Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Multivariate Analysis 91 (2004) 74–89
Journal of Multivariate Analysis
www.elsevier.com/locate/jmva

Quantile regression for longitudinal data

Roger Koenker*

Department of Economics, University of Illinois at Urbana-Champaign, Box 111-1206 South-Sixth St., Champaign, IL 61820, USA

Received 11 November 2003

Abstract

The penalized least squares interpretation of the classical random effects estimator suggests a possible way forward for quantile regression models with a large number of "fixed effects". The introduction of a large number of individual fixed effects can significantly inflate the variability of estimates of other covariate effects. Regularization, or shrinkage of these individual effects toward a common value can help to modify this inflation effect. A general approach to estimating quantile regression models for longitudinal data is proposed employing ℓ_1 regularization methods. Sparse linear algebra and interior point methods for solving large linear programs are essential computational tools.

AMS 1991 subject classifications: 62J05; 62J07; 62G35

Keywords: Quantile regression; Penalty methods; Shrinkage; L-statistics; Random effects; Robust estimation; Hierarchical models

1. Introduction

Recent contributions to the literature on linear and nonlinear mixed models have emphasized the strong link with penalty methods for nonparametric function estimation. Shrinkage of highly overparameterized models toward simpler, plausible models suggested by prior smoothness considerations shares many common features with the shrinkage of nominal effects toward common values based on prior beliefs about their exchangeability. The dominant paradigm in the random effects, mixed model literature has been a Gaussian structure in which covariates exert a pure location shift effect on the response variable. In some applica-

* Fax: +001-217-244-6678.
E-mail address: roger@ysidro.econ.uiuc.edu (R. Koenker).

0047-259X/$ - see front matter © 2004 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmva.2004.05.006

tions it is of interest to explore a broader class of covariate effects, while still accounting for individual specific effects. Such models enable the investigator to explore various forms of heterogeneity associated with the covariates under less stringent distributional assumptions.

The almost exclusive focus on least squares estimators under Gaussian conditions for longitudinal data analysis can be taken as a challenge: Can a more flexible, more robust approach to longitudinal data analysis be forged outside the Gaussian random effects framework? I will argue that quantile regression might play a constructive role in such a development.

The construction of infant and adolescent growth charts provides a motivating application in which a functional component as well as ordinal and nominal factors may appear. It is of obvious importance to construct reference growth charts that accurately represent the conditional quantiles of the growth distribution without unduly constraining the estimation process by unverifiable distributional assumptions. Several authors including Cox and Jones in the discussion of Cole [2] have suggested that quantile regression methods may offer advantages over parametric approaches to the analysis of such growth charts. A challenge in these applications is to respect the longitudinal structure of most growth data allowing individual specific effects while allowing covariates to play a more flexible role.

The quantile regression problems that will be considered generally involve a large number of cross-sectional "individuals" observed over a relatively short number of time periods. Typical reference growth charts are based on several hundred individuals with about 10–20 measurements per individual. When each cross-sectional observation is allowed an individual specific location shift effect the parametric dimension of the resulting estimation problem can be quite large. Computational methods that exploit the inherently sparse nature of the linear algebra for interior point solution of the resulting linear programming problems play an essential role.

2. Models and methods

Consider the classical linear random effects model,

$$y_{ij} = x_{ij}^{\top}\beta + \alpha_i + u_{ij} \quad j = 1, \ldots, m_i, \quad i = 1, \ldots, n, \tag{2.1}$$

which we will write in matrix form as,

$$y = X\beta + Z\alpha + u.$$

The matrix Z represents an incidence matrix that identifies the n distinct individuals in the sample. In the growth curve setting the subscript i would index individual patients, and the subscript j would index the m_i distinct measurements made on the ith patient. We begin by recalling an instructive characterization of the random effects estimator under Gaussian conditions.

2.1. Gaussian random effects as penalized least squares

Suppose u and α are independent Gaussian vectors with $u \sim \mathcal{N}(0, R)$ and $\alpha \sim \mathcal{N}(0, Q)$. Observing that $v = Z\alpha + u$ has covariance matrix

$$Evv^{\top} = R + ZQZ^{\top}$$

we can immediately deduce that the minimum variance unbiased estimator of β is,

$$\hat{\beta} = (X^\top (R + ZQZ^\top)^{-1} X)^{-1} X^\top (R + ZQZ^\top)^{-1} y.$$

This estimator is certainly not very appealing from a robustness standpoint, but the optimization problem that gives rise to $\hat{\beta}$ is suggestive of a larger class of possible candidate estimators under non-Gaussian conditions.

Proposition. *$\hat{\beta}$ solves* $\min_{(\alpha,\beta)} \|y - X\beta - Z\alpha\|^2_{R^{-1}} + \|\alpha\|^2_{Q^{-1}}$, *where* $\|x\|^2_A = x^\top Ax$.

Proof. Differentiating we obtain the normal equations,

$$X^\top R^{-1} X\hat{\beta} + X^\top R^{-1} Z\hat{\alpha} = X^\top R^{-1} y$$

$$Z^\top R^{-1} X\hat{\beta} + (Z^\top R^{-1} Z + Q^{-1})\hat{\alpha} = Z^\top R^{-1} y.$$

Solving, we have $\hat{\beta} = (X^\top \Omega^{-1} X)^{-1} X^\top \Omega^{-1} y$ where

$$\Omega^{-1} = R^{-1} - R^{-1} Z (Z^\top R^{-1} Z + Q^{-1})^{-1} Z^\top R^{-1}.$$

But $\Omega = R + ZQZ^\top$, see e.g. Rao [15, p. 33]. □

This result has a long history. Robinson [16] attributes the normal equations above to Henderson [5]. Goldberger [4] introduced the terminology "best linear unbiased predictor", subsequently rendered as BLUP, to describe the estimator $\hat{\beta}$ and its associated "estimator" $\hat{\alpha}$ of the random effects. The implicit estimation of the random effects may appear strange, but viewing the random effects estimator as a penalized least squares estimator opens the door to the consideration of alternative penalties. By shrinking the unconstrained $\hat{\alpha}$'s toward a common value we achieve not only improved performance of the individual fixed-effect estimates, but also improve the performance of the estimate of β. In the Bayesian paradigm the penalty formulation is natural, as emphasized by Lindley and Smith [13], and many subsequent authors. Alternatives to the Gaussian penalty $\|\alpha\|^2_{Q^{-1}}$, such as those proposed below would simply reflect differences in prior beliefs about the distribution of the α's.

