# ESTIMATING AND SELECTING VARIABLES FOR A CONDITIONAL QUANTILE REGRESSION MODEL USING PENALTY FUNCTIONS

Mohammed Ibrahim Zamel

Dhi-Qar Education Directorate, Iraq
*Corresponding author, email: mohamedz1402@uowasit.edu.iq

doi: 10.17977/um066.v6.i6.2026.1

**Keywords**

Quantile regression
Penalized quantile regression
Variable Selection

**Abstract**

Penalised quantile regression is an effective approach for variable selection and parameter estimation, particularly when dealing with high-dimensional data containing a large number of explanatory variables. In such situations, some explanatory variables may have little or no effect on the response variable, leading to overly complex models that are difficult to interpret and may reduce estimation efficiency. To address this issue, penalisation techniques are incorporated into the quantile regression framework to simultaneously perform parameter estimation and variable selection. This study focuses on comparing two widely used penalised quantile regression estimators, namely the Least Absolute Shrinkage and Selection Operator (LASSO) and the Minimax Concave Penalty (MCP). The comparison is conducted through Monte Carlo simulation experiments implemented in the R programming environment under different sample sizes, model dimensions, and quantile levels. Explanatory variables are generated from a multivariate normal distribution, while the response variable is constructed based on the generated predictors and random error terms. The performance of the estimators is evaluated using the Mean Squared Error (MSE) criterion to assess estimation accuracy. In addition, variable selection performance is examined through the False Positive Rate (FPR) and False Negative Rate (FNR), which measure the ability of each estimator to correctly identify relevant and irrelevant variables. The simulation results indicate that the MCP estimator generally outperforms the LASSO estimator across most experimental scenarios. Specifically, MCP produces lower MSE values, indicating superior estimation accuracy, while also achieving lower FPR and FNR values, demonstrating greater effectiveness in selecting the true explanatory variables. These findings suggest that MCP is a more reliable and efficient method for estimation and variable selection in penalised quantile regression models.

## 1. Introduction

Linear regression is often inapplicable in various scientific studies due to non-compliance with regression conditions. A viable alternative is Quantile Regression (QR), introduced by Koenker and Bassett in 1978, which has garnered increased interest in recent years. Assuming we possess a random sample$(Y_1, X_1), \dots, (Y_n, X_n)$ ، $(0 < \theta < 1)$ The linear regression model can be articulated as follows (Amin et al., 2015).

$$\mathbf{Y_i = \acute{X}_i \beta + \epsilon_i} \qquad \mathbf{i = 1, 2, \dots, n} \qquad (\mathbf{1})$$

Yi represents the response variable

Xi represents explanatory variables (independent variables).

β means model parameters, $\beta = (\beta_1, \beta_2, \dots, \beta_p)$ and its dimensions p .

$\epsilon_i$ = Represents a random error that is normally distributed with an average

zero and variance $\sigma^2$

$\sigma^2)\ \ \epsilon_i \sim N(0\ ,$

Therefore, the parameters of the quantile regression model $\beta = (\beta_1, \beta_2, \dots, \beta_p) \in \mathcal{R}^P$ can be estimated using the following formula:

$$\widehat{\boldsymbol{\beta}}_{(\boldsymbol{\theta})} = \mathbf{arg\,min}_{\boldsymbol{\beta}\in\mathbb{R}^{\mathbf{p}}} \sum_{\mathbf{i=1}}^{\mathbf{n}} \boldsymbol{\rho}_{\boldsymbol{\theta}}\left(\mathbf{y_i} - \acute{\mathbf{X}}_{\mathbf{i}}\boldsymbol{\beta}\right) \qquad (\mathbf{2})$$

$\rho_\theta(t)$ It is called the tuning function and is expressed in the following form:

$$\boldsymbol{\rho}_{\boldsymbol{\theta}}(\mathbf{t}) = \begin{cases} \boldsymbol{\theta}\mathbf{t} & \mathbf{if}\ \ \mathbf{t} \geq \mathbf{0} \\ -(\mathbf{1}-\boldsymbol{\theta})\mathbf{t} & \mathbf{if}\ \ \mathbf{t}\ \geq \mathbf{0} \end{cases} \qquad (\mathbf{3})$$

$( 0 < \theta < 1 )$

The selection of factors is crucial when numerous explanatory variables are present. The influence of independent variables on the response variable remains uncertain, given the presence of numerous explanatory variables that exert minimal or negligible effects. The primary objective is to achieve a parsimonious model, characterised by a limited number of explanatory variables, referred to as a sparse model.

