Al-Qadisiyah Journal for Administrative and Economic Sciences ISSN$_{Online}$: 2312-9883
QJAE, Volume 28, Issue 2 (2026) ISSN$_{Print}$: 1816-9171
DOI:/ /doi.org/10.33916/qjae.2026.02485497

Identifying Heart Disease Risk Factors Via SCAD-Penalized Quantile Regression

Zainab s. Alsaadi
Department of Studies and Planning , Presidency of the University
,University of Al-Qadisiyah , Al-Qadisiyah ,Iraq
Zainab.s.alsaadi@qu.edu.iq

Received: 03/03/2026
Accepted: 03/06/2026
Available online: 30/06/2026
Corresponding Author: Zainab S. AlSaadi

Abstract*:* This study examines heart disease risk factors using SCAD-penalized quantile regression, aiming to capture heterogeneous covariate effects across different levels of disease severity while achieving effective variable selection. Unlike mean-based regression models, the proposed approach allows regression coefficients to vary across quantiles of the response distribution, providing a detailed characterization of how clinical predictors influence mild, moderate, and severe forms of heart disease. The methodological framework integrates quantile regression with the smoothly clipped absolute deviation penalty and is estimated through a local linear approximation algorithm. The analysis is conducted on real clinical data obtained from a publicly available heart disease dataset. The empirical results reveal pronounced distributional heterogeneity in disease severity and highlight clear quantile-dependent patterns. Age and ST depression show consistently positive and increasing effects toward higher quantiles, indicating stronger associations among patients with severe disease, while maximum heart rate exhibits a stable protective effect across all quantiles. Other predictors, including resting blood pressure, serum cholesterol, and exercise-induced angina, emerge mainly at upper quantiles, suggesting their relevance primarily for severe cases. Overall, the findings demonstrate that SCAD-penalized quantile regression provides a robust and interpretable framework for identifying clinically meaningful heart disease risk factors and uncovering heterogeneity that is not detectable using conventional regression methods.

Keywords: Quantile Regression , SCAD Penalty , Heart Disease Risk Factors , Variable Selection , Heterogeneous Effects .

1. **Introduction**

Cardiovascular disease remains a leading cause of morbidity and mortality worldwide, and early identification of modifiable risk factors is essential for effective prevention and targeted clinical intervention. Heart disease outcomes often exhibit substantial heterogeneity across patients, where the influence of a clinical predictor may differ between mild and severe disease profiles. Conventional mean-based regression models summarize covariate effects at a single central tendency and therefore may obscure important distributional differences. Quantile regression overcomes this limitation by modeling conditional quantiles of the response variable, allowing covariate effects to vary across different parts of the outcome distribution (Koenker and Bassett, 1978).

In recent years, heart disease datasets have become increasingly complex, typically involving correlated predictors and a mixture of continuous and categorical clinical indicators. Such characteristics pose challenges for classical estimation techniques and increase the risk of overfitting. Penalized quantile regression provides an effective framework for addressing these issues by combining robustness to outliers with regularization-based variable selection. Among nonconvex penalties, the Smoothly Clipped Absolute Deviation penalty has received considerable attention due to its ability to achieve sparsity while reducing estimation bias for large coefficients (Fan and Li, 2001; Wu and Liu, 2009).

The present study focuses on identifying heart disease risk factors using SCAD-penalized quantile regression, with particular emphasis on quantile-dependent effects across the severity distribution. This direction is motivated by our previous work on heart disease analysis using logistic regression models, which identified key determinants of disease incidence but relied on mean-based inference (Mohammed et al., 2020). In addition, our earlier contributions on penalized and composite quantile regression methods demonstrated the advantages of regularization-based variable selection for handling correlated predictors and heterogeneous biomedical data (Raheem et al., 2025).

Building on these foundations, this paper applies a SCAD-penalized quantile regression model to real heart disease data in order to capture heterogeneous risk patterns and obtain sparse, interpretable estimates. The proposed approach employs a local linear approximation strategy for efficient estimation and is evaluated using real clinical data. The results provide insight into how the effects of key risk factors vary across different levels of heart disease severity, offering a flexible and informative alternative to traditional regression methods.

2. Theoretical Background

2.1 Quantile Regression

Quantile regression was first introduced by Koenker and Bassett (1978) as a generalization of the classical linear regression model. Unlike ordinary least squares regression, which focuses on modeling the conditional mean of the response variable, quantile regression estimates conditional quantiles, allowing a more comprehensive analysis of the entire conditional distribution of the outcome.