2.2. Quantile regression with fixed effects

Contemplating the extension of the model (2.1) to models for conditional quantile functions we must first confront the question: What role should the α's play? Generally, the α's would be intended to capture some individual specific source of variability, or "unobserved heterogeneity," that was not adequately controlled for by other covariates in the model. For example, in a study of the effect of a dietary intervention on blood pressure, it would be desirable to estimate departures from individuals' idiosyncratic levels. If the number of observations m_i were large for each individual then we might even hope to estimate a *distributional* shift $\alpha_i(\tau)$ for each individual. This would certainly be useful for *groups* of individuals: a distributional shift for men versus women, or for blacks versus whites. However, in most applications the m_i, the number of observations on each individual, would be relatively modest and then it is quite unrealistic to attempt to estimate a τ-dependent, *distributional*, individual effect. At best we may be able to estimate an individual specific location-shift effect, and even this may strain credulity.

We will consider the following model for the conditional quantile functions of the response of the jth observation on the ith individual y_{ij},

$$Q_{y_{ij}}(\tau|x_{ij}) = \alpha_i + x_{ij}^\top \beta(\tau) \quad j = 1, \ldots m_i, \quad i = 1, \ldots, n. \tag{2.2}$$

In this formulation the α's have a pure location shift effect on the conditional quantiles of the response. The effects of the covariates, x_{ij} are permitted to depend upon the quantile, τ, of interest, but the α's do not.

To estimate the model (2.2) for several quantiles simultaneously, we propose solving,

$$\min_{(\alpha,\beta)} \sum_{k=1}^{q} \sum_{j=1}^{n} \sum_{i=1}^{m_i} w_k \rho_{\tau_k}(y_{ij} - \alpha_i - x_{ij}^\top \beta(\tau_k)) \tag{2.3}$$

where $\rho_\tau(u) = u(\tau - I(u < 0))$, denotes the piecewise linear quantile loss function of Koenker and Bassett [11]. The weights w_k control the relative influence of the q quantiles $\{\tau_1, \ldots, \tau_q\}$, on the estimation of the α_i parameters. The choice of the weights, w_k, and the associated quantiles τ_k, is somewhat analogous to the choice of discretely weighted L-statistics, as for example in Mosteller [14]. In the Monte-Carlo section below we use Tukey's trimean as a prototype assigning weights, 0.25, 0.5, and 0.25 to the quartiles. Koenker [10] considered an analogous situation in which only the intercept parameter was permitted to depend upon τ and the slope parameters associated with the included covariates were constrained to be identical for several τ's. In this case the slope parameters are estimated were estimated as regression L-statistics. In the present instance, it is the α parameters that are estimated as discretely weighted L-statistics.

Solving the problem (2.3) may appear somewhat quixotic when the dimensions n, m and q are large. In least squares applications the usual strategy would be to transform y and X to deviations from individual means, and then compute $\hat{\beta}$ from the transformed data. For quantile regression this decomposition of projections is not available and we are required to deal directly with the full problem. Fortunately, in typical applications the problem is quite sparse, that is the design matrix of the full problem is mostly zeros. Storing the dense version of this matrix with all the zeros treated as double precision floats may well be infeasible, but standard sparse matrix storage schemes only require space for the non-zero elements and their indexing locations. This dramatically reduces the memory requirements in large problems.

Interior point methods for solving (2.3) proceed iteratively by solving a sequence of diagonally weighted least squares steps using a Cholesky factorization. The sparsity of the design is typically preserved quite well in this factorization, as noted by Saunders [18], and the computational effort is roughly proportional to the number of non-zero elements. Implementations of this approach for the public domain dialect R, Ihaka and Gentleman [6], of Chambers [1] S language are discussed in Koenker and Ng [12] and are available on CRAN at www.r-project.org.

2.3. Penalized quantile regression with fixed effects

We have seen that the optimal estimator for the Gaussian prototype model (2.1) involves shrinking the $\hat{\alpha}$'s toward a common value. When the x_{ij} component of the model contains

an intercept, as we will henceforth assume, this common value can be taken to be the conditional central tendency of the response at a point determined by the centering of the other covariates. In the quantile regression version of the model (2.2) this would be some corresponding conditional quantile of the response, although we would require further conditions including symmetry of the τ_k's and the w_k's to be specified.

Particularly when n is large relative to the m_i's shrinkage may be advantageous in controlling the variability introduced by the large number of estimated α parameters. For the quantile loss function, ρ_τ it is convenient to consider the ℓ_1 penalty,

$$P(\alpha) = \sum_{i=1}^{n} |\alpha_i|$$

in place of the conventional Gaussian penalty. This choice maintains the linear programming form of the problem and also preserves the sparsity of the resulting design matrix. Several authors, notably Tibshirani [20] and Donoho et al. [3], have pointed out that ℓ_1 shrinkage offers some statistical advantages over more conventional Gaussian ℓ_2 penalties in addition to its computational advantages.

We will consider estimators solving the penalized version of (2.3)

$$\min_{(\alpha,\beta)} \sum_{k=1}^{q} \sum_{j=1}^{n} \sum_{i=1}^{m_i} w_k \rho_{\tau_k}(y_{ij} - \alpha_i - x_{ij}^\top \beta(\tau_k)) + \lambda \sum_{i=1}^{n} |\alpha_i|. \tag{2.4}$$

For $\lambda \to 0$ we obtain the fixed effects estimator described above, while as $\lambda \to \infty$ the $\hat{\alpha}_i \to 0$ for all $i = 1, 2, \ldots, n$ and we obtain an estimate of the model purged of the fixed effects. Note that since the x_{ij} component is assumed to contain an intercept, in either case we will also have q, τ-specific, estimates of the intercept. If we consider the special case that $m_i \equiv m$ for all i, we can write the design matrix for a single quantile as,

$$[X \dot{:} I_n \otimes e_m]$$

where $X = (x_{ij})$ is nm by p, and e_m denotes an m-vector of ones. The design matrix for $q > 1$ may be written as,

$$[W \otimes X \dot{:} w \otimes (I_n \otimes e_m)].$$

Appending the penalty term we have the augmented design matrix,

$$\begin{bmatrix} W \otimes X & w \otimes (I_n \otimes e_m) \\ 0 & \lambda I_n \end{bmatrix}.$$

which has dimension $qnm + n$ by $qp + n$. The corresponding response vector is $\tilde{y} = ((w \otimes y)^\top 0_n^\top)^\top$. These dimensions may seem even more daunting than before, but again we should emphasize that the sparsity of the design matrix comes to the rescue. Even quite large problems of this type can be handled successfully on rather modest machines.