Identifying relevant variables can be challenging, potentially resulting in the omission of crucial explanatory factors. These issues necessitate proper and optimal resolution. Penalised quantile regression serves as an effective instrument to tackle these challenges (Fan & Li, 2001).

Wu and Liu (2009) examined penalised divisional regression utilising the SCAD and ADAPTIVE LASSO estimators through the DCA algorithm.

The findings from both real and simulated data corroborated that the employed estimators possess Oracle properties (Wu & Liu, 2009), and Wang and Li (2012) examined Penalised quantile regression in the context of high-dimensional data. The data frequently encounter heterogeneity issues due to the proliferation of explanatory variables, and a penalty function (SCAD) was utilised to identify the relevant variables. In linear regression, estimating the conditional distribution of the response variable at various points is unfeasible; therefore, quantile regression is employed. However, it is not applicable when the number of explanatory variables is extensive. To mitigate this issue, Penalised quantile regression is utilised. This research seeks to compare the Lasso and MCP estimators of the quantile regression model to identify the most accurate estimate, utilising simulations based on mean squared error, false positive rate, and false negative rate standards.

## 1.1. The Importance of Quantile Regression

Quantile regression analyzes the relationship between one or more explanatory variables and the response variable across different points or regions. In contrast, conventional regression based on the (OLS) method focuses on estimating the conditional mean of the response variable at given values of the explanatory variables.

Quantile regression works on minimize the sum of squared errors while also reducing the mean absolute deviation, and it estimates conditional piecewise functions by minimizing the sum of absolute errors. This approach addresses several limitations commonly associated with the ordinary least squares (OLS) method, which is often highly sensitive to outliers that can substantially distort the results. In contrast, segmented regression estimators are more robust to extreme observations, as this robustness arises from the minimization of a piecewise loss function. Within this framework, quantile regression further enhances robustness by allowing the estimation of different parts of the conditional distribution rather than focusing solely on the mean. in general, the concept of univariate quantile regression can be extended to conditional segmentations by considering one or more covariates within the quantile regression framework. Suppose that we have a random variable (Y) with a probability distribution function, as expressed in the following form (Zhang et al., 2010; Hao, Naiman, & Naiman, 2007):

$$F(y|X = x) = pr(Y \leq y|X = x)$$

Quantile regression offers several important advantages. First, the use of quantile regression yields values of the dependent variable that are closer to the original observed values. In addition, it can be applied when the assumptions of standard linear regression are not satisfied, particularly in situations where the random errors are not normally distributed. Furthermore, quantile regression provides more accurate and more efficient information than ordinary regression. Another important advantage is its ability to measure the relationship between the dependent variable and the explanatory variables at different points of the conditional distribution. Moreover, it provides a comprehensive and precise explanation of the relationship between the response variable Y and the explanatory variables X, offering a clear and overall picture through fitting more than one regression line in modeling the conditional distribution $Y|X = \{X_1, X_2, \dots, X_P\}$ in different segments (Wu & Liu, 2009).

## 1.2. Penalized Quantile Regression

The quantity of explanatory variables rises while the sample size diminishes, the estimation of the model's parameters becomes challenging to execute. Selecting relevant variables is a considerable challenge, representing a critical issue in statistical modelling. In this instance, quantile regression is inapplicable; therefore, the suitable option is to employ a regression approach designed for high-dimensional data, specifically the Penalised Quantile Regression method (Yousif & Housain, 2021)

The penalty function is incorporated into the divisional loss function, resulting in the Objective function as expressed in the following formula (Li & Zhu, 2008):

$$\mathbf{Q_T(\beta) = \sum_{i=1}^{n} \rho_\theta(y_i - \acute{X}_i\beta) + \lambda \sum_{j=1}^{p} p_\lambda(|\beta_j|)} \qquad \mathbf{(4)}$$

$p_\lambda(.)$ :Penalty function

$\lambda$ : penalty parameter and its value $\lambda \geq 0$ ،it is called the tuning parameter.