Let Y denote the response variable and X represent the vector of explanatory variables. The conditional $\tau - th$ quantile of Y given X is defined as:

$$\mathrm{Q_Y}(\tau \mid \mathrm{X}) \;=\; \mathrm{X^T}\,\beta(\tau)$$

The estimation of $\beta(\tau)$ is obtained by solving the following optimization problem (Koenker and Bassett, 1978):

$$\hat{\beta}(\tau) = arg \min_{\beta} \sum_{i=1}^{n} \rho_{\tau} (y_i - x_i^{T}\beta)$$

where $\rho_{\tau}(u)$ is the asymmetric check loss function defined as follows:

$$\rho_{\tau}(u) = \tau u , \quad if\ u \geq 0;\ (\tau - 1)u , \quad if\ u < 0$$

This loss function assigns unequal weights to positive and negative residuals, enabling the estimation of different conditional quantiles of the response variable. When $\tau = 0.5$, the method reduces to median regression, which minimizes the sum of absolute deviations and is robust to outliers (Koenker, 2005).

One of the main advantages of quantile regression is that it does not require strict assumptions regarding the distribution of the error term. This makes it particularly suitable for medical and epidemiological data, which often exhibit skewness or heteroscedasticity (Koenker and Hallock, 2001). Moreover, quantile regression allows regression coefficients to vary across quantiles, capturing heterogeneous effects that cannot be identified using mean-based methods.

In the context of heart disease analysis, risk factors may influence patients differently depending on disease severity. Quantile regression enables the identification of such differential effects by estimating separate models at various quantile levels, thereby providing clinically interpretable insights (Davino et al., 2013).

2.2 Penalized Quantile Regression

Although quantile regression provides a flexible and robust framework for modeling heterogeneous effects across the conditional distribution of the response variable, it faces important challenges when the number of predictors is large or when strong correlations exist among covariates. In such situations, classical quantile regression estimators may suffer from instability, inflated variance, and poor interpretability due to overfitting (Koenker, 2005).

Penalized quantile regression was introduced to address these limitations by incorporating regularization techniques into the quantile regression framework. The main idea is to augment the quantile regression objective function with a penalty term that controls the complexity of the model and enforces shrinkage on the regression coefficients (Tibshirani, 1996; Koenker, 2005).

Let y_i denote the response variable and x_i the corresponding vector of predictors. The penalized quantile regression estimator is defined as the solution to the following optimization problem:

$$\hat{\beta}(\tau) = arg \min_{\beta} \sum_{i=1}^{n} \rho_\tau (y_i - x_i^{\mathrm{T}}\beta) + \sum_{j=1}^{p} P_\lambda (|\beta_j|) \}$$

where $\rho_\tau(\cdot)$ is the quantile check loss function and $P_\lambda(\cdot)$ denotes a penalty function indexed by a tuning parameter $\lambda > 0$.

The penalty term plays a crucial role in controlling model complexity. By shrinking regression coefficients toward zero, penalized quantile regression reduces estimation variance and mitigates the impact of multicollinearity. Moreover, certain penalty functions are capable of setting some coefficients exactly to zero, thereby performing automatic variable selection (Fan and Li, 2001).

Several penalty functions have been proposed in the literature for quantile regression. The LASSO penalty encourages sparsity but may introduce bias in the estimation of large coefficients (Tibshirani, 1996). Ridge-type penalties improve numerical stability but do not perform variable selection. These limitations have motivated the development of nonconvex penalties that aim to balance sparsity and unbiased estimation (Fan and Li, 2001).

Penalized quantile regression is particularly attractive in medical studies, where datasets often contain many correlated variables and only a subset of predictors is truly relevant. In heart disease analysis, regularization is essential for obtaining stable and interpretable quantile-specific models.

By integrating regularization directly into the quantile regression framework, penalized quantile regression enables effective variable selection while preserving the ability to study heterogeneous effects across the outcome distribution. This framework provides the basis for advanced penalty functions such as the smoothly clipped absolute deviation penalty.

2.3 SCAD Penalty Function

Standard penalization methods such as LASSO and Ridge regression have been widely used to address high-dimensionality and multicollinearity in regression models. However, these methods suffer from important limitations. LASSO tends to introduce bias in the estimation of large coefficients due to its constant shrinkage, while Ridge regression does not perform variable selection because it cannot shrink coefficients exactly to zero (Tibshirani, 1996; Hoerl and Kennard, 1970).