3. Asymptopia

The existence of the parameter α whose dimension, n, is tending to infinity raises some fundamentally new issues for the asymptotic analysis of the quantile regression estimator. To address these issues it seems prudent to begin with a relatively simple setting in which we focus on estimation of a single conditional quantile function. We will restrict attention to balanced designs with $m_i = m$ for all $i = 1, \ldots, n$. Then, since $Z = I_n \otimes e_m$, we have $Z^\top Z = mI_n$. We will begin by considering the fixed effect estimator, and then turn to the penalized estimator. Both m and n will be assumed to tend to infinity. Convergence in distribution will be denoted by the symbol $\rightsquigarrow$.

Consider the objective function

$$V_{mn}(\delta) = \sum_{j=1}^{m} \sum_{i=1}^{n} \rho_\tau(y_{ij} - \xi_{ij}(\tau) - z_{ij}^\top \delta_0/\sqrt{m} - x_{ij}^\top \delta_1/\sqrt{mn}) - \rho_\tau(y_{ij} - \xi_{ij}(\tau))]$$

where $\xi_{ij}(\tau) = \alpha_i + x_{ij}^\top \beta(\tau)$. Note that

$$\hat{\delta} = \begin{pmatrix} \hat{\delta}_0 \\ \hat{\delta}_1 \end{pmatrix} = \begin{pmatrix} \sqrt{m}(\hat{\alpha} - \alpha) \\ \sqrt{mn}(\hat{\beta}(\tau_1) - \beta(\tau_1)) \end{pmatrix}$$

minimizes the function V_{mn}. We will impose the following regularity conditions:

A1. The y_{ij} are independent with conditional distribution functions, F_{ij}, given x_{ij}, and differentiable conditional densities, $0 < f_{ij} < \infty$, with bounded derivatives f'_{ij}, at $\xi_{ij}(\tau)$, for $j = 1, \ldots, m,\ i = 1, \ldots, n$,

A2. Let $\omega = \tau(1 - \tau)$ and denote $\Phi = \mathrm{diag}(f_{ij}(\xi_{ij}(\tau)))$. The limiting forms of the following matrices are positive definite:

$$D_0 = \lim_{\substack{m\to\infty \\ n\to\infty}} \frac{\omega}{m} \begin{pmatrix} Z^\top Z & Z^\top X/\sqrt{n} \\ X^\top Z/\sqrt{n} & X^\top X/n \end{pmatrix}$$

$$D_1 = \lim_{\substack{m\to\infty \\ n\to\infty}} m^{-1} \begin{pmatrix} Z^\top \Phi Z & Z^\top \Phi X/\sqrt{n} \\ X^\top \Phi Z/\sqrt{n} & X^\top \Phi X/n \end{pmatrix}.$$

A3. $\max_{\substack{1 \leqslant i \leqslant n \\ 1 \leqslant j \leqslant m}} ||x_{ij}|| < M$.

The condition A1 is now quite familiar in the quantile regression literature. Condition A2 is not, but if one supposes for a moment that the model is of the pure location shift form (2.1), then D_1 simplifies somewhat and A2 reduces to a condition on the matrices $X^\top X/(mn)$ and $Z^\top Z/m$. We have seen that the latter is equal to I_n, and the former condition is again familiar. If $Z^\top X = 0$ so that there is no "between" variability in the x's, then the expressions simplify considerably, but this case is quite atypical, and generally we would expect that there would be some non-orthogonality between the individual effects and the other covariates and thus some potential improvement in the estimation of β's due to control of the α's by shrinkage toward a common value. These expectations are confirmed in the next section through a

small simulation experiment. Condition A3 could be relaxed at the cost of some added complication of the argument.

Theorem 1. *Under conditions* A1–A3, *with* $n^a/m \to 0$ *for some* $a > 0$, *the* $\hat{\delta}_1$ *component of the minimizer,* $\hat{\delta}$, *converges in distribution to a Gaussian random vector with mean zero and covariance matrix given by the lower p by p block of the matrix* $D_1^{-1} D_0 D_1^{-1}$.

Proof. Two distinct arguments will be given. The first should be regarded as purely heuristic, since it overlooks the complications introduced by the infinite dimensional nature of α. (An alternative view of the first argument is that it applies to situations in which m tends to infinity, and n is fixed.) The second explicitly concentrates out the α parameter focusing on the finite dimensional asymptotic behavior of $\hat{\beta}(\tau)$. The equivalence between the two results is established with the aid of a matrix identity formulated as Lemma 1.