## 1.3. Least absolute shrinkage and selection operator (Lasso)

The one of the penal estimation methods for quantile regression and was proposed by Tibshirani (1996), which works on estimating and selecting the variable (Variable Selection) at the same time. The basic principle of this method is to minimize the Sum of squared errors according to the constraint that represents the absolute sum of the parameters.

$$\mathbf{\sum_{j=1}^{p} |\beta_j|} \qquad \mathbf{(5)}$$

The (lasso) function makes some parameters exactly equal to zero when the penalty parameter is large, and the other parameters are reduced according to a certain amount, and the penalty parameter (λ) is used to control the amount of shrinkage. Li and Zhu (2008) proposed a penalty function (lasso) in the quantile regression (QR) by adding the penalty function (Lasso) to the Partition loss function to obtain the Partition target function.

From this we obtain the (lasso) estimators $\hat{\beta}_{\text{LASSO}} = (\hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_P)$ in quantile regression (QR) According to the following formula (Li & Zhu, 2008)

$$\mathbf{\hat{\beta}_{(LASSO,QR)} = \arg\min_\beta \left\{ \sum_{i=1}^{n} \rho_\theta(y_i - \acute{X}_i\beta) + \lambda \sum_{j=1}^{p} |\beta_j| \right\}} \qquad \mathbf{(6)}$$

λ: Represents the penalty paramete

also called (tuning parameter), It works by controlling and prioritizing between the partition loss function and the penalty limits, i.e., it controls the Size parameters.

$\lambda p(.)$: Represents the penalty function.

to calculate the Lasso estimator, we use the (SNCD) algorithm (Yi & Huang, 2017).

### 1.4. Minimax Convex Penalty (MCP)

The concave penalty function proposed by Zhang (2010) is distinguished by its rapid parameter estimation, continuity, unbiasedness, and precision. Furthermore, it simultaneously simplifies the model and enhances its accuracy, thereby outperforming the Lasso method in both estimation speed and accuracy. The MCP method can be written as follows:

$$\mathbf{p_{\lambda,\gamma}}(|\boldsymbol{\beta}|) = \begin{cases} \boldsymbol{\lambda}|\boldsymbol{\beta}| - \frac{\beta^2}{2\gamma} & |\boldsymbol{\beta}| \leq \boldsymbol{\gamma\lambda} \\ \frac{1}{2}\boldsymbol{\gamma\lambda^2} & |\boldsymbol{\beta}| > \boldsymbol{\gamma\lambda} \end{cases} \qquad \mathbf{(7)}$$

$\gamma > 1$

The objective function for the Penalised Quantile Regression of a Minimum Concave Penalty (MCP) estimator can be articulated as follows (Wang, Wu, & Li, 2012):

$$\widehat{\boldsymbol{\beta}}_{\mathbf{MCP,QR}} = \mathbf{argmin}_{\boldsymbol{\beta}}\left\{\sum_{\mathbf{i=1}}^{\mathbf{n}} \boldsymbol{\rho_\theta}(\mathbf{y_i} - \mathbf{X_i^T}\boldsymbol{\beta}) + \boldsymbol{\lambda}\sum_{\mathbf{j=1}}^{\mathbf{p}} \mathbf{p_{\lambda\gamma}}(\boldsymbol{\beta})\right\} \qquad (8)$$

## (Comparison criteria)

There are several criteria that can measure efficiency when estimating regression functions, and the following criteria have been used:

## criterion (MSE)

It represents the mean squared error and is given as follows Hodson, Over & Foks, 2021):

$$\mathbf{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{g}(\beta_1^T X \dots \beta_d^T X)\right)^2 \qquad \mathbf{(9)}$$

$\hat{g}(\beta_1^T X \dots \beta_d^T X)$: represents the estimated value of the link function

n: Represents the number of observations.

