



Article ID 1007-1202(2025)02-0169-15 DOI <https://doi.org/10.1051/wujns/2025302169>

Cite this article: QI Hui, XUE Yaxin. Subgroup Analysis of a Single-Index Threshold Penalty Quantile Regression Model Based on Variable Selection[J]. *Wuhan Univ J of Nat Sci*, 2025, 30(2): 169-183.

Subgroup Analysis of a Single-Index Threshold Penalty Quantile Regression Model Based on Variable Selection

□ QI Hui^{1,2}, XUE Yaxin³

1. School of Statistics and Mathematics, Zhongnan University of Economics and Law, Wuhan 430073, Hubei, China;

2. Institute of Information Engineering, Sanming University, Sanming 365004, Fujian, China;

3. Department of Biostatistics, School of Public Health, Fudan University, Shanghai 200032, China

Abstract: In clinical research, subgroup analysis can help identify patient groups that respond better or worse to specific treatments, improve therapeutic effect and safety, and is of great significance in precision medicine. This article considers subgroup analysis methods for longitudinal data containing multiple covariates and biomarkers. We divide subgroups based on whether a linear combination of these biomarkers exceeds a predetermined threshold, and assess the heterogeneity of treatment effects across subgroups using the interaction between subgroups and exposure variables. Quantile regression is used to better characterize the global distribution of the response variable and sparsity penalties are imposed to achieve variable selection of covariates and biomarkers. The effectiveness of our proposed methodology for both variable selection and parameter estimation is verified through random simulations. Finally, we demonstrate the application of this method by analyzing data from the PA.3 trial, further illustrating the practicality of the method proposed in this paper.

Key words: longitudinal data; subgroup analysis; threshold model; quantile regression; variable selection

CLC number: O212.1

0 Introduction

Subgroups are usually defined based on biomarkers such as blood sugar, blood pressure, gene or protein expression levels, etc^[1]. When evaluating the efficacy of new therapy on patients in clinical studies, these characteristic differences may lead to different responses of the same therapy to different patients^[2]. Therefore, when

physicians assess therapeutic effect, they should not only consider the average effect across the overall population, but also the heterogeneous effects within subgroups. Many methods for subgroup identification have proposed in research literature, such as method based on the tree structure^[3], global model^[4], clustering analysis^[5], etc. In clinical practice, a straightforward and easily interpretable approach is often preferred, which involves dividing subgroups according to whether a single continuous

Received date: 2024-06-28 © Wuhan University 2025

Foundation item: Supported by the Natural Science Foundation of Fujian Province(2022J011177, 2024J01903) and the Key Project of Fujian Provincial Education Department(JZ230054)

Biography: QI Hui, male, Associate professor, Ph. D. candidate, research direction: biostatistics, high dimensional statistics. E-mail: qh19810@126.com

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

biomarker exceeds a certain threshold. This is the threshold model we focus on.

However, in the threshold model described above, a single variable is used to classify subgroups, which can result in a limited representation of subgroup characteristics and may be compromised by the inappropriate choice of the threshold variable. Numerous studies suggest that subgroups may arise from the combined influence of multiple variables, for instance, Jiang *et al.*^[6] demonstrated that multiple single nucleotide polymorphisms (SNPs) can frequently alter the risk of developing specific diseases, whereas any single SNP alone lacks this capability; Fan *et al.*^[7] showed that a risk score defined as a function of multiple predictors is useful for identifying subgroups of AIDS patients who benefit more from treatment. Similarly, Vander Weele *et al.*^[8] discussed the importance of using multiple variables to identify optimal treatment subgroups. He *et al.*^[9] and Wei *et al.*^[10] extended this model to survival and longitudinal data, respectively. While they utilized penalization methods to identify significant biomarkers, they did not perform variable selection for the included covariates.

Moreover, the majority of current studies on threshold model primarily focus on traditional mean regression. However, sometimes we place greater emphasis on sample data at other quantiles, for example, we prefer to conduct heterogeneity analysis on patients with worse condition and provide targeted medical solutions. Consequently, we consider introducing quantile regression to analyze the threshold model. The threshold quantile regression was initially introduced by Cai *et al.*^[11] and applied to quantile self-exciting threshold autoregressive time series models. Galvao *et al.*^[12] studied the threshold quantile autoregressive processes. Furthermore, Lee *et al.*^[13] developed the general tests for threshold effects in regression models, Su and Xu^[14] presented a systematic estimation and inference procedure for threshold quantile regression. On this basis, Zhang *et al.*^[15] proposed a more general single-index threshold quantile regression model. However, there is no existing literature that addresses the variable selection for covariates and biomarkers in single-index threshold quantile regression.