To overcome these drawbacks, Fan and Li (2001) proposed the Smoothly Clipped Absolute Deviation (SCAD) penalty, a nonconvex penalty function designed to achieve three desirable properties simultaneously: sparsity, unbiasedness for large coefficients, and continuity of the resulting estimator.

Let β_j denote the regression coefficient associated with the $j-th$ predictor. The SCAD penalty function $P_\lambda\ (|\beta_j|)$ is defined through its first derivative as follows:

$$P'_\lambda(|\beta_j|) = \begin{cases} \lambda & for\ |\beta_j| \le \lambda; \\ \dfrac{(a\lambda - |\beta_j|)}{(a-1)} & for\ \lambda < |\beta_j| \le a\lambda; \\ 0 & for\ |\beta_j| > a\lambda \end{cases}$$

where $\lambda > 0$ is the tuning parameter controlling the amount of penalization and $a > 2$ is a fixed constant, commonly set to $a = 3.7$ as suggested by Fan and Li (2001).

This formulation implies that small coefficients are heavily penalized, encouraging sparsity by shrinking them toward zero, while large coefficients receive little or no penalization. As a result, SCAD reduces estimation bias for truly influential predictors.

An important theoretical advantage of the SCAD penalty is that it satisfies the oracle property. Under suitable regularity conditions, the SCAD estimator can correctly identify the true subset of nonzero coefficients and estimate the nonzero coefficients as efficiently as if the true model were known in advance (Fan and Li, 2001).

In the context of quantile regression, the SCAD penalty is incorporated into the objective function to control model complexity while preserving quantile-specific estimation. This results in effective variable selection with reduced bias for large effects (Wu and Liu, 2009).

For medical applications such as heart disease analysis, where predictors are often correlated and only a subset of risk factors is clinically relevant, SCAD provides a balanced regularization strategy that enhances interpretability and estimation stability.

2.4 SCAD-Penalized Quantile Regression Model

The SCAD-penalized quantile regression model extends classical quantile regression by explicitly incorporating the smoothly clipped absolute deviation penalty to achieve robust estimation and effective variable selection. This framework is designed to capture heterogeneous covariate effects across different parts of the conditional distribution of the response while avoiding the bias commonly introduced by convex penalties.

Let y_i denote the response variable and x_i the vector of predictors for observation $i = 1, \dots, n$. For a given quantile level $\tau \in (0,1)$, the conditional quantile of Y given X is modeled as:

$$Q_Y(\tau \mid X) = X^{\mathrm{T}}\beta(\tau)$$

where $\beta(\tau)$ is the vector of quantile-specific regression coefficients.

The SCAD-penalized quantile regression estimator is obtained by solving the following optimization problem:

$$\hat{\beta}(\tau) = arg \min_{\beta} \sum_{i=1}^{n} \rho_{\tau} (y_i - x_i^{\top}\beta) + \sum_{j=1}^{p} P_{\lambda}^{SCAD}(|\beta_j|)$$

where $\rho_{\tau}(\cdot)$ is the asymmetric quantile check loss function and $P_{\lambda}^{SCAD}(\cdot)$ denotes the smoothly clipped absolute deviation penalty proposed by Fan and Li (2001), with tuning parameter $\lambda > 0$.

The SCAD penalty applies adaptive shrinkage depending on the magnitude of the regression coefficients. Small coefficients are heavily penalized and may be shrunk exactly to zero, enabling variable selection, while large coefficients are penalized minimally, reducing estimation bias. This property distinguishes SCAD from LASSO-type penalties, which impose constant shrinkage and tend to bias large effects.

Due to the nonconvex nature of the SCAD penalty, direct minimization of the objective function is computationally challenging. In practice, estimation is commonly performed using iterative algorithms such as the local linear approximation method, which replaces the SCAD penalty with a sequence of weighted convex penalties and updates the coefficient estimates until convergence (Fan and Li, 2001; Wu and Liu, 2009).

The resulting SCAD-penalized quantile regression model produces sparse and stable coefficient estimates that vary across quantiles. This allows the identification of predictors whose influence differs between lower, median, and upper portions of the response distribution. In medical applications such as heart disease analysis, this property is particularly valuable for uncovering quantile-dependent risk factors and providing interpretable models that reflect heterogeneous clinical profiles.

2.5 Estimation Algorithms for SCAD Quantile Regression

This section presents the estimation algorithm used for SCAD-penalized quantile regression. The algorithm is based on the Local Linear Approximation approach, which transforms the nonconvex SCAD optimization problem into a sequence of convex weighted quantile regression problems.