The model that gives the least value to the MSE is best.

## (FPR) and (FNR) criterion

The calculated models are assessed for dispersion using the False Positive Rate (FPR) and the False Negative Rate (FNR). The false positive rate refers to coefficients that are zero in the actual model but non-zero in the estimation, whereas the false negative rate pertains to coefficients that are non-zero in the actual model but zero in the estimation (Zhang et al., 2010).

$$\mathbf{FPR}(\widehat{\boldsymbol{\beta}}) = \frac{|\{\mathbf{j} \in \{\mathbf{1}, \dots, \mathbf{p}\}: \widehat{\boldsymbol{\beta}}_\mathbf{j} \neq \mathbf{0} \cap \boldsymbol{\beta}_\mathbf{j} = \mathbf{0}\}|}{|\{\mathbf{j} \in \{\mathbf{1}, \dots, \mathbf{p}\}: \boldsymbol{\beta}_\mathbf{j} = \mathbf{0}\}|} \qquad (\mathbf{10})$$

$$\mathbf{FNR}(\widehat{\boldsymbol{\beta}}) = \frac{|\{\mathbf{j} \in \{\mathbf{1}, \dots, \mathbf{p}\}: \widehat{\boldsymbol{\beta}}_\mathbf{j} = \mathbf{0} \cap \boldsymbol{\beta}_\mathbf{j} \neq \mathbf{0}\}|}{|\{\mathbf{j} \in \{\mathbf{1}, \dots, \mathbf{p}\}: \boldsymbol{\beta}_\mathbf{j} \neq \mathbf{0}\}|} \qquad (\mathbf{11})$$

## 2. Methods

This study employed a Monte Carlo simulation approach to compare the performance of penalized quantile regression estimators, namely LASSO and MCP. The simulation experiments were conducted using the R programming language. The objective was to evaluate the estimation accuracy and variable selection capability of both estimators under different sample sizes and model dimensions.

The quantile regression model was estimated using penalized techniques to address high-dimensional data problems. Two penalty functions were considered: the Least Absolute Shrinkage and Selection Operator (LASSO) and the Minimax Concave Penalty (MCP). These methods simultaneously perform parameter estimation and variable selection by shrinking irrelevant coefficients toward zero.

The explanatory variables were generated from a multivariate normal distribution, while the error terms followed a normal distribution with mean zero. The response variable was obtained from the linear combination of explanatory variables and random errors. Two simulation scenarios were considered: Model 1 (P = 10, n = 25) and Model 2 (P = 25, n = 100). The estimators were evaluated using Mean Squared Error (MSE) for estimation performance and False Positive Rate (FPR) and False Negative Rate (FNR) for variable selection performance.

### 2.1. Simulation

The simulation method helps in selecting different random sample sizes as well as the ability to the importance of the simulation comes in choose various values for random errors. There are different randomness, as the numbers used are independent in different experiments. Methods of simulation, including the analog method, the mixed procedure and the Monte Carlo procedure.

Which is one of the most important and famous and most used simulation methods it relies on generating random samples from a hypothetical theoretical community that is similar to the real population and generating observations of most probabilistic distributions. The Monte Carlo method can be defined as "a numerical method used to process experiments Using a computer that includes multiple types of logical and mathematical relations to describe the behavior and structure of a real system" (Ingalls, 2011)

To compare the penal estimators (Lasso, MCP), simulation experiments are conducted utilising the Monte Carlo approach and rely on the R programming language, as detailed below:

### 2.2. Generate explanatory variables

We depend on the normal distribution of multiple variables with an average of (mean=0) and variance (Σ) to generate the explanatory variables P as follows:

$$\mathbf{X} \sim \mathbf{MN}(\mathbf{0}, \boldsymbol{\Sigma})$$
$$\boldsymbol{\Sigma}_{\mathbf{ij}} = \boldsymbol{\rho}^{|\mathbf{i}-\mathbf{j}|} \quad , \quad \boldsymbol{\rho} = \mathbf{0.5}$$

### 2.3. Generate random errors

The process of generating random errors is done according to the natural distribution with an average of (mean=0) and a variance ( $\sigma^2$).

$$\boldsymbol{\varepsilon}_{\mathbf{i}} \sim \mathbf{N}(\mathbf{0}, \boldsymbol{\sigma}^{\mathbf{2}}) \quad , \mathbf{i} = \mathbf{1}, \mathbf{2}, \dots, \mathbf{n}$$

### 2.4. dependent variable

The dependent variable can be calculated by multiplying the previously generated X explanatory variable matrix by the default parameter vector plus the prior random error limit ( $\varepsilon_i$).

