Statistics and Probability Letters 223 (2025) 110431

Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

Variable-selection consistency of linear quantile regression by validation set approach

Suin Kim, Sarang Lee, Nari Shin, Yoonsuh Jung *

Departments of Statistics, University of Korea, 145, Anam-ro, Seoul, 02841, South Korea

ARTICLE INFO

Keywords:
Check loss
Cross-validation
High-dimensional quantile regression
Model selection

ABSTRACT

We consider the problem of variable selection in the quantile regression model by cross-validation. Although cross-validation is commonly used in quantile regression for model selection, its theoretical justification has not yet been verified. In this work, we prove that cross-validation with the check loss function can lead to variable-selection consistency in quantile regression. Specifically, we investigate its asymptotic properties in linear quantile regression and its penalized version under both fixed and diverging number of parameters. For penalized models, penalties with the oracle property combined with cross-validation are shown to provide variable-selection consistency. In general, one of the crucial requirements for this consistency to hold is that the validation set size should be asymptotically equivalent to the total number of observations, which is also required in the conditional mean linear regression.

1. Introduction

When measuring the prediction accuracy of conditional mean regression, it is natural to use the squared error loss as the validation function. Shao (1993) showed that the squared error loss in cross-validation (CV) leads to variable-selection consistency under certain conditions. One of these conditions requires the size of validation set to be asymptotically equivalent to the total number of observations. When fitting a conditional quantile regression at quantile q, the check loss function, $\rho_q(u) = (q - I(u < 0))u$, is employed (Koenker and Bassett, 1978). Therefore, quantile regression models commonly use the check loss function as a validation function. However, there is a lack of theoretical support for employing the check loss function in CV for model (or tuning parameter) selection. In this work, we focus on establishing the variable-selection consistency of CV under quantile regression models.

The main contribution of this paper is to demonstrate that cross-validation can achieve variable-selection consistency under linear quantile regression models. A key finding is that allocating the majority of the observations to model validation leads to more accurate model selection. This contrasts with the common practice among researchers of allocating more observations for training. This observation parallels the findings of Shao (1993) in the context of conditional mean regression, and we arrive at the same conclusion for both linear and penalized linear quantile regression models. However, it is crucial to maintain a reasonable number of observations for training rather than allocating an overwhelming majority to validation.

2. Variable selection in quantile regression

We consider the linear model:

$$y_i = x_i'\beta_q + \epsilon_i, \text{ for } i = 1, \ldots, n, \tag{1}$$

* Corresponding author.
E-mail addresses: ksi2002@korea.ac.kr (S. Kim), lsr0403@korea.ac.kr (S. Lee), shinnari98@naver.com (N. Shin), yoons77@korea.ac.kr (Y. Jung).

https://doi.org/10.1016/j.spl.2025.110431
Received 27 October 2024; Received in revised form 11 March 2025; Accepted 1 April 2025
Available online 15 April 2025
0167-7152/© 2025 Elsevier B.V. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

where y_i is an independent random variable with distribution function F_i and x_i is a vector of fixed size p. Throughout this paper, we assume that each F_i is absolutely continuous and has a continuous probability density function f_i for $i = 1, \ldots, n$. The vector $\beta_q = (\beta_{q,1}, \ldots, \beta_{q,p})'$ represents the true qth quantile regression coefficients. The ϵ_i is the error term which satisfies $P(\epsilon_i \le 0) = q$. The quantile regression estimate $\hat{\beta}_q$ is defined as the minimizer of $\sum_{i=1}^{n} \rho_q(y_i - x_i'\beta_q)$. The asymptotic properties of $\hat{\beta}_q$ are well established in Koenker and Bassett (1978) and Koenker (2005) under model (1). Our focus here is to investigate whether a true model can be selected with probability tending to one using cross-validation (CV).

To perform CV, we randomly split the data into two parts. The first part, consisting of n_c observations, is used for model construction. The remaining $n_v = n - n_c$ observations are used for model validation. Let $\xi_i = \xi_i(q)$ be the true conditional quantile of y_i, given by $x_i'\beta_q$, and let M be the selected model with $|M|$ number of variables. Here, $|\cdot|$ denotes cardinality.

