Journal of the Korean Data & Information Science Society
2010, **21**(2), 309–316

한국데이터정보과학회지

Support vector quantile regression for longitudinal data[†]

Changha Hwang[1]

Department of Statistics, Dankook University

Received 3 January 2010, revised 8 February 2010, accepted 12 February 2010

Abstract

Support vector quantile regression (SVQR) is capable of providing more complete description of the linear and nonlinear relationships among response and input variables. In this paper we propose a weighted SVQR for the longitudinal data. Furthermore, we introduce the generalized approximate cross validation function to select the hyperparameters which affect the performance of SVQR. Experimental results are then presented, which illustrate the performance of the proposed SVQR.

Keywords: Generalized approximate cross validation function, kernel function, longitudinal data, support vector quantile regression.

1. Introduction

Quantile regression introduced by Koenker and Bassett (1978) is gradually involving into an ensemble of practical statistical methods for estimating and conducting inference about models for conditional quantile functions. Quantile regression is an increasingly popular method for estimating the quantiles of a distribution conditional on the values of covariates. Regression quantiles are robust against the influence of outliers and, taken several at a time, they give a more complete picture of the conditional distribution than a single estimate of the center.

Just as classical linear regression methods based on minimizing the sum of squared residuals enable one to estimate a wide variety of models for conditional mean functions, quantile regression methods offer a mechanism for estimating models for the conditional median function, and the full range of other conditional quantile functions. By supplementing the estimation of conditional mean functions with techniques for estimating an entire family of conditional quantile functions, quantile regression is capable of providing a more complete statistical analysis of the stochastic relationships among random variables. The introductions and current research areas of the quantile regression can be found in Koenker and Hallock (2001), Yu *et al.* (2003).

[†] This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2009-0072101).

[1] Professor, Department of Statistics, Dankook University, Gyeonggido 448-701, Korea.
E-mail: chwang@dankook.ac.kr

The minimization problem associated with linear quantile regression is in essence the linear programming (LP) optimization problem, which is based on simplex algorithm or interior point algorithm. The current state of algorithms for nonlinear quantile regression is far less satisfactory. The widely used algorithm is interior point algorithm. Nonlinear quantile regression poses new algorithmic challenge. Refer to Koenker and Park (1996) and Koenker and Hallock (2001) for the algorithms.

For data that are clustered and/or longitudinal, mixed-effect regression models are becoming increasingly popular (Hedeker and Gibbons, 2006; Wu and Zhang, 2006; Hwang, 2008; Shim and Seok, 2008; Shim *et al.*, 2009). Mixed-effect models constitute both fixed and random effects. In clustered data, subjects are clustered within an organization such as a hospital, school, clinic or firm. In longitudinal data where individuals are repeatedly assessed, measurements are clustered within individuals. For clustered data the random effects represent cluster effects, while for longitudinal data the random effects represent subject effects. For longitudinal data, Koenker (2004) proposed a nonlinear quantile function including subject-specific bias. Geraci and Bottai (2007) proposed a nonlinear quantile function based on the asymmetric Laplace distribution.

In this paper we propose a weighted SVQR for longitudinal data. To select appropriate parameters for the achievement of high generalization performance, a commonly used method is minimizing the cross validation (CV) function. But selecting parameters using CV function is computationally formidable. Yuan (2006) proposed the generalized approximate cross validation (GACV) function for quantile spline estimation with a differentiable modified check function. Li *et al.* (2007) obtained the trace of hat matrix of estimated quantile, which leads to obtain GACV function of SVQR with a check function. See Hwang (2007), Hwang (2008), Shim and Seok (2008) and Shim *et al.* (2009) for GACV of some other kernel machines

The rest of this paper is organized as follows. In Section 2 we give a review of SVQR. In Section 3 we propose SVQR for longitudinal data and present the model selection method using GACV function. In Section 4 we perform the numerical studies through examples. In Section 5 we give the conclusions.