In practical applications, incorporating excessive number of biomarkers for subgroup division may diminish the interpretability of the subgroups, and including excessive number of covariates may also introduce multicollinearity. Consequently, we prefer to select the variables most pertinent to the response from the model's ex-

planatory variables. Common variable selection methods include those based on penalties. This method, initially proposed by Tibshirani in 1996^[16], is known as Lasso estimation, which uses an L1 norm penalty. Subsequent researches include: Fused Lasso penalty^[17], Adaptive Lasso^[18], Graphical Lasso^[19] and so on. Furthermore, as Lasso compresses all coefficients equally, the resulting estimator is often biased. To this end, Fan *et al.*^[20] introduced the smoothly clipped absolute deviation (SCAD) penalty and proved that the estimation result satisfied the oracle property.

Our paper extends the single-index threshold quantile regression^[15] to longitudinal data and introduces the SCAD penalty for variable selection of covariates and biomarkers. To this end, we propose an efficiency algorithm similar to that proposed by Zhang *et al.*^[21], which transforms the problem of estimating the covariate regression coefficients and threshold parameters into a penalized linear quantile regression. Furthermore, we obtain a degraded linear quantile regression through pseudo observations. In addition, the penalty for covariates and biomarkers should be performed separately. We combine the two-step grid search algorithm^[22] to select the compression parameters that minimize the Bayesian information criterion (BIC), and then substitute the pre-selected compression parameters into the iterative algorithm for parameter estimation. Finally, random simulations and example analysis illustrate the practicality of our method.

1 Models and Estimation

We assume that y_{ij} is the response variable of the i -th individual at the j -th observation, $i = 1, \dots, n, j = 1, \dots, m$, total observations $N = nm$, $\mathbf{z}_{ij} = (z_{1,ij}, \dots, z_{p,ij})^T$ is the covariates with dimension of p , where $z_{1,ij} = 1$, $\tilde{\mathbf{z}}_{ij}$ is a subset of $\mathbf{z}_{ij}$, and $\mathbf{x}_i = (x_{i0}, \dots, x_{id})^T$ is the baseline biomarker including the intercept term. We consider the following single-index threshold quantile regression with $\tau \in (0, 1)$:

$$y_{ij} = \mathbf{z}_{ij}^T \boldsymbol{\beta}(\tau) + \tilde{\mathbf{z}}_{ij}^T \boldsymbol{\delta}(\tau) I(\mathbf{x}_i^T \boldsymbol{\gamma}(\tau) > 0) + \varepsilon_{ij},$$

$$i = 1, \dots, n, j = 1, \dots, m.$$

The regression coefficients $\boldsymbol{\beta}(\tau)$ characterize the baseline effect of $\mathbf{z}_{ij}$ at the τ -quantile of y_{ij} , and the regression coefficients $\boldsymbol{\delta}(\tau)$ describe the heterogeneity of $\tilde{\mathbf{z}}_{ij}$ within the subgroup, the threshold coefficients $\boldsymbol{\gamma}(\tau)$ determine the subgroup based on whether the linear combi-

nation $\mathbf{x}_i^\top \boldsymbol{\gamma}(\tau)$ exceeds 0. We consider time-invariant biomarkers to prevent time-varying biomarkers from dividing repeated observations of the same sample into different subgroups due to the influence of treatment. In addition, the error terms $(\varepsilon_{i1}, \dots, \varepsilon_{im})$ and $(\varepsilon_{k1}, \dots, \varepsilon_{km})$ are typically assumed to be independent for any $i \neq k$ stemming from the index set $\{1, \dots, n\}$, but correlated within $(\varepsilon_{i1}, \dots, \varepsilon_{im})$ for any fixed $i \in \{1, \dots, n\}$. Assume $P(\varepsilon_{ij} < 0 | \mathbf{z}_{ij}, \mathbf{x}_i) = \tau$.