Part 1. The function V_{mn} can be decomposed into two parts using the identity of Knight [7],

$$\rho_\tau(u - v) - \rho_\tau(u) = -v\psi_\tau + \int_0^v (I(u \leqslant s) - I(u \leqslant 0))\, ds$$

where $\psi_\tau(u) = \tau - I(u < 0)$ denotes the quantile influence function. We will write,

$$V_{mn}(\delta) = V_{mn}^{(1)}(\delta) + V_{mn}^{(2)}(\delta)$$

where $v_{ij} = z_{ij}^\top \delta_0 + x_{ij}^\top \delta_1/\sqrt{n}$ and,

$$V_{mn}^{(1)}(\delta) = -m^{-1/2} \sum_j \sum_i \psi_\tau(y_{ij} - \xi_{ij}(\tau_k)) v_{ij}$$

$$V_{mn}^{(2)}(\delta) = -m^{-1/2} \sum_j \sum_i \int_0^{v_{ij}} (I(y_{ij} \leqslant \xi_{ij}(\tau) + t/\sqrt{m}) - I(y_{ij} \leqslant \xi_{ij}(\tau)))\, dt.$$

The first term is asymptotically Gaussian. Let $\Psi_k = \mathrm{diag}(\psi_\tau(y_{ij} - \xi_{ij}(\tau)))$ and note that $E\Psi e_{mn} e_{mn}^\top \Psi = \omega I_{mn}$. Conditions A2 and A3 imply a Lindeberg condition and we have,

$$V_{mn}^{(1)}(\delta) = -m^{-1/2}(Z^\top \Psi \delta_0 + X^\top \Psi \delta_1/\sqrt{n}) \rightsquigarrow -B\delta.$$

The second term is asymptotically quadratic in δ. Note that

$$\begin{aligned}
EV_{mn}^{(2)}(\delta) &= m^{-1} \sum_j \sum_i \int_0^{v_{ij}} \sqrt{m}(F_{ij}(\xi_{ij}(\tau) + t/\sqrt{m}) - F_{ij}(\xi_{ij}(\tau)))\, dt \\
&= m^{-1} \sum_j \sum_i \int_0^{v_{ij}} f_{ij}(\xi_{ij}(\tau)) t\, dt + o(1) \\
&= \frac{1}{2m} \sum_j \sum_i f_{ij}(\xi_{ij}(\tau))(z_{ij}^\top \delta_0 + x_{ij}^\top \delta/\sqrt{n})^2 + o(1) \\
&= \frac{1}{2m}(\delta_0^\top Z^\top \Phi Z \delta_0 + 2\delta_0^\top Z^\top \Phi X \delta_1/\sqrt{n} + \delta_1^\top X^\top \Phi X \delta_1/n) + o(1) \\
&\to \frac{1}{2}\delta^\top D_1 \delta.
\end{aligned}$$

The variance of $V_{mn}^{(2)}(\delta)$ converges to zero by Condition A3. The limiting form of the objective function is thus

$$V_0(\delta) = -\delta^\top B + \frac{1}{2}\delta^\top D_1 \delta$$

where B is a zero mean Gaussian vector with covariance matrix D_0. In finite-dimensional settings, i.e. with n fixed in the present instance, convexity of the objective function, V_{mn}, and the uniqueness of the minimum of V_0, yields uniformity in δ. so $\hat{\delta}$ converges to the argmin of V_0 completing the argument as in Knight and Fu [9].

Part 2. Given the infinite dimensional nature of α there may be some legitimate doubt about the validity of the foregoing approach. A more rigorous argument can be made by explicitly replacing the $\hat{\alpha}$'s by their Bahadur representation and thereby concentrating out their effect, expressing the objective function solely in terms of the finite dimensional parameter β. Using the reparameterization of the previous argument, note that for any fixed δ_1, we can consider the behavior of $\hat{\delta}_{0i}$, which depends only on the m observations for the ith subsample. It follows from the argument of Ruppert and Carroll [17] that uniformly for $\| \delta_1 \| < \Delta_1$ and $|\delta_{0i}| < \Delta_0$,

$$\| g_i(\delta_{0i}, \delta_1) - g_i(0,0) - E(g_i(\delta_{0i}, \delta_1) - g_i(0,0)) \| = o_p(1)$$

where

$$g_i(\delta_{0i}, \delta_1) = -m^{-1/2} \sum_{j=1}^{n} \psi_\tau(y_{ij} - \xi_{ij}(\tau) - x_{ij}^\top \delta_1/\sqrt{nm} - \delta_{0i}/\sqrt{m})$$

with $\psi_\tau(u) = \tau - I(u < 0)$. Expanding we have,

$$\begin{aligned} E(g_i(\delta_{0i}, \delta_1)) &= m^{-1/2} \sum_{j=1}^{n} [F_{ij}(\xi_{ij}(\tau) + x_{ij}^\top \delta_1/\sqrt{nm} + \delta_{0i}/\sqrt{m}) - \tau] \\ &= m^{-1/2} \sum_{j=1}^{n} f_{ij}(\xi_{ij}(\tau))[x_{ij}^\top \delta_1/\sqrt{nm} + \delta_{0i}/\sqrt{m}] + o_p(1). \end{aligned}$$

Optimality of the $\hat{\delta}_{0i}$ implies that $g_i(\delta_{0i}, \delta_1) = o(m^{-1})$, letting $\bar{f}_i = m^{-1} \sum_{j=1}^{m} f_{ij}$,

$$\hat{\delta}_{0i} = \bar{f}_i^{-1} \left[m^{-1} \sum_{j=1}^{m} f_{ij}(\xi_{ij}(\tau)) x_{ij}^\top \delta_1/\sqrt{n} + m^{-1/2} \sum_{j=1}^{m} \psi_\tau(y_{ij}(\tau) - \xi_{ij}(\tau)) \right] + R_{mi}.$$

Substituting the $\hat{\delta}_{0i}$'s, we will denote,

$$G(\delta_1) = \frac{1}{\sqrt{mn}} \sum_{i=1}^{n} \sum_{j=1}^{m} x_{ij} \psi_\tau(y_{ij}(\tau) - \xi_{ij}(\tau) - x_{ij}^\top \delta_1/\sqrt{nm} + \hat{\delta}_{0i}/\sqrt{m}).$$