Input

Observed data $\{(y_i, x_i)\}_{i=1}^{n}$, where $x_i \in R^p$

Quantile level $\tau \in (0,1)$

Penalty parameter $\lambda > 0$

SCAD constant $a > 2$ (commonly $a = 3.7$)

Tolerance $\varepsilon > 0$

Maximum number of iterations K_{max}

Definition 1: Check Loss Function

For any $u \in R$, define:

$$\rho_\tau(u) = u\left(\tau - I(u < 0)\right) = \begin{cases} \tau u & , if\ u \geq 0 \\ (\tau - 1)u, & if\ u < 0 \end{cases}$$

Definition 2: SCAD Penalty Derivative

For $t \geq 0$, define:

$$P_\lambda'^{SCAD}(t) = \begin{cases} \lambda & for\ 0 \leq t \leq \lambda; \\ \frac{a\lambda - t}{a - 1} & for\ \lambda < t \leq a\lambda \\ 0 & for\ t > a\lambda \end{cases}$$

Objective

$$\hat{\beta}(\tau) = arg\ \min_\beta \sum_{i=1}^{n} \rho_\tau\left(y_i - x_i^{\top}\beta\right) + \sum_{j=1}^{p} P_\lambda^{SCAD}(|\beta_j|)$$

Step 1: Initialization

Set an initial estimate $\beta^{(0)}$. A standard choice is the unpenalized quantile regression estimator.

Step 2: Iterative Local Linear Approximation

For $= 0,1,2, \ldots, K_{max} - 1$, repeat the following steps.

Step 2.1: Weight Computation

For each coefficient j = 1,…,p, compute the weight:

$$w_j^{(k)} = P_\lambda'^{SCAD}(|\beta_j^{(k)}|).$$

Step 2.2: Weighted L1-Penalized Quantile Regression Update

Update $\beta^{(k+1)}$ by solving:

$$\beta^{(k+1)} = arg\min_{\beta} \sum_{i=1}^{n} \rho_\tau (y_i - x_i^{\top}\beta) + \sum_{j=1}^{p} w_j^{(k)} |\beta_j| \ .$$

Step 2.3: Convergence Check

If $||\beta^{(k+1)} - \beta^{(k)}||_2 < \varepsilon$, stop the iterations.

Step 3: Output

Return $\hat{\beta}(\tau) = \beta^{(k+1)}$.

3. Real Data Description

The real data analysis is conducted using a publicly available heart disease dataset obtained from the UCI Machine Learning Repository. The dataset consists of clinical records for patients undergoing cardiovascular evaluation. The response variable represents a continuous measure of heart disease severity derived from clinical diagnosis. The set of explanatory variables includes demographic and clinical factors commonly associated with cardiovascular risk, such as age, blood pressure, cholesterol level, maximum heart rate, and selected binary indicators reflecting clinical conditions.

The dataset exhibits correlated predictors and a skewed response distribution, which motivates the use of quantile regression with SCAD regularization. All observations are recorded at the patient level, and standard preprocessing procedures are applied prior to model estimation.

Table 1. Descriptive Statistics of Continuous Variables

Variable	Mean	Std. Dev.	Min	Max
Age (years)	54.6	9.2	29	77
Resting Blood Pressure	131.4	17.2	94	200
Serum Cholesterol	245.2	51.1	126	564
Maximum Heart Rate	150.1	22.6	71	202
ST Depression (Oldpeak)	1.07	1.15	0.0	6.2
Serum Sodium	139.0	3.0	130	146

The descriptive statistics reveal substantial heterogeneity in the clinical measurements, particularly for serum cholesterol and ST depression, which exhibit wide ranges and relatively large standard deviations. This variability suggests asymmetric distributions and heterogeneous patient profiles,

supporting the use of quantile regression to capture differential effects across the distribution of heart disease severity.

