked to the nonparametric maximum likelihood (NPML) estimate of the mixing distribution $f_{b,q}(\cdot)$, see among others [2]. According to (3) and conditional on the k-th component of the finite mixture, the M-quantile regression model of order q can be written as follows:

$$MQ_q(y_{ith} \mid \mathbf{x}_{it}, \mathbf{b}_{kh,q}; \psi) = \mathbf{x}'_{ith}\beta_{h,q} + \mathbf{w}'_{ith}\zeta_{kh,q}. \tag{3}$$

and the individual likelihood function (2) becomes

$$L_i(\cdot) = \frac{1}{B_q(\sigma)} \sum_{k=1}^{K} \prod_{h=1}^{H} \prod_{t=1}^{T_i} \exp\left\{-\rho_q\left[y_{ith} - MQ_q(y_{ith} \mid \mathbf{x}_{ith}, \zeta_{ih,q}; \psi)\right]\right\} \pi_{k,q}. \tag{4}$$

The full data likelihood is given by the product of the individual likelihood functions (4) and model parameter estimates can be derived by directly maximizing the resulting equation. Here, we propose an indirect approach, based on a EM algorithm [6]. Let $\Phi_q = \left\{\beta_{1,q}, \ldots, \beta_{H,q}, \zeta_{11,q}, \ldots, \zeta_{K1,q}, \ldots, \zeta_{1H,q}, \ldots, \zeta_{KH,q}, \sigma, \pi_{1,q}, \ldots, \pi_{K,q}\right\}$ represent the “global” set of model parameters for the q-th quantile, $q \in (0,1)$. The log-likelihood function for complete data is defined as

$$\ell_c(\Phi_q)\sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik,q}\sum_{h=1}^{H}\sum_{t=1}^{T_i}\left\{\log\left[f_q(y_{ith}\mid\beta_{h,q},\zeta_{kh,q},\sigma)\right]+\log(\pi_{k,q})\right\} \tag{5}$$

where $z_{ik,q}$, $i=1,\ldots,n$, $k=1,\ldots,K$, denotes the indicator variable for the i-th individual in the k-th component of the mixture when the q-th quantile is considered. Parameter estimates are derived by alternating two separate steps. In the E-step, the unobservable indicators are replaced by the corresponding posterior expectations. In the M-step, the conditional expectation of the complete data log-likelihood given the observed responses $\mathbf{y}_i$ and the current parameter estimates is maximised with respect to model parameters. The E- and the M- steps are repeatedly alternated until the difference between two subsequent likelihoods is lower than a fixed constant ε. For a given level $q\in[0,1]$, the algorithm is run with a fixed number of mixture components; once it reaches convergence, K is increased to $K+1$ and the algorithm is run again. To avoid being trapped in local maxima, we may consider multiple starting points. A formal comparison between models corresponding to different choices for K is performed using penalized likelihood criteria.

References

[1] M. Aitkin. A general maximum likelihood analysis of overdispersion in generalized linear models. *Statistics and Computing*, 6(3):251–262, 1996.

[2] M. Aitkin. A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics*, 55(1):117–128, 1999.

[3] M. Alfó, M.G. Ranalli, and N. Salvati. Finite mixtures of quantiles and M-quantile models. *Statistics and Computing*, page in press, 2016.

[4] A. Bianchi and N. Salvati. Asymptotic properties and variance estimators of the M-quantile regression coefficients estimators. *Communications in Statistics - Theory and Methods*, 44:2416–2429, 2015.

[5] J. Breckling and R. Chambers. *M*-quantiles. *Biometrika*, 75:761–771, 1988.

[6] A.P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B*, 39(1):1–38, 1977.

[7] P.J. Huber. Robust estimation of a location parameter. *Ann. Math. Stat.*, 35:73–101, 1964.