A model M can be classified into one of the following categories: a true model $M^* = \{j \in \{1, \ldots, p\} : \beta_{q,j} \neq 0\}$, an over-specified model M_o (where $M_o \supset M^*$ and $M_o \neq M^*$), and an under-specified model M_u (where $M_u \not\supset M^*$). M_o includes all the variables in M^* and at least one variable with a zero coefficient, whereas M_u misses at least one variable with a nonzero coefficient. We assume that a unique M^* exists.

With these notations, β_q^M represents the regression coefficients for model M, and $x_{i,M} \in \mathbb{R}^{|M|}$ are the corresponding covariates for the ith observation. Note that their dimensions vary and depend on the model under consideration. Since we assume p is fixed in this section, which will be relaxed later, the number of over-specified and under-specified models is finite. Now, we assume the following:

(C1) $f_i(\xi_i)$ is uniformly bounded away from 0 and ∞ at ξ_i for $i = 1, 2, \ldots, n$.
(C2) $n_c \to \infty$ and $n_v/n \to 1$ as $n \to \infty$.
(C3) $\max_{i=1,\ldots,n} \|x_i\|/\sqrt{n} \to 0$.
(C4) $\lim_{n\to\infty} n^{-1} \sum_i x_i x_i' =: \Sigma_0$ is positive definite.
(C5) $\lim_{n\to\infty} n^{-1} \sum_i f_i(\xi_i) x_i x_i' =: \Sigma_1$ is positive definite.
(C6) $\liminf_{n\to\infty} \sum_{i=1}^{n} \rho_q(y_i - x_{i,M_u}'\beta_q^{M_u})/n = \liminf_{n\to\infty} k_n > 0$ for $0 < q < 1$.

Condition (C2) implies that more accurate validation is necessary for variable-selection consistency compared to model training, as the number of observations in the validation set, n_v, exceeds the number of observations in the training set, n_c. This is one of our key conditions and is also crucial for variable-selection consistency under conditional mean regression (Shao, 1993). Since Σ_0 and Σ_1 are assumed to exist in conditions (C4) and (C5), their submatrices for any model M also exist. These conditions imply that the variables are linearly independent with probability tending to one. Failure to meet these conditions results in asymptotic multicollinearity. Let X_M be a design matrix of model M. For an over-specified model (M_o), we assume that the true regression surface can be recovered, i.e., $X_{M_o}\beta_q^{M_o} = X\beta_q$. Besides, it is natural to assume that under the under-specified model, M_u, $X_{M_u}\beta_q^{M_u}$ cannot recover $X\beta_q$. Thus, for each M_u, we consider a minimal model identifiability condition (C6), which is analogous to $\sum_{i=1}^{n}(y_i - x_{i,M_u}'\beta_q^{M_u})^2/n > 0$ as $n \to \infty$ under squared error loss (Zhang, 1993).

Under model (1), we first fit the linear quantile regression using the training data. With the estimated regression coefficients, the validation error is measured by calculating the check loss between the observed y_i and the predicted $\hat{y}_i$ in the validation set. The averaged cross-validation score (CVS) for a model M is given by:

$$CVS(M) = \frac{1}{n_v} \sum_{i\in\mathcal{V}} \rho_q(y_i - x_{i,M}'\hat{\beta}_{(-\mathcal{V})}^{M}), \tag{2}$$

where $\mathcal{V}$ is an index set of validation set, and $\hat{\beta}_{(-\mathcal{V})}^{M}$ is the estimate of β_q from the training data using model M.

Given a selected model M, the difference in CVS between models M and M^* is given by:

$$CVS(M) - CVS(M^*) = \frac{1}{n_v} \sum_{i\in\mathcal{V}} \left\{ \rho_q(y_i - x_{i,M}'\hat{\beta}_{(-\mathcal{V})}^{M}) - \rho_q(y_i - x_{i,M^*}'\hat{\beta}_{(-\mathcal{V})}^{M^*}) \right\}.$$

Lemma 2.1 states that the CVS from an over-specified model is larger than that from the true model with probability tending to one. As a result, it follows that any over-specified model will, with a probability tending to one, not be selected by cross-validation.