Again, uniformly for $\| \delta_1 \| < \Delta_1$, one can show that,

$$\| G(\delta_1) - G(0) - E(G(\delta_1) - G(0)) \| = o_p(1),$$

and at the minimizer, $G(\hat{\delta}_1) = o((mn)^{-1})$. Expanding, as above,

$$\begin{aligned} E(G(\delta_1) - G(0)) &= \frac{1}{\sqrt{mn}} \sum_{i=1}^{n} \sum_{j=1}^{m} f_{ij} x_{ij} (x_{ij}^\top \delta_1 / \sqrt{nm} + \hat{\delta}_{0i} / \sqrt{m}) \\ &= \frac{1}{mn} \sum_{i=1}^{n} \sum_{j=1}^{m} f_{ij} x_{ij} \left(x_{ij}^\top \delta_1 - \bar{f}_i^{-1} m^{-1} \sum_{j=1}^{m} f_{ij} x_{ij}^\top \delta_1 \right) \\ &\quad + \frac{1}{\sqrt{mn}} \sum_{i=1}^{n} \sum_{j=1}^{m} f_{ij} x_{ij} \bar{f}_i^{-1} m^{-1/2} \sum_{j=1}^{m} \psi_\tau(y_{ij} - \xi_{ij}(\tau)) \\ &\quad + \frac{1}{\sqrt{mn}} \sum_{i=1}^{n} \sum_{j=1}^{m} f_{ij} x_{ij} R_{mi} / \sqrt{m} + O(m^{-1/2}) \end{aligned}$$

where the order of the final term is controlled by the bound on the derivative of the conditional density. Setting the foregoing expression equal to $G(0)$ and solving for $\hat{\delta}_1$ yields, in somewhat more convenient matrix notation,

$$\hat{\delta}_1 = (X^\top M_{\tilde{Z}}^\top \Phi M_{\tilde{Z}} X)^{-1} (X^\top M_{\tilde{Z}}^\top \Psi + R_{mn})$$

where $M_{\tilde{Z}} = I - P_{\tilde{Z}}$, $P_{\tilde{Z}} = Z(Z^\top \Phi Z)^{-1} \Phi$, and Ψ denotes the mn vector $(\psi(y_{ij} - \xi_{ij}(\tau)))$, and

$$R_{mn} = \frac{1}{\sqrt{mn}} \sum_{i=1}^{n} \sum_{j=1}^{m} f_{ij} x_{ij} R_{mi} / \sqrt{m} + O(m^{-1/2}).$$

The remainder term, R_{mn}, has dominant component that comes from the Bahadur representation of the $\hat{\alpha}$'s. By A1 and A3, we have for a generic constant K,

$$R_{mn} = m^{-1/4} \frac{K}{\sqrt{n}} \sum_{i=1}^{n} R_{0i} + o_p(m^{-1/4}).$$

The analysis of Knight [8] shows that the summands converge in distribution, that is as $m \to \infty$, we have $m^{1/4} R_{mi} \rightsquigarrow R_{0i}$, where the R_{0i} are functionals of Brownian motion. Independence of the y_{ij}, and the condition on the growth of m ensures that contribution of the remainder is negligible. Denoting the limiting form of the matrices:

$$\tilde{D}_1 = \lim_{\substack{m \to \infty \\ n \to \infty}} (mn)^{-1} X^\top M_{\tilde{Z}}^\top \Phi M_{\tilde{Z}} X$$

and

$$\tilde{D}_0 = \lim_{\substack{m \to \infty \\ n \to \infty}} \frac{\omega}{mn} X^\top M_{\tilde{Z}}^\top M_{\tilde{Z}} X$$

we have, neglecting the penalty term of the objective function,

$$\tilde{\delta}_1 \rightsquigarrow \mathcal{N}(0, \tilde{D}_1^{-1}\tilde{D}_0\tilde{D}_1^{-1}).$$

The next lemma verifies that this form of the covariance matrix is identical to the lower diagonal block of the matrix $D_1^{-1}D_0D_1^{-1}$ derived previously. □

Lemma 1. $(D_1^{-1}D_0D_1^{-1})_{22} = \tilde{D}_1^{-1}\tilde{D}_0\tilde{D}_1^{-1}$.

Proof. Standard partitioned inverse formulae give,

$$\begin{aligned} mn(D_1^{-1}D_0D_1^{-1})_{22} &= \begin{pmatrix} -FE^{-1} & E^{-1} \end{pmatrix} \begin{pmatrix} Z^\top Z & Z^\top X \\ X^\top Z & X^\top X \end{pmatrix} \begin{pmatrix} -FE^{-1} \\ E^{-1} \end{pmatrix} \\ &= E^{-1}[F^\top Z^\top ZF - X^\top ZF - F^\top Z^\top X + X^\top X]E^{-1} \end{aligned}$$

where $E = X^\top \Phi X - X^\top \Phi Z(Z^\top \Phi Z)^{-1}Z^\top \Phi X = mn\tilde{D}_1$, and $F = (Z^\top \Phi Z)^{-1}Z^\top \Phi X = P_{\tilde{Z}}X$. The result then follows by noting that the term in square brackets is equal to $X^\top M_{\tilde{Z}}^\top M_{\tilde{Z}}X$. □

3.1. Asymptotics for the penalized quantile regression estimator

To explore the asymptotic behavior of the penalized quantile regression estimator solving (2.3) we will maintain the assumption of a balanced design and consider simultaneously estimating q quantiles. Let

$$\begin{aligned} V_{mn}(\delta) = \sum_{k=1}^{q}\sum_{j=1}^{m}\sum_{i=1}^{n} w_k[\rho_{\tau_k}(y_{ij} - \xi_{ij}(\tau_k) - z_{ij}^\top\delta_0/\sqrt{m} - x_{ij}^\top\delta_k/\sqrt{mn}) \\ -\rho_{\tau_k}(y_{ij} - \xi_{ij}(\tau_k))] + \lambda_m \sum_{i=1}^{n} |\alpha_i - \delta_{0i}/\sqrt{m}| - |\alpha_i| \end{aligned}$$

where $\xi_{ij}(\tau_k) = \alpha_i + x_{ij}^\top\beta(\tau_k)$. Note that

$$\hat{\delta} = \begin{pmatrix} \hat{\delta}_0 \\ \hat{\delta}_1 \\ \vdots \\ \hat{\delta}_q \end{pmatrix} = \begin{pmatrix} \sqrt{m}(\hat{\alpha} - \alpha) \\ \sqrt{mn}(\hat{\beta}(\tau_1) - \beta(\tau_1)) \\ \vdots \\ \sqrt{mn}(\hat{\beta}(\tau_q) - \beta(\tau_q)) \end{pmatrix}$$

minimizes the function V_{mn}. We will impose the modified regularity conditions:

B1. The y_{ij} are independent with conditional distribution functions, F_{ij}, given x_{ij}, and differentiable conditional densities, $0 < f_{ij} < \infty$, with bounded derivatives f'_{ij}, at $\xi_{ij}(\tau)$, for $j = 1, \ldots, m,\ i = 1, \ldots, n$,

B2. Let Ω denote the q by q matrix with typical element $\tau_k \wedge \tau_l - \tau_k\tau_l$ and $\Phi_k = \mathrm{diag}(f_{ij}(\xi_{ij}(\tau_k)))$. The limiting forms of the following matrices are positive definite:

$$D_0 = \lim_{\substack{m\to\infty \\ n\to\infty}} m^{-1} \begin{pmatrix} w^\top \Omega w Z^\top Z & w^\top \Omega W \otimes Z^\top X/\sqrt{n} \\ W\Omega w \otimes X^\top Z/\sqrt{n} & W\Omega W \otimes X^\top X/n \end{pmatrix}$$

$$D_1 = \lim_{\substack{m\to\infty \\ n\to\infty}} m^{-1} \begin{pmatrix} \sum w_k Z^\top \Phi_k Z & w_1 Z^\top \Phi_1 X/\sqrt{n} & \cdots & w_q Z^\top \Phi_q X/\sqrt{n} \\ w_1 X^\top \Phi_1 Z/\sqrt{n} & w_1 X^\top \Phi_1 X/n & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ w_q X^\top \Phi_q Z/\sqrt{n} & 0 & \cdots & w_q X^\top \Phi_q X/n \end{pmatrix}.$$

B3. $\max_{\substack{1\leqslant i\leqslant n \\ 1\leqslant j\leqslant m}} ||x_{ij}|| < M$.

Theorem 2. *Under conditions* B1–B3, *provided that* $\lambda_m/\sqrt{m} \to \lambda_0$, *and* $n^a/m \to 0$ *for some* $a > 0$, *the first component* $\hat{\delta}_1$ *minimizing* V_{mn} *has the same limiting distribution as the first component of the minimizer of*,

$$V_0(\delta) = -\delta^\top B + \frac{1}{2}\delta^\top D_1 \delta + \lambda_0 \delta^\top s$$

where B *is a zero mean Gaussian vector with covariance matrix* D_0, *and* $s = (s_0^\top 0_{pq}^\top)^\top$ *and* $s_0 = (\mathrm{sgn}(\alpha_i))$.

Proof. A sketch of the heuristic form of the argument for the previous result is provided. The function V_{mn} can be decomposed into three parts using the identity of Knight [7],

$$\rho_\tau(u - v) - \rho_\tau(u) = -v\psi_\tau + \int_0^v (I(u \leqslant s) - I(u \leqslant 0))\, ds$$

where $\psi_\tau(u) = \tau - I(u < 0)$ denotes the quantile influence function. We will write,

$$V_{mn}(\delta) = V_{mn}^{(1)}(\delta) + V_{mn}^{(2)}(\delta) + V_{mn}^{(3)}(\delta),$$

where $v_{ijk} = z_{ij}^\top \delta_0 + x_{ij}^\top \delta_k/\sqrt{n}$ and,

$$V_{mn}^{(1)}(\delta) = -m^{-1/2} \sum_k \sum_j \sum_i w_k \psi_{\tau_k}(y_{ij} - \xi_{ij}(\tau_k)) v_{ijk}$$

$$V_{mn}^{(2)}(\delta) = -m^{-1/2} \sum_k \sum_j \sum_i w_k \int_0^{v_{ijk}} (I(y_{ij} \leqslant \xi_{ij}(\tau_k) + t/\sqrt{m}) - I(y_{ij} \leqslant \xi_{ij}(\tau_k)))\, dt$$

$$V_{mn}^{(3)}(\delta) = \lambda_m \sum_i |\alpha_i - \delta_{0i}/\sqrt{m}| - |\alpha_i|.$$

The first term is asymptotically Gaussian. Let $\Psi_k = \mathrm{diag}(\psi_{\tau_k}(y_{ij} - \xi_{ij}(\tau_k)))$ and note that $\Psi_k e_{mn} e_{mn}^\top \Psi_l = (\tau_k \wedge \tau_l - \tau_k\tau_l) I_{mn}$. Conditions A2 and A3 imply a Lindeberg condition

and we have,

$$V_{mn}^{(1)}(\delta) = -m^{-1/2}\sum_k w_k(Z^\top \Psi_k \delta_0 + X^\top \Psi_k \delta_k) \rightsquigarrow -B\delta.$$

The second term is asymptotically quadratic in δ. Note that

$$\begin{aligned} &EV_{mn}^{(2)}(\delta) \\ &= m^{-1}\sum_k\sum_j\sum_i w_k \int_0^{v_{ijk}} \sqrt{m}(F_{ij}(\xi_{ij}(\tau_k) + t/\sqrt{m}) - F_{ij}(\xi_{ij}(\tau_k)))\,dt \\ &= m^{-1}\sum_k\sum_j\sum_i w_k \int_0^{v_{ijk}} f_{ij}(\xi_{ij}(\tau_k))t\,dt + o(1) \\ &= \frac{1}{2m}\sum_k\sum_j\sum_i w_k f_{ij}(\xi_{ij}(\tau_k))(z_{ij}^\top\delta_0 + x_{ij}^\top\delta_k/\sqrt{n})^2 + o(1) \\ &= \frac{1}{2m}\sum_k w_k(\delta_0^\top Z^\top \Phi_k Z\delta_0 + 2\delta_0^\top Z^\top \Phi_k X\delta_k/\sqrt{n} + \delta_k^\top X^\top \Phi_k X\delta_k/n) + o(1) \\ &\to \frac{1}{2}\delta^\top D_1\delta. \end{aligned}$$

The variance of $V_{mn}^{(2)}(\delta)$ converges to zero by Condition A3. Finally,

$$V_{mn}^{(3)}(\delta) = \frac{\lambda_m}{\sqrt{m}}\sum_{i=1}^n \delta_{0i}\,\mathrm{sgn}(\alpha_i) \to \lambda_0\delta_0^\top s.$$