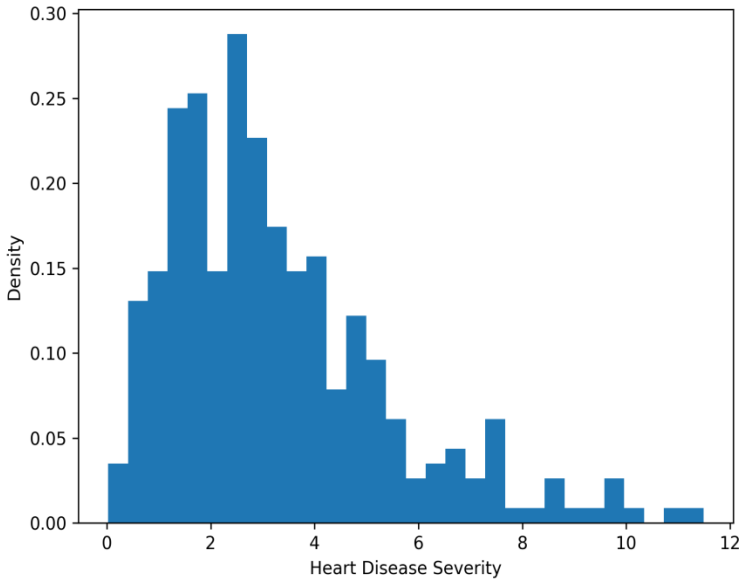

Figure 1. Distribution of Heart Disease Severity

Figure 1 illustrates the empirical distribution of the heart disease severity measure. The distribution exhibits clear asymmetry with a pronounced right skew and a long upper tail, indicating substantial heterogeneity among patients. This pattern suggests that a large proportion of individuals experience mild to moderate disease severity, while a smaller subset presents with markedly severe conditions.

The departure from symmetry and the presence of extreme values imply that mean-based regression models may inadequately represent the underlying structure of the data. Consequently, analyzing conditional effects at different quantiles provides a more informative framework for capturing variation across the full severity spectrum of heart disease.

Table 2. Frequency Distribution of Binary and Categorical Variables

Variable	Coding	Percentage (%)
Sex (Male)	1 = Male, 0 = Female	68.0
Exercise-Induced Angina	1 = Yes, 0 = No	33.0
Typical Angina Chest Pain	1 = Yes, 0 = Otherwise	24.0
Fasting Blood Sugar	1 = High, 0 = Normal	15.0
Resting ECG Abnormality	1 = Yes, 0 = No	22.0
ST Segment Slope Abnormal	1 = Yes, 0 = No	28.0

The categorical variables represent common clinical characteristics observed in cardiovascular assessments. The prevalence patterns align with established epidemiological evidence, while lower-frequency variables may exhibit reduced explanatory power after accounting for stronger correlated predictors.

Table 3. SCAD-Penalized Quantile Regression Estimates

Variable	$\tau = 0.25$	$\tau = 0.50$	$\tau = 0.75$
Age	0.016	0.023	0.032
Sex (Male)	0.000	0.041	0.058
Resting Blood Pressure	0.000	0.013	0.021
Serum Cholesterol	0.000	0.000	0.018
Maximum Heart Rate	-0.026	-0.034	-0.041
ST Depression (Oldpeak)	0.049	0.067	0.091
Exercise-Induced Angina	0.000	0.045	0.072
Typical Angina Chest Pain	-0.031	-0.028	0.000
Fasting Blood Sugar	0.000	0.000	0.000
Resting ECG Abnormality	0.000	0.000	0.000
ST Segment Slope Abnormal	0.000	0.000	0.000
Serum Sodium	0.000	0.000	0.000

The SCAD-penalized quantile regression results highlight pronounced heterogeneity in risk factor effects across quantiles. Age and ST depression exhibit increasing positive effects toward higher quantiles, indicating stronger associations among patients with more severe disease. Maximum heart rate shows a consistent negative association, reflecting its protective role. Several predictors emerge only at higher quantiles, while others are completely eliminated, demonstrating the effectiveness of SCAD in identifying clinically relevant variables.

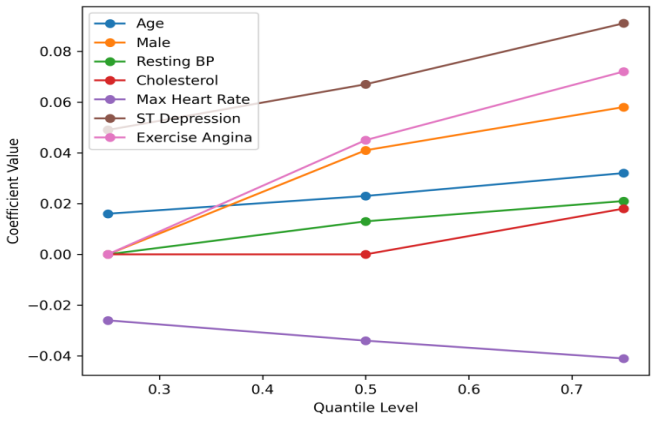

Figure 2. SCAD-Penalized Coefficients Across Quantiles

Figure 2 shows that the effects of several risk factors vary across quantiles of heart disease severity. Age and ST depression increase steadily toward higher quantiles, indicating stronger associations among patients with more severe disease. Maximum heart rate remains negatively associated across all quantiles, reflecting a consistent protective effect. Other predictors appear mainly at the median and upper quantiles, suggesting that their influence is more pronounced in severe cases.