2. Support vector quantile regression

Let the training data set D be denoted by $(\boldsymbol{x}_i, y_i)_{i=1}^n$, with each input $\boldsymbol{x}_i \in R^d$ including a constant 1 and the response $y_i \in R$, where the response variable y_i is related to the input vector $\boldsymbol{x}_i$. Here the feature mapping function $\phi(\cdot) : R^d \to R^{d_f}$ maps the input space to the higher dimensional feature space where the dimension d_f is defined in an implicit way. An inner product in feature space has an equivalent kernel in input space, $\phi(\boldsymbol{x}_i)'\phi(\boldsymbol{x}_j) = K(\boldsymbol{x}_i, \boldsymbol{x}_j)$ (Mercer, 1909). Several choices of the kernel $K(\cdot, \cdot)$ are possible. We consider the nonlinear regression case, in which the quantile regression function $q(\boldsymbol{x})$ of the response given $\boldsymbol{x}$ can be regarded as a nonlinear function of input vector $\boldsymbol{x}$. ig With a check function $\rho_\theta(\cdot)$, the estimator of the θ-th quantile regression function can be defined as any solution to the optimization problem,

$$\min \frac{1}{2}\boldsymbol{w}'\boldsymbol{w} + C\sum_{i=1}^{n} \rho_\theta(y_i - q(\boldsymbol{x}_i)) \tag{2.1}$$

where $\rho_\theta(r) = \theta r I_{(r\geq 0)} + (1-\theta) r I_{(r<0)}$. We can express the regression problem by formulation for SVQR as follows:

$$\min \frac{1}{2}\boldsymbol{w}'\boldsymbol{w} + C\sum_{i=1}^{n}(\theta\xi_i + (1-\theta)\xi_i^*) \tag{2.2}$$

subject to

$$\begin{aligned} &y_i - \boldsymbol{w}'\phi(\boldsymbol{x}_i) - b \leq \xi_i, \\ &\boldsymbol{w}'\phi(\boldsymbol{x}_i) + b - y_i \leq \xi_i^*, \\ &\xi_i, \xi_i^* \geq 0, \end{aligned}$$

where C is a regularization parameter penalizing the training errors. We construct a Lagrange function as follows:

$$\begin{aligned} L =& \frac{1}{2}\boldsymbol{w}'\boldsymbol{w} + C\sum_{i=1}^{n}(\theta\xi_i + (1-\theta)\xi_i^*) - \sum_{i=1}^{n}\alpha_i(\xi_i - y_i + \boldsymbol{w}'\phi(\boldsymbol{x}_i) + b) \\ &- \sum_{i=1}^{n}\alpha_i^*(\xi_i^* + y_i - \boldsymbol{w}'\phi(\boldsymbol{x}_i) - b) - \sum_{i=1}^{n}(\eta_i\xi_i + \eta_i^*\xi_i^*). \end{aligned} \tag{2.3}$$

We notice that the positivity constraints $\alpha_i, \alpha_i^*, \eta_i, \eta_i^* \geq 0$ should be satisfied. After taking partial derivatives of equation (2.3) with regard to the primal variables $(\boldsymbol{w}, \xi_i, \xi_i^*)$ and plugging them into equation (2.3), we have the optimization problem below.

$$\max -\frac{1}{2}\sum_{i,j=1}^{n}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)K(\boldsymbol{x}_i, \boldsymbol{x}_j) + \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)y_i \tag{2.4}$$

with constraints

$$\sum_{i=1}^{n}(\alpha_i - \alpha_i^*) = 0, \alpha_i \in [0, \theta C], \alpha_i^* \in [0, (1-\theta)C].$$

Solving the above equation with the constraints, the optimal Lagrange multipliers, α_i and α_i^* are determined. And then the estimator of the θ-th quantile regression function given the input vector $\boldsymbol{x}$ is obtained as follows,

$$\widehat{q}_\theta(\boldsymbol{x}) = \sum_{i=1}^{n}K(\boldsymbol{x}_i, \boldsymbol{x})(\alpha_i - \alpha_i^*) + b. \tag{2.5}$$

The optimal value of b is obtained by employing Karush-Kuhn-Tucker conditions (Kuhn and Tucker, 1951) as follows,

$$b = \frac{1}{n_s}\sum_{i\in I_s}(y_i - K(\boldsymbol{x}_i, \boldsymbol{x})(\boldsymbol{\alpha} - \boldsymbol{\alpha}^*))$$

where $I_s = \{i = 1, 2, \cdots, n : (\theta-1)C < \alpha_i - \alpha_i^* < \theta C\}$, n_s is the size of I_s.