Convexity of the objective function, V_{mn}, and the uniqueness of the minimum of V_0 yields uniformity in δ completing the argument as above. □

4. Monte Carlo

In this section a very brief glimpse into the finite sample behavior of the penalized quantile regression estimator is offered. I begin by contrasting the shrinkage effect of ℓ_1 and ℓ_2 penalty methods. Consider a simple example with $n = 50$ and $m = 5$ and response generated by the model,

$$y_{ij} = \alpha_i + u_{ij}$$

with α_i's iid from the χ_3^2 distribution, and u_{it} iid also from χ_3^2. In the left panel of Fig. 1 we illustrate the estimated, $\hat{\alpha}_i$'s as a function of the regularization parameter λ. Here we have used the estimator (2.3) with weights $w = (0.25, 0.50, 0.25)$ on the three quartiles. The x_{ij}'s were generated as Gaussian according to (4.3) below. In the right panel we illustrate the corresponding shrinkage effects for the ℓ_2 Gaussian penalty method. The ℓ_1 shrinkage method is more tolerant of large discrepancies; note that the gradient condition involves only the signs of the estimated effects, not their magnitudes, so highly unusual α_i's can

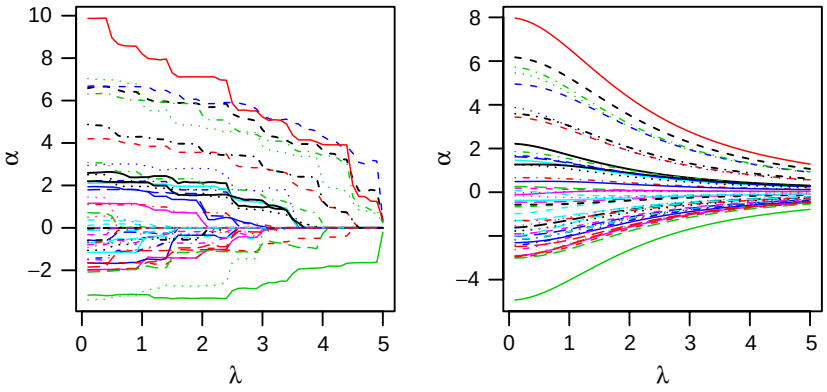

Fig. 1. Shrinkage of the fixed effect parameter estimates, $\hat{\alpha}_i$. The left panel illustrates an example of the ℓ_1 shrinkage effect. The right panel illustrates an example of the ℓ_2 shrinkage effect.

Table 1
Location-shift model

	LS	PLS	LSFE	QR	PQR	QRFE
$\mathcal{N}$						
Bias	0.0031	0.0048	0.0056	0.0048	0.0067	0.0047
RMSE	0.0847	0.0604	0.0668	0.0977	0.0781	0.0815
t_3						
Bias	−0.0062	−0.0054	−0.0051	−0.0063	−0.0101	−0.0082
RMSE	0.1377	0.1031	0.1143	0.1274	0.0881	0.0921
χ^2_3						
Bias	−0.0068	0.0002	0.0032	−0.0052	0.0063	0.0072
RMSE	0.2155	0.1503	0.1650	0.2362	0.1506	0.1513

be substantially shrunken toward zero without the extreme prejudice implied by the ℓ_2 criterion.

Two simple versions of our basic model are considered in the simulation experiments. In the first, reported in Table 1, the scalar covariate, x_{ij}, exerts a pure location shift effect. In the second, reported in Table 2, x_{ij} has a both a location and scale shift effect. In the former case the response, y_{ij}, is generated by the model,

$$y_{ij} = \alpha_i + x_{ij}\beta + u_{ij} \tag{4.1}$$

while in the latter case,

$$y_{ij} = \alpha_i + x_{ij}\beta + (1 + x_{ij}\gamma)u_{ij}. \tag{4.2}$$

Without loss of generality we will take $\beta = 0$. Interest will focus on the effect of the covariate, x_{ij}, at the median. Sample sizes are fixed, with $n = 50$, and $m = 5$ for both versions of the model. In the first version of the model the covariate effect is clearly zero, in the second version of the model it depends on the choice of the quantile of interest and the

Table 2
Location-scale-shift model

	LS	PLS	LSFE	QR	PQR	QRFE
$\mathcal{N}$						
Bias	0.0000	0.0010	0.0012	−0.0020	−0.0021	−0.0022
RMSE	0.0559	0.0501	0.0542	0.0638	0.0526	0.0556
t_3						
Bias	−0.0045	0.0000	0.0008	−0.0044	−0.0015	0.0021
RMSE	0.0948	0.0806	0.0870	0.0758	0.0620	0.0693
χ^2_3						
Bias	0.0617	0.0609	0.0608	0.0317	−0.0055	−0.0128
RMSE	0.1608	0.1292	0.1368	0.1627	0.1042	0.1092

form of the error distribution. In all cases the reported entries are based on 400 replications of the simulations.

A critical aspect governing the performance of penalty methods in these settings is the "between" versus "within" variability of the covariate. A convenient way to summarize this is the interclass correlation coefficient. If we generate x_{ij}'s as

$$x_{ij} = \gamma_i + v_{ij} \tag{4.3}$$

with γ_i and v_{ij} independent and identically distributed over i and i, j, respectively, then the interclass correlation coefficient,

$$\rho_x = \sigma_\gamma^2/(\sigma_\gamma^2 + \sigma_u^2)$$

see e.g. Scheffé [20, p. 223] is just the ordinary correlation coefficient between any two x_{ij} and x_{ik} observations with $j \neq k$. We take $\rho_x = 0.5$ in our simulations.

We consider three variants of model 1. In all three variants the x_{ij}'s are generated from (4.3) with both γ_i's and v_{ij}'s as Gaussian with unit variance. The response y is then generated from (4.1). In the first variant both the α_i's and u_{ij}'s are standard Gaussian, in the second variant both are Student t on three degrees of freedom, and in the third variant both are central χ^2_3. The interclass correlation coefficient of the response is 0.50 for all three variants.