Lemma 2.1. *Suppose conditions (C1)-(C6) hold, then for any over-specified model M_o, we have*

$$P(CVS(M_o) - CVS(M^*) > 0) \to 1, \text{ as } n \to \infty. \tag{3}$$

Lemma 2.2 further demonstrates that under-specified models are likewise unlikely to be selected by cross-validation as the sample size increases.

Lemma 2.2. *Under conditions (C1)–(C6), for any under-specified model M_u, we have*

$$P(CVS(M_u) - CVS(M^*) > 0) \to 1, \text{ as } n \to \infty. \tag{4}$$

By synthesizing the findings of Lemmas 2.1 and 2.2, the subsequent theorem establishes the variable-selection consistency of cross-validation employing the check loss function.

Theorem 2.3. *Under conditions (C1)–(C6), the model $\hat{M}$ selected by the cross-validation satisfies*

$$\lim_{n\to\infty} P(\hat{M} = M^*) = 1.$$

The proof follows straightforwardly from the results of Lemmas 2.1 and 2.2.

3. Variable selection in penalized quantile regression

3.1. Fixed number of parameters

In this section, we consider penalized quantile regression under linear model (1) within the regularization framework:

$$\min_{\beta_q} \sum_{i=1}^{n} \rho_q(y_i - x_i'\beta_q) + n\sum_{j=1}^{p} p_\lambda(\beta_{q,j}), \tag{5}$$

where p is fixed and $p < n$. LASSO estimator, proposed by Tibshirani (1996), is known to be inconsistent under squared error loss (Leng et al., 2006; Meinshausen and Bühlmann, 2006; Zou, 2006). In contrast, estimators with SCAD (Fan and Li, 2001) and adaptive LASSO (Zou, 2006) penalties have an oracle property under linear quantile regression models, as proved by Wu and Liu (2009) for the case when $p < n$.

In Section 2, we required a consistent estimator to investigate the variable-selection consistency of CV. Therefore, we employ consistent estimators such as SCAD and adaptive LASSO in this section, excluding LASSO. The SCAD penalty is defined as:

$$p_\lambda(|\theta|) = \begin{cases} \lambda|\theta|, & \text{if} \quad 0 \le |\theta| < \lambda, \\ \frac{(a^2-1)\lambda^2 - (|\theta| - a\lambda)^2}{2(a-1)}, & \text{if} \quad \lambda \le |\theta| < a\lambda, \\ \frac{(a+1)\lambda^2}{2}, & \text{if} \quad |\theta| \ge a\lambda, \end{cases} \tag{6}$$

where $a = 3.7$ is suggested by Fan and Li (2001). The adaptive LASSO quantile regression estimator finds the minimizer of

$$\sum_{i=1}^{n} \rho_q(y_i - x_i'\beta_q) + \lambda\sum_{j=1}^{p} \hat{w}_j|\beta_{q,j}|, \tag{7}$$

where the weights are often estimated as $\hat{w}_j = 1/(|\tilde{\beta}_{q,j}|^\gamma + e)$ for $j = 1, \dots, p$ and $\gamma, e > 0$ are some constants. Here, $\tilde{\beta}_{q,j}$ can be the traditional quantile regression estimator without the penalty in (5) or a penalized estimator (such as ridge or LASSO).

In practice, variable selection is conducted by identifying a penalty parameter λ that minimizes the CVS in (2). We denote the penalty parameter selected using the training set as λ_{n_c}. Although the penalized estimator $\hat{\beta}^M_{(-v)}$ depends on λ_{n_c} from the training set, we omit this dependence to avoid notational complexity.

We now introduce additional assumptions for λ alongside conditions (C1)-(C6) given in Section 2. Note that condition (C7) applies to SCAD and (C8) applies to adaptive LASSO.

(C7) (SCAD) $\lambda_{n_c} \to 0$ and $\sqrt{n_c}\lambda_{n_c} \to \infty$ as $n_c \to \infty$.
(C8) (Adaptive LASSO) $\sqrt{n_c}\lambda_{n_c} \to \infty$ and $n_c^{(\gamma+1)/2}\lambda_{n_c} \to \infty$.

Conditions (C7) and (C8) are employed by Wu and Liu (2009) to establish the consistency of the SCAD and adaptive LASSO, respectively. These conditions ensure $\sqrt{n_c}$-consistent estimator in the training set.