In the nonlinear case, $\boldsymbol{w}$ is no longer explicitly given. However, it is uniquely defined in the weak sense by the dot products. Here the linear regression model can be regarded as a special case of the nonlinear regression model by using identity feature mapping function, that is, $\phi(\boldsymbol{x}) = \boldsymbol{x}$ which implies the linear kernel such that $K(\boldsymbol{x}_1, \boldsymbol{x}_2) = \boldsymbol{x}_1'\boldsymbol{x}_2$.

3. Weighted SVQR

In this section we propose a weighted SVQR for longitudinal data by incorporating the weights obtained from median regression into SVQR represented in Section 2.

Karlsson (2008) used the weighted objective function to take correlation of $\mathbf{e}_i = (e_{i1}, \cdots, e_{in_i})'$ into the estimation of quantiles,

$$\min \sum_{i=1}^{N} \sum_{j=1}^{n_i} \omega_{ij} \rho_\theta (y_{ij} - f(\boldsymbol{x}_{ij}, \boldsymbol{\beta})),$$

where ω_{ij} is the weight on the j th observation of the i th subject and $f(\boldsymbol{x}_{ij}, \boldsymbol{\beta})$ is a known nonlinear parametric function.

For the longitudinal data, we define the estimator of the θ-th quantile regression function as any solution to the optimization problem,

$$\min \frac{1}{2} \boldsymbol{w}'\boldsymbol{w} + C \sum_{i=1}^{N} \sum_{j=1}^{n_i} \omega_{ij} \rho_\theta (y_{ij} - q(\boldsymbol{x}_{ij})) \tag{3.1}$$

Given ω_{ij} for $i = 1, \cdots, N, j = 1, \cdots, n_i$, we have the optimization problem below.

$$\max -\frac{1}{2} \sum_{i,k=1}^{N} \sum_{j=1}^{n_i} \sum_{l=1}^{n_k} (\alpha_{ij} - \alpha_{ij}^*)(\alpha_{kl} - \alpha_{kl}^*) K^{ij,kl} + \sum_{i=1}^{N} \sum_{j=1}^{n_i} (\alpha_{ij} - \alpha_{ij}^*) y_{ij} \tag{3.2}$$

with constraints

$$, \sum_{i=1}^{N} \sum_{j=1}^{n_i} (\alpha_{ij} - \alpha_{ij}^*) = 0, \alpha_{ij} \in [0, \omega_{ij}\theta C], \alpha_{ij}^* \in [0, \omega_{ij}(1-\theta)C],$$

where K is the $N_n \times N_n$ kernel matrix , $N_n = \sum_{i=1}^{N} n_i$ and $K^{ij,kl}$ is an element of K corresponding to $\boldsymbol{x}_{ij}$ and $\boldsymbol{x}_{kl}$. Here we use two types of weights proposed by Karlsson (2008),

$$e_{ij} = y_{ij} - \widehat{q_{0.5}}(\boldsymbol{x}_{ij}), u_i = \frac{1}{n_i} \sum_{j=1}^{n_i} |e_{ij} - \bar{e_{i.}}|, u_i = \sqrt{\frac{1}{n_i} \sum_{j=1}^{n_i} (e_{ij} - \bar{e}_{i.})^2}, \omega_{ij} = \frac{1}{u_i} \tag{3.3}$$

where $\widehat{q}_{0.5}(\boldsymbol{x}_k)$ is obtained from (3.1) with $\omega_{ij} = 1$.