For the recognizability of $\boldsymbol{\gamma}(\tau)$, reference Zhang *et al*^[15] and Horowitz^[23], assuming that at least one threshold variable has a nonzero coefficient, and the corresponding probability distribution is absolutely continuous with respect to the Lebesgue measure conditional on the other threshold variables. The variables in $\mathbf{x}_i$ can be rearranged so that x_{1i} meets this condition, then the model can be rewritten as:

$$Q_{\boldsymbol{\gamma}}(\tau | \mathbf{z}_{ij}, \mathbf{x}_i) = \mathbf{z}_{ij}^\top \boldsymbol{\beta}(\tau) + \tilde{\mathbf{z}}_{ij}^\top \boldsymbol{\delta}(\tau) I(x_{1i} + \mathbf{x}_{2i}^\top \boldsymbol{\psi}(\tau) > 0),$$

where $\mathbf{x}_i = (x_{1i}, \mathbf{x}_{2i}^\top)^\top$, $\boldsymbol{\psi}(\tau) = \frac{\boldsymbol{\gamma}_{-1}(\tau)}{\boldsymbol{\gamma}_1(\tau)}$ with $\boldsymbol{\gamma}_1(\tau)$ and $\boldsymbol{\gamma}_{-1}(\tau)$ corresponding to the coefficients of x_{1i} and $\mathbf{x}_{2i}$ respectively. If $\boldsymbol{\gamma}_1(\tau)$ is positive, the normalization without altering the coefficients of $\boldsymbol{\beta}(\tau)$ and $\boldsymbol{\delta}(\tau)$. If $\boldsymbol{\gamma}_1(\tau)$ is negative, we should redefined $\boldsymbol{\beta}(\tau)$ and $\boldsymbol{\delta}(\tau)$ to ensure the consistency of the model^[15]. Moreover, for the convenience of expression, we remove (τ) in the parameters below.

In order to estimate all parameters $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \boldsymbol{\delta}^\top, \boldsymbol{\psi}^\top)^\top$, a computationally efficient method is to ignore the possible correlation between repeated observations within the longitudinal data, that is, to minimize the following objective function under the independence framework:

$$\ell_n(\boldsymbol{\theta}) = \frac{1}{nm} \sum_{i=1}^n \sum_{j=1}^m \rho_\tau \left\{ y_{ij} - \mathbf{z}_{ij}^\top \boldsymbol{\beta} - \tilde{\mathbf{z}}_{ij}^\top \boldsymbol{\delta} I(x_{1i} + \mathbf{x}_{2i}^\top \boldsymbol{\psi} > 0) \right\},$$

where $\rho_\tau(u) = \{\tau - I(u < 0)\}u$ is the quantile loss function. In order to address the challenge of discontinuity in the objective function, similar to Seo and Linton^[24], we employ an integrated kernel function to smooth this characteristic function, for example, the cumulative distribution function of the standard normal distribution $\Phi(\cdot)$, then we can derive the smoothing objective function as following:

$$S_n(\boldsymbol{\theta}; h_n) = \frac{1}{nm} \sum_{i=1}^n \sum_{j=1}^m \rho_\tau \left\{ y_{ij} - \mathbf{z}_{ij}^\top \boldsymbol{\beta} - \tilde{\mathbf{z}}_{ij}^\top \boldsymbol{\delta} \Phi \left(\frac{x_{1i} + \mathbf{x}_{2i}^\top \boldsymbol{\psi}}{h_n} \right) \right\}.$$

In refer to Zhang *et al*^[15], this article selects the bandwidth $h_n = 0.5 \log(N) / \sqrt{N}$.