The simulation process is carried out according to the following method:

σ: it takes the two values ( 0.5 ,1 )

We depend on:

the first model: $\beta = (3,1.5,0,2,0, \dots ,0)$ , (P=10, n=25)

the second model: $\beta = (1,1,1,-2,-2,-2,0, \dots ,0)$, $(P = 25 , n = 100)$

as for the default values (Quantile), they are $\theta = (0.3,0.5,0.7)$

The comparison between the estimators is based on the Mean Squared Errors (MSE) criterion, through which the estimators can be compared in terms of estimation. As for the selection of variables, the False Positive Rate (FNR) and False Negative Rate (FPR) criteria are used.

## 3. Results and Discussion

The results of the simulation are as shown in the following tables:

**Table 1. Simulation results for the first model when (P=10, n=25)**

| θ | σ | Estimators | MSE | FPR | FNR |
|---|---|---|---|---|---|
| 0.3 | 0.5 | LASSO | 0.0326 | 0.131 | 0 |
| | | MCP | 0.0228 | 0.005 | 0 |
| | 1 | LASSO | 0.5695 | 0.302 | 0.106 |
| | | MCP | 0.4368 | 0.057 | 0.24 |
| 0.5 | 0.5 | LASSO | 0.0633 | 0.148 | 0 |
| | | MCP | 0.0204 | 0.005 | 0 |
| | 1 | LASSO | 0.2154 | 0.24 | 0 |
| | | MCP | **0.1284** | **0.028** | **0.013** |
| 0.7 | 0.5 | LASSO | 0.0403 | 0.006 | 0 |
| | | MCP | 0.0401 | 0 | 0 |
| | 1 | LASSO | 0.0931 | 0.194 | 0 |
| | | MCP | 0.0399 | 0.057 | 0 |

**Table 2. Simulation results for the second model when (P=25, n=100)**

| θ | σ | Estimators | MSE | FPR | FNR |
|---|---|---|---|---|---|
| 0.3 | 0.5 | LASSO | 0.0066 | 0.167 | 0 |
| | | MCP | 0.0015 | 0.14 | 0 |
| | 1 | LASSO | 0.0508 | 0.16 | 0 |
| | | MCP | 0.0367 | 0.136 | 0.086 |
| 0.5 | 0.5 | LASSO | 0.0123 | 0.174 | 0 |
| | | MCP | 0.0022 | 0.136 | 0 |
| | 1 | LASSO | 0.0911 | 0.1454 | 0.08 |
| | | MCP | 0.1049 | 0.133 | 0.229 |
| 0.7 | 0.5 | LASSO | 0.0849 | 0.152 | 0 |
| | | MCP | 0.0671 | 0.136 | 0 |
| | 1 | LASSO | 0.0614 | 0.165 | 0.02 |
| | | MCP | 0.0308 | 0.136 | 0.066 |

The simulation results presented in Tables 1 and 2 demonstrate the comparative performance of the LASSO and MCP estimators under different sample sizes, model dimensions, and quantile levels. The comparison was conducted using the Mean Squared Error (MSE) criterion to evaluate estimation accuracy and the False Positive Rate (FPR) and False Negative Rate (FNR) criteria to assess variable selection performance. For the first model (P = 10, n = 25), the MCP estimator consistently produced lower MSE values than the LASSO estimator across most quantile settings, indicating superior estimation accuracy. Similarly, for the second model (P = 25, n = 100), MCP generally achieved smaller estimation errors and exhibited better stability as the dimensionality of the explanatory variables increased. These findings suggest that MCP is more capable of producing accurate parameter estimates, even in situations where the number of explanatory variables is relatively large.