The estimator of the θ -th quantile regression function given the input vector $\boldsymbol{x}_{ij}$ is obtained as follows.

$$\widehat{q_\theta}(\boldsymbol{x}_{ij}) = \sum_{k=1}^{N} \sum_{l=1}^{n_k} K^{ij,kl} (\alpha_{kl} - \alpha_{kl}^*) + b,$$

where $K^{ij,kl}$ is the element of K corresponding to $\boldsymbol{x}_{ij}$ and $\boldsymbol{x}_{kl}$.

The functional structures of SVQR is characterized by hyperparameters - the regularization parameter C and the kernel parameters. We define the cross validation (CV) function used in SVQR for longitudinal data as,

$$CV(\boldsymbol{\lambda}) = \frac{1}{N_n}\sum_{i=1}^{N}\sum_{j=1}^{n_i}\omega_{ij}\rho_\theta(y_{ij} - \widehat{q}_\theta^{(-ij)}(\boldsymbol{x}_{ij}|\boldsymbol{\lambda})), \tag{3.4}$$

where $\boldsymbol{\lambda}$ is a set of hyperparameters and $\widehat{q}_\theta^{(-ij)}(\boldsymbol{x}_{ij}|\boldsymbol{\lambda})$ is the θ -th quantile regression function estimated without $(\boldsymbol{x}_{ij}, y_{ij})$. Since for each candidates of hyperparameters, $\widehat{q}_\theta^{(-ij)}(\boldsymbol{x}_{ij}|\boldsymbol{\lambda})$ for $i = 1, \cdots, N$, $j = 1, \cdots, n_i$, should be evaluated, selecting parameters using CV function is computationally formidable.

For convenience, we now rearrange y_{ij}'s using single index and then denote each response by $y_k, k = 1, \cdots, N_n$ That is, y_{ij} 's are denoted as follows:

$$y_1 = y_{11}, \cdots, y_{n_1} = y_{1,n_1}, y_{n+1} = y_{21}, \cdots, y_{2n_2} = y_{2,n_2}, \cdots, y_{N_n} = y_{N,n_N}.$$

We also rearrange (ω_{ij}, $\boldsymbol{x}_{ij}$)'s and then denote these pairs using single index in accordance with y_{ij}'s. Then the estimator of the θ -th quantile regression function given the input vector $\boldsymbol{x}$ is obtained as follows.

$$\widehat{q}_\theta(\boldsymbol{x}) = \sum_{k=1}^{N_n} K(\boldsymbol{x}_k, \boldsymbol{x})(\alpha_k - \alpha_k^*) + b,$$

and the CV function (3.4) can be rewritten as

$$CV(\boldsymbol{\lambda}) = \frac{1}{N_n}\sum_{k=1}^{N_n}\omega_k\rho_\theta(y_k - \widehat{q}_\theta^{(-k)}(\boldsymbol{x}_k|\boldsymbol{\lambda})). \tag{3.5}$$

By leaving-out-one lemma (Craven and Wahba, 1979),

$$(y_k - \widehat{q}_\theta^{(-k)}(\boldsymbol{x}_k|\boldsymbol{\lambda})) - (y_k - \widehat{q}_\theta(\boldsymbol{x}_k|\boldsymbol{\lambda})) = \widehat{q}_\theta(\boldsymbol{x}_k|\boldsymbol{\lambda}) - \widehat{q}_\theta^{(-k)}(\boldsymbol{x}_k|\boldsymbol{\lambda}) \approx \frac{\partial\widehat{q}_\theta(\boldsymbol{x}_k|\boldsymbol{\lambda})}{\partial y_k}(y_k - \widehat{q}_\theta^{(-k)}(\boldsymbol{x}_k|\boldsymbol{\lambda}))$$

we have

$$(y_k - \widehat{q}_\theta^{(-k)}(\boldsymbol{x}_k|\boldsymbol{\lambda})) \approx \frac{y_k - \widehat{q}_\theta(\boldsymbol{x}_k|\boldsymbol{\lambda})}{1 - \dfrac{\partial\widehat{q}_\theta(\boldsymbol{x}_k|\boldsymbol{\lambda})}{\partial y_k}}.$$