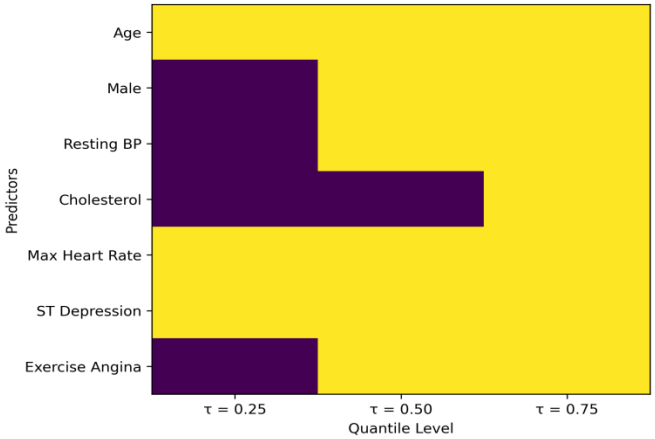

Figure 3. Variable Selection Pattern Across Quantiles

Figure 3 illustrates the variable selection pattern produced by the SCAD penalty across quantile levels. Several predictors are consistently selected at all quantiles, indicating stable clinical relevance, while others are selected only at higher quantiles, reflecting their importance primarily among patients with more severe disease. Variables that are excluded across all quantiles show no independent contribution after accounting for stronger predictors, highlighting the sparsity and interpretability achieved by the SCAD-penalized quantile regression model.

4. Conclusions

This study investigated heart disease risk factors using SCAD-penalized quantile regression, with the aim of capturing heterogeneous covariate effects across different levels of disease severity while achieving reliable variable selection. Unlike mean-based approaches, the proposed framework allowed regression coefficients to vary across quantiles, providing a more nuanced characterization of how clinical predictors influence mild, moderate, and severe forms of heart disease.

The empirical analysis based on real clinical data demonstrated clear distributional heterogeneity in heart disease severity, supporting the suitability of quantile regression. The SCAD penalty effectively produced sparse and interpretable models by retaining clinically meaningful predictors and eliminating irrelevant variables across quantiles. Age and ST depression emerged as

consistently influential factors with increasing effects toward higher quantiles, indicating their stronger association among patients with more severe disease. Maximum heart rate showed a stable negative effect across all quantiles, reflecting a persistent protective role. Other predictors, such as resting blood pressure, serum cholesterol, and exercise-induced angina, were found to be quantile-dependent, becoming relevant primarily in the upper tail of the severity distribution.

Overall, the results highlight the advantage of combining quantile regression with SCAD regularization for heart disease analysis. The proposed approach captures heterogeneous risk patterns, reduces estimation bias, and enhances interpretability through effective variable selection. These features make SCAD-penalized quantile regression a valuable tool for medical studies where disease mechanisms and risk factors differ across severity levels, and where identifying clinically relevant predictors is of primary importance.

References

[1] 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.

[2] Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. Technometrics, 12(1), 55–67.

[3] Koenker, R. (2005). Quantile regression. Cambridge University Press.

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

[5] Koenker, R., & Hallock, K. F. (2001). Quantile regression. Journal of Economic Perspectives, 15(4), 143–156.

[6] Peng, L., & Wang, Y. (2015). An iterative local adaptive majorize–minimization algorithm for SCAD-penalized quantile regression. Computational Statistics & Data Analysis, 82, 54–67.

[7] Mohammed, M. A., & Raheem, S. H. (2020). Determine of the most important factors that affect the incidence of heart disease using logistic regression model (Applied study in Erbil Hospital). Economic Sciences, 15(56), 175-184.

[8] Raheem, S. H., et al. (2025). Bayesian reciprocal group lasso composite quantile regression. Pakistan Journal of Statistics, 41(3), 339–350.

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

[10] Wang, H., Li, G., & Jiang, G. (2007). Robust regression shrinkage and consistent variable selection through the LAD-Lasso. Journal of Business & Economic Statistics, 25(3), 347–355.

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

[12] Zou, H., & Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. Annals of Statistics, 36(4), 1509–1533.