Six estimators are considered: three from the least squares family, three from the quantile regression family. The ordinary least squares estimator (LS) simply ignores the α_i effects entirely, maximally shrinking all of these estimates to zero. The penalized least squares estimator (PLS) is the classical random effects estimator for the model (2.1) using the (known) optimal variance ratio. The least squares fixed effects, or "within" estimator (LSFE) simply implements the unpenalized least squares estimator of the model (2.1). Correspondingly, the ordinary quantile regression estimator (QR) fully shrinks the $\hat{\alpha}_i$'s to zero, the fixed effects estimator (QRFE) shrinks them not at all, and the penalized quantile regression estimator (PQR) shrinks them with λ chosen to be the ratio of scale parameters σ_u/σ_α. The quantile regression estimators minimize the objective function (2.4) with weights (0.25, 0.5, 0.25) associated with the three quartiles. In the subsequent tables, however, we focus exclusively on the performance of the median slope estimate, as a way to compare with the least squares estimation of the slope of the conditional mean relationship. Bias is computed in each case

with respect to the true slope parameter, which in all but one case is zero. The exceptional case is described in more detail below.

Table 1 reports the results of the location shift simulations. Bias is small in all cases. In the Gaussian setting we see roughly the anticipated efficiency loss due to estimating the median rather than the mean. The gain from penalization, while not overwhelming, is certainly worthwhile. In the t_3 setting the penalized quantile regression estimators do considerably better than their least squares competitors. In the χ^2_3 case the penalized quantile regression estimator does only slightly better than the unpenalized fixed effects procedure, but both are competitive with the penalized least squares results.

In the location-scale version of the model we adopt the same three distributions for generating the α_i's and the u_{ij}'s. In the location-scale model it is important that the resulting linear quantile functions do not cross, an eventuality we avoid by now taking the x_{ij}'s as χ^2_3 instead of Gaussian, thus ensuring that the scale parameter will be positive. In the Gaussian and t_3 cases, since we are focusing on the estimation of the median effect, by symmetry the effect of the covariate x_{ij} on median response is still zero. However, in the χ^2_3 case the median effect is,

$$\beta(1/2) = \beta + \gamma Q_u(1/2),$$

which in our case with $\beta = 0$ and $\gamma = 1/10$, is 0.236.

In Table 2 we report the results of the location-scale model simulations. Again, we see that the quantile regression estimators perform quite well in the t_3 case, but they now are also quite competitive even in the Gaussian case, a finding that may be attributed to the effect of the heteroscedasticity in this formulation of the model. It is also apparent that imposing more aggressive shrinkage is helpful in these cases. The comparison of performance in the χ^2_3 case is somewhat difficult, since the procedures are inherently estimating different functions. The quantile regression methods are all intended to estimate the conditional median function and do reasonably well in the sense that the bias is still very modest. The least squares estimators are targeting the conditional mean function, which is now nonlinear in x_{ij}, so we have evaluated both bias and root mean square error as if the least squares methods were also estimating the conditional median function. This obviously puts the least squares methods at some disadvantage.

5. Extensions

Many issues remain to be investigated. As in most problems of regularization there are serious issues about the choice of the regularization parameter, λ; only a prima facie case has been made that *some* degree regularization is desirable, deciding precisely how much shinkage poses challenging new questions. There are many variants of the model that would extend the oneway layout structure for the fixed effects. These include the incorporation of ordinal factors and nonparametric smoothing components. The analysis of the performance of the methods for fixed m_i sample sizes is also a critical direction for future research. Applications to reference growth curves would appear to be a natural laboratory for further development of quantile regression models for longitudinal data.

Acknowledgements

The author would like to express his appreciation to Steve Portnoy, Xuming He, Gib Bassett, Carlos Lamarche and two anonymous referees for helpful comments related to this work. This research was supported in part by NSF Grant SES 02-40781.

References

[1] J.M. Chambers, Programming with Data: A Guide to the S Language, Springer, Berlin, 1998.
[2] T.J. Cole, Fitting smoothed centile curves to reference data, J. Roy. Statist. Soc. A 151 (1988) 385–406.
[3] D. Donoho, S. Chen, M. Saunders, Atomic decomposition by basis pursuit, SIAM J. Sci. Comput. 20 (1998) 33–61.
[4] A. Goldberger, Best linear unbiased prediction in the generalized linear regression model, J. Amer. Statist. Assoc. 57 (1962) 369–375.
[5] C. Henderson, Estimation of Genetic Parameters (Abstract), Ann. Math. Statist. 21 (1950) 309–310.
[6] R. Ihaka, R. Gentleman, R: a language for data analysis and graphics, J. Comput. Graphical Statist. 5 (1996) 299–314.
[7] K. Knight, Ann. Statist. 26 (1998) 755–770.
[8] K. Knight, Comparing conditional quantile estimators: first and second order considerations, preprint, 2001
[9] K. Knight, W. Fu, Asymptotics for Lasso-type estimators, Ann. Statist. 28 (2000) 1356–1378.
[10] R. Koenker, A note on *L*-estimators for linear models, Statist. Probab. Lett. 2 (1984) 323–325.
[11] R. Koenker, G. Bassett, Regression quantiles, Econometrica 46 (1978) 33–50.
[12] R. Koenker, P. Ng, SparseM: a sparse linear algebra package for *R*, J. Statist. Software 8 (2003) 1–9.
[13] D. Lindley, A. Smith, Bayes estimates for the linear model, J. Roy. Statist. Soc. (B) 34 (1972) 1–41.
[14] F. Mosteller, On some useful "inefficient" statistics, Ann. Math. Statist. 17 (1946) 377–408.
[15] C. Rao, Linear Statistical Inference and its Applications, Wiley, New York, 1973.
[16] G. Robinson, That BLUP is a good thing: the estimation of random effects, Statist. Sci. 6 (1991) 15–31.
[17] D. Ruppert, R. Carroll, Trimmed least squares estimation in the linear model, J. Am. Stat. Assoc. 75 (1980) 828–838.
[18] M.A. Saunders, Major Cholesky would feel proud, ORSA J. Comput. 6 (1994) 23–27.
[19] Scheffé. H. Analysis of Variance, Wiley, New York, 1959.
[20] R. Tibshirani, Regression shrinkage and selection via the lasso, J. Roy. Statist. Soc. B 58 (1996) 267–288.