Theorem 3.1. *Suppose conditions (C1)-(C6) hold. In addition, we impose (C7) for SCAD penalty and (C8) for adaptive LASSO penalty. Then, for quantile regression with SCAD or adaptive LASSO penalties, the model $\hat{M}$ selected by cross-validation satisfies*

$$\lim_{n\to\infty} P(\hat{M} = M^*) = 1.$$

Theorem 3.1 demonstrates that CV is consistent in variable selection for penalized quantile regression with SCAD and adaptive LASSO penalties. It should be noted that conditions (C4) and (C5) in Section 2 imply that $p < n$ and that p is fixed. Amin et al. (2015) investigated the asymptotic properties of SCAD-penalized quantile regression model for diverging p, while still requiring $p < n$ as n increases. On the other hand, Wang et al. (2012) established the asymptotic properties of SCAD and MCP under $p > n$ for local minima, but were unable to identify an oracle estimator among multiple minima. Zhang and Zhang (2012) and Wang et al. (2013) provided general theory and oracle properties of SCAD and MCP under squared error loss. The following subsection examines the case of diverging p with $p < n$ for the quantile regression model with SCAD penalty.

3.2. Diverging number of parameters

We consider the same model given in (5) with the SCAD penalty but assume that $p = p_n$ diverges as n increases, under $p < n$. In addition, we partition x_i' as $x_i' = (u_i', v_i')$, where u_i' represents the active part of x_i' for which the corresponding components of the coefficient vector β_q are nonzero. The dimensions of u_i and v_i are s_n and t_n, respectively, with $s_n + t_n = p_n$. We assume conditions (C1)–(C6) and introduce the following additional assumptions for model construction.

(D1) $\max_{i=1,\ldots,n_c} \|v_i\| = O_p(\sqrt{p_{n_c}/n_c} + \lambda_{n_c})$ where $\lambda_{n_c} \to 0$.

(D2) $\lambda_{n_c}^{-1}\sqrt{p_{n_c}/n_c} \to 0$ as $\lambda_{n_c} \to 0$.

(D3) $p_{n_c}^3/n_c \to \infty$ as $n_c \to \infty$.

(D4) $\lim_{n_c\to\infty} {n_c}^{-1} \sum_i u_{i,M} u'_{i,M} = \Sigma_{2,M}$ is positive definite.

Conditions (D1)–(D4), initially assumed by Amin et al. (2015) for the consistency of the SCAD estimator, are modified here to suit model construction using a training set of size n_c. Condition (D4) is also assumed in Huang et al. (2008). When the number of true nonzero regression coefficients, s_n, diverges, we need to replace p_n^3/n in condition (D3) with the stronger condition $s_n^3 p_n^3/n$. Then, Theorem 3.1 in Amin et al. (2015) proved that $\|\hat{\beta}_q - \beta_q\| = O_p(\sqrt{p_n/n} + \lambda_n)$. This aligns with the results of He and Shao (2000) for increasing dimensions, showing $\|\hat{\beta}_q - \beta_q\| = O_p(\sqrt{p_n/n})$ under model (5) without the penalty term. Under these conditions, we consider a model selection procedure using CV with the SCAD-penalized quantile regression. Lemmas 3.2 and 3.3 show that over-specified models and under-specified models are not selected by CV with probability tending to one.

Lemma 3.2. *Consider the SCAD-penalized linear quantile regression model (5). Suppose conditions (C1)–(C6) and (D1)–(D4) hold. Then for any over-specified model M_o, we have*

$$P(CVS(M_o) - CVS(M^*) > 0) \to 1, \text{ as } n \to \infty. \tag{8}$$

Lemma 3.3. *Consider the SCAD-penalized linear quantile regression model (5). Suppose conditions (C1)–(C6) and (D1)–(D4) hold. Then for any under-specified model M_u, we have*

$$P(CVS(M_u) - CVS(M^*) > 0) \to 1, \text{ as } n \to \infty \tag{9}$$

By combining Lemmas 3.2 and 3.3, we establish the variable-selection consistency of SCAD-penalized quantile regression when the number of parameters diverges.