Regarding variable selection, the simulation results indicate that MCP outperformed LASSO in identifying the true underlying model. In most scenarios, MCP achieved lower FPR values, implying

a reduced tendency to incorrectly retain irrelevant explanatory variables, while also maintaining relatively low FNR values, reflecting its ability to preserve important variables in the model. The superior performance of MCP can be attributed to its concave penalty structure, which reduces estimation bias while preserving sparsity. In contrast, the LASSO estimator tends to shrink all coefficients uniformly, which may lead to biased estimates and less accurate variable selection when the penalty parameter increases. Overall, the results confirm that the MCP estimator provides a more efficient balance between estimation accuracy and variable selection effectiveness, making it a preferable choice for penalized quantile regression models involving high-dimensional data.

## 4. Conclusions

The findings of this study indicate that the MCP estimator exhibits superior performance compared with the LASSO estimator in penalized quantile regression models. Across the simulation scenarios considered, MCP consistently achieved lower Mean Squared Error (MSE) values, demonstrating greater estimation accuracy and stability. Furthermore, the MCP estimator proved to be more effective in variable selection by producing lower False Positive Rate (FPR) and False Negative Rate (FNR) values, thereby improving its ability to correctly identify relevant explanatory variables while excluding irrelevant ones. This superior performance can be attributed to the ability of the MCP penalty function to reduce estimation bias while maintaining model sparsity. Overall, the simulation results suggest that MCP provides a more efficient balance between accurate parameter estimation and reliable variable selection, making it a suitable and robust approach for analyzing high-dimensional data within the quantile regression framework.

## References


Amin, M., Song, L., Thorlie, M. A., & Wang, X. (2015). SCAD-penalized quantile regression for high-dimensional data analysis and variable selection. *Statistica Neerlandica, 69*(3), 212–235.

Buhai, S. (2005). Quantile regression: Overview and selected applications. *Ad Astra, 4*(4), 1–17.

Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association, 96*(456), 1348–1360.

Hao, L., & Naiman, D. Q. (2007). *Quantile regression* (Vol. 149). Thousand Oaks, CA: Sage Publications.

Hodson, T. O., Over, T. M., & Foks, S. S. (2021). Mean squared error, deconstructed. *Journal of Advances in Modeling Earth Systems, 13*(12), e2021MS002681.

Ingalls, R. G. (2011). Introduction to simulation. In *Proceedings of the 2011 Winter Simulation Conference* (pp. 1374–1388). Piscataway, NJ: IEEE.

Koenker, R., & Bassett, G. (1978). Regression quantiles. *Econometrica, 46*(1), 33–50.

Li, Y., & Zhu, J. (2008). L1-norm quantile regression. *Journal of Computational and Graphical Statistics, 17*(1), 163–185.

Tibshirani, R. J. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological), 58*(1), 267–288.

Wang, L., Wu, Y., & Li, R. (2012). Quantile regression for analyzing heterogeneity in ultra-high dimension. *Journal of the American Statistical Association, 107*(497), 214–222.

Wu, Y., & Liu, Y. (2009). Variable selection in quantile regression. *Statistica Sinica, 19*(2), 801–817.

Yi, C., & Huang, J. (2017). Semismooth Newton coordinate descent algorithm for elastic-net penalized Huber loss regression and quantile regression. *Journal of Computational and Graphical Statistics, 26*(3), 547–557.

Yousif, A. H., & Housain, W. J. (2021). Atan regularized in quantile regression for high dimensional data. In *Journal of Physics: Conference Series* (Vol. 1818, No. 1, Article 012098). Bristol, England: IOP Publishing.

Zhang, C. H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics, 38*(2), 894–942.

Zhang, J. M., Shi, Q. C., Xu, F. Z., Fu, Y. L., Wang, S. M., Gu, W., ... & Hu, W. P. (2010). False positive rate and false negative rate of the 12-item General Health Questionnaire and related factors. *Chinese Mental Health Journal*.