Then the approximate cross validation (ACV) function can be obtained as

$$ACV(\boldsymbol{\lambda}) = \frac{1}{N_n}\sum_{k=1}^{N_n}\omega_k\left(\frac{y_k - \widehat{q}_\theta(\boldsymbol{x}_k|\boldsymbol{\lambda})}{1 - \dfrac{\partial\widehat{q}_\theta(\boldsymbol{x}_k|\boldsymbol{\lambda})}{\partial y_k}}\right) = \frac{1}{N_n}\sum_{k=1}^{N_n}\omega_k\frac{\rho_\theta(y_k - \widehat{q}_\theta(\boldsymbol{x}_k|\boldsymbol{\lambda}))}{1 - h_{kk}},$$

where H is the hat matrix such that $\widehat{q}_\theta(\boldsymbol{x}|\boldsymbol{\lambda}) = H\boldsymbol{y}$ with the (k,l) thelement $h_{kl} = \partial\widehat{q}_\theta(\boldsymbol{x}_k)/\partial y_l$. By Li *et al.*(2007) we have $tr(H)$ = a size of the set E,

$$E = \{k = 1, \cdots, N_n | y_k - q_\theta(\boldsymbol{x}_k) = 0, 0 <= \alpha_k <= \omega_k\theta C, 0 <= \alpha_k^* <= \omega_k(1-\theta)C\}.$$

Replacing h_{ii} by their average $tr(H)/n$, the generalized approximate cross validation (GACV) function can be obtained as

$$GACV(\boldsymbol{\lambda}) = \frac{\sum_{k=1}^{N_n} \omega_k \rho_\theta(y_k - \widehat{q}(\boldsymbol{x}_k|\boldsymbol{\lambda}))}{N_n - |E|}, \tag{3.6}$$

where $|E|$ is a size of the set E.

4. Numerical studies

In this section, we illustrate the performance of the proposed quantile regression estimation through the simulated example on the nonlinear quantile regression case. We set 2 subjects and 200 observations in each subject for a data set. We generate 100 data sets for the numerical studies. The univariate input observations x are equally spaced ranging from 0 to 1, the corresponding responses y are drawn from a univariate normal distribution with the variance that varies with subject as follows,

$$y_{ij} \sim N\left(\sin(1.5\pi x_{ij}), 2i+1\right), i = 1, 2, j = 1, 2, \cdots, 200.$$

The radial basis kernel function is utilized in this example, which is

$$K(x_1, x_2) = \exp(-\frac{1}{\sigma^2}||x_1 - x_2||^2).$$

We use the weight w_{ij} computed from the equation (3.3) with $u_i = \sqrt{\frac{1}{n_i}\sum_{j=1}^{n_i}(e_{ij} - \bar{e}_{i.})^2}$. Figure 4.1 shows a family of quantile functions estimated by the weighted SVQR and unweighted SVQR which ignores the subject effects. The estimated quantile regression functions for $\theta = 0.1, 0.5, 0.9$ are superimposed on the scatter plot. To illustrate the estimation performance of the weighted SVQR, we compare it with SVQR via 100 data sets, where the mean squared error (MSE) is used as the prediction erformance measure defined by

$$MSE = \frac{1}{100}\sum_{l=1}^{100}\sum_{i=1}^{2}\sum_{j=1}^{200}(\widehat{q}_\theta^{(l)}(x_{ij}) - q_\theta^{(l)}(x_{ij}))^2 for \theta = 0.1, 0.5, 0.9.$$

The averages of 100 MSEs and their standard deviations from the weighted SVQR and the unweighted SVQR are obtained in Table 4.1. In this example, we can see that the weighted SVQR provides better estimation performance than SVQR for longitudinal data.