Theorem 3.4. *Suppose conditions (C1)–(C6) and (D1)–(D4) hold. Then under SCAD-penalized quantile regression, the model $\hat{M}$ selected by cross-validation satisfies*

$$\lim_{n\to\infty} P(\hat{M} = M^*) = 1.$$

4. Simulation studies

We evaluate the finite-sample performance of cross-validation (CV) in quantile regression. Although no theoretical proof of variable-selection consistency has existed, it is intuitive that variable-selection performance in QR improves with increasing sample size. Therefore, our simulations do not seek to demonstrate variable-selection consistency explicitly. Instead, motivated by the condition $n_v/n \to 1$ as $n \to \infty$, we specifically examine how varying the proportion of the validation set influences variable selection and estimation accuracy. We examine validation proportions of 20%, 40%, 60%, and 80% across quantiles $q = 0.1, 0.2, 0.3, 0.4, 0.5$. Simulation results for non-penalized quantile regression indicate that increasing the validation proportion generally leads to proper variable selection and precise estimation. For penalized quantile regression, the validation proportion substantially improves estimation accuracy. However, maintaining a sufficient number of training observations is crucial for model reliability. Detailed results and additional analyses are provided in the supplementary materials.

5. Discussion

In this study, we examine the variable-selection consistency of CV in linear and penalized linear quantile regression models. Our study considers both fixed and diverging numbers of variables. We confirm that a higher validation set proportion is necessary to prevent the selection of misspecified models, both theoretically and empirically. The theoretical relationship between oracle estimators and variable-selection consistency through cross-validation is an intriguing area for future research. Given that the oracle properties of the quantile regression estimator for $p > n$ have been proven by Fan et al. (2014), the variable-selection consistency of cross-validation under $p > n$ remains an open question for future investigation.

Acknowledgments

Jung's work has been partially supported by National Research Foundation of Korea (NRF) grants funded by the Korean government (MSIT) (No. 2022R1F1A1071126 and No. RS-2022-NR068754).

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.spl.2025.110431.

Data availability

Data will be made available on request.

References

Amin, M., Song, L., Thorlie, M.A., Wang, X., 2015. SCAD-penalized quantile regression for highdimensional data analysis and variable selection. Stat. Neerl. 69 (3), 212–235.
Fan, J., Fan, Y., Barut, E., 2014. Adaptive robust variable selection. Ann. Statist. 42 (1), 324–351.
Fan, J., Li, R., 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. J. Amer. Statist. Assoc. 96 (456), 1348–1360.
He, X., Shao, Q.-M., 2000. On parameters of increasing dimensions. J. Multivariate Anal. 73 (1), 120–135.
Huang, J., Horowitz, J.L., Ma, S., 2008. Asymptotic properties of bridge estimators in sparse high-dimensional regression models. Ann. Statist. 36 (2), 587–613.
Koenker, R., 2005. Quantile Regression. Cambridge University Press.
Koenker, R., Bassett, G., 1978. Regression quantiles. Econometrica 46 (1), 33–50.
Leng, C., Lin, Y., Wahba, G., 2006. A note on the lasso and related procedures in model selection. Statist. Sinica 16 (4), 1273–1284.
Meinshausen, N., Bühlmann, P., 2006. High-dimensional graphs and variable selection with the lasso. Ann. Statist. 34 (3), 1436–1462.
Shao, J., 1993. Linear model selection by cross-validation. J. Amer. Statist. Assoc. 88 (422), 486–494.
Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B 58 (1), 267–288.
Wang, L., Kim, Y., Li, R., 2013. Calibrating nonconvex penalized regression in ultra-high dimension. Ann. Statist. 41 (5), 2505–2536.
Wang, L., Wu, Y., Li, R., 2012. Quantile regression for analyzing heterogeneity in ultra-highdimension. J. Amer. Statist. Assoc. 107 (497), 214–222.
Wu, Y., Liu, Y., 2009. Variable selection in quantile regression. Statist. Sinica 19 (2), 801–817.
Zhang, P., 1993. Model selection via multifold cross validation. Ann. Statist. 21 (1), 299–313.
Zhang, C.-H., Zhang, T., 2012. A general theory of concave regularization for high-dimensional sparse estimation problems. Statist. Sci. 27 (4), 576–593.
Zou, H., 2006. The adaptive lasso and its oracle properties. J. Amer. Statist. Assoc. 101 (476), 1418–1429.