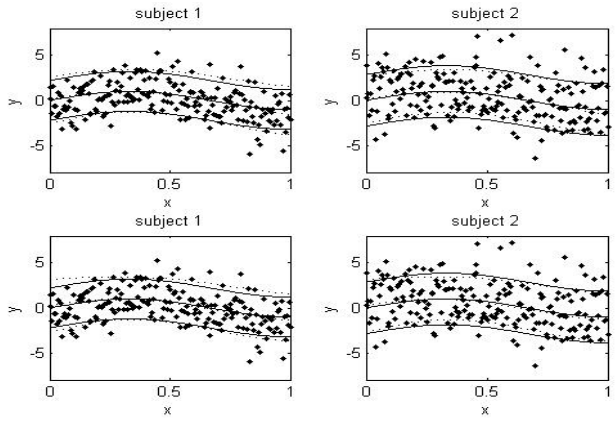

Figure 4.1 An illustration of the proposed weighted SVQR (Upper) and SVQR (Lower) for one of 100 data sets. True quantile regression function (solid line) and the estimated quantile regression function (dotted line) for $\theta = 0.1, 0.5, 0.9$ are superimposed on the scatter plot.

Table 4.1 The averages of 100 MSEs and their standard deviations in parenthesis of the proposed weighted SVQR and the unweighted SVQR for data sets.

θ	weighted SVQR	unweighted SVQR
0.1	0.2110(0.0750)	0.2118(0.0831)
0.5	0.0549(0.0290)	0.0562(0.0307)
0.9	0.2060(0.0686)	0.2147(0.0775)

5. Conclusions

In this paper, we dealt with estimating the nonlinear quantile regression function by SVQR for longitudinal data and obtained GACV function for the proposed procedure. Through the example we showed that the proposed procedure can be useful in analyzing longitudinal data.

References

Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerical Mathematics*, **31**, 377-403.

Geraci, M. and Bottai, M. (2007). Quantile regression for longitudinal data using the asymmetric Laplace distribution. *Biostatistics*, **8**, 140-154.

Hedeker, D. and Gibbons, R. D. (2006). *Longitudinal Data Analysis*, John Wiley & Sons.

Hwang, C. (2007). Kernel machine for Poisson regression. *Journal of Korean Data & Information Science Society*, **18**, 767-772 .

Hwang, C. (2008). Mixed effects kernel binomial regression. *Journal of Korean Data & Information Science Society*, **19**, 1327-1334 .

Karlsson, A. (2008). Nonlinear quantile regression estimation for longitudinal data. *Communications in Statistics - Simulation and Computation*, **37**, 114-131.

Kimeldorf, G. S. and Wahba, G. (1971). Some results on Tchebycheffian spline functions. *Journal of Mathematical Analysis and its Applications*, **33**, 82-95.

Koenker, R. (2004). Quantile regression for longitudinal data. *Journal of Multivariate Analysis*, **91**, 74-89.
Koenker, R. and Bassett. G. (1978). Regression quantile. *Econometrica*, **46**, 33-50.
Koenker, R. and Hallock, K. F. (2001). Quantile regression. *Journal of Economic Perspectives*, **40**, 122-142.
Koenker, R. and Park, B. J. (1996). An interior point algorithm for nonlinear quantile regression. *Journal of Econometrics*, **71**, 265-283.
Li, Y., Liu, Y. and Zhu, J. (2007). Quantile regression in reproducing kernel hilbert spaces. *Journal of the American Statistical Association*, **102**, 255-268.
Mercer, J. (1909). Functions of positive and ngative type and their cnnection with theory of integral equations. *Philosophical Transactions of Royal Society*, 415-446.
Shim , J. and Seok, K. H. (2008). Kernel poisson regression for longitudinal data. *Journal of Korean Data & Information Science Society*, **19**, 1353-1360 .
Shim , J., Kim, T. Y., Lee, S. and Hwang, C. (2009). Credibility estimation via kernel mixed effects model. *Journal of Korean Data & Information Science Society* **20**, 445 -452 .
Smola, A. and Scholkopf, B. (1998). On a kernel-based method for pattern recognition, regression, approximation and operator inversion. *Algorithmica*, **22**, 211-231.
Wu, H. and Zhang, J. (2006). *Nonparametric regression methods for longitudinal data analysis*, Wiley.
Yu, K., Lu, Z. and Stander, J. (2003). Quantile regression: Applications and current research area. *The Statistician*, **52**, Part3, 331-350.
Yuan, M. (2006). GACV for quantile smoothing splines. *Computational Statistics and Data Analysis*, **50**, 813-829.