Furthermore, it is hoped to select important covariates $\mathbf{z}_{ij}$ and threshold variables $\mathbf{x}_{2i}$. In practice, a few exposure variables of interest are included in $\tilde{\mathbf{z}}_{ij}$, such as treatment group indices or drug doses, thus variable selection for interaction terms is not considered. Consequently, the objective function, which includes corresponding penalty terms, is formulated as:

$$Q_n(\boldsymbol{\theta}; h_n) = S_n(\boldsymbol{\theta}; h_n) + \sum_{l=1}^p p_{\lambda_1}(|\beta_l|) + \sum_{l=1}^d p_{\lambda_2}(|\psi_l|),$$

where $p_\lambda(\cdot)$ is the penalty function, and λ_1 and λ_2 are the shrinkage adjustment parameters that control the partial linear regression parameters and threshold parameters, respectively. We consider the following non-convex SCAD penalty function:

$$\begin{aligned} p_\lambda(|\alpha|) &= \lambda |\alpha| I(0 < |\alpha| < \lambda) \\ &\quad + \frac{(a^2 - 1)\lambda^2 - (|\alpha| - a\lambda)^2}{2(a - 1)} I(\lambda < |\alpha| < a\lambda) \\ &\quad + \frac{(a^2 + 1)\lambda^2}{2} I(|\alpha| \geq a\lambda), \end{aligned}$$

when $\lambda > 0$, and $a = 3.7$. SCAD penalty approaches the optimal Bayesian risk^[20], therefore, the nuisance parameters involved in the model only are $\lambda_1 > 0$ and $\lambda_2 > 0$.

Since the objective function $Q_n(\boldsymbol{\theta}; h_n)$ is non-convex and not differentiable everywhere, we use the local linear approximation algorithm of Zou and Li^[25] to transform the non-convex SCAD penalty into a convex approximation form. Specifically,

$$\begin{aligned} p_{\lambda_1}(|\beta_l|) &\approx p_{\lambda_1}(|\hat{\beta}_l^{(0)}|) + p'_{\lambda_1}(|\hat{\beta}_l^{(0)}|)(|\beta_l| - |\hat{\beta}_l^{(0)}|), \text{ for } \beta_l \approx \hat{\beta}_l^{(0)}, \\ p_{\lambda_2}(|\psi_l|) &\approx p_{\lambda_2}(|\hat{\psi}_l^{(0)}|) + p'_{\lambda_2}(|\hat{\psi}_l^{(0)}|)(|\psi_l| - |\hat{\psi}_l^{(0)}|), \text{ for } \psi_l \approx \hat{\psi}_l^{(0)}. \end{aligned}$$

Moreover, even if we employ a profile estimation similar to Zhang *et al*^[15] to estimate the slope $(\boldsymbol{\beta}^\top, \boldsymbol{\delta}^\top)^\top$ and threshold parameters $\boldsymbol{\psi}$ separately, the objective function is still non-convex with respect to $\boldsymbol{\psi}$ with fixing $(\boldsymbol{\beta}^\top, \boldsymbol{\delta}^\top)^\top$. Owing to the slow computation speed of genetic algorithms and the sensitivity of the Nelder-Mead method to initial values, therefore, we employ the first-order Taylor linear approximation to Φ ,

$$\Phi \left(\frac{x_{1i} + \mathbf{x}_{2i}^\top \boldsymbol{\psi}}{h_n} \right) \approx \Phi \left(\frac{x_{1i} + \mathbf{x}_{2i}^\top \hat{\boldsymbol{\psi}}^{(0)}}{h_n} \right) + \Phi' \left(\frac{x_{1i} + \mathbf{x}_{2i}^\top \hat{\boldsymbol{\psi}}^{(0)}}{h_n} \right) \frac{\mathbf{x}_{2i}^\top}{h_n} (\boldsymbol{\psi} - \hat{\boldsymbol{\psi}}^{(0)}),$$

where $\Phi'(\cdot)$ indicates the probability density function of the standard normal distribution. Accordingly, with initial estimation for $\hat{\boldsymbol{\delta}}$ and $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\psi}})$, we can derive a new objective function $\tilde{Q}(\boldsymbol{\beta}, \boldsymbol{\psi})$ with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\psi}$:

$$\tilde{Q}(\boldsymbol{\beta}, \boldsymbol{\psi}) = \sum_{i=1}^n \sum_{j=1}^m \rho_\tau \left(\tilde{y}_{ij} - \mathbf{z}_{ij}^\top \boldsymbol{\beta} - \tilde{\mathbf{x}}_{ij}^\top \boldsymbol{\psi} \right)$$

$$+N \sum_{l=1}^p p'_{\lambda_1}(|\hat{\beta}_l|)|\beta_l| + N \sum_{l=1}^d p'_{\lambda_2}(|\hat{\psi}_l|)|\psi_l|,$$

and

$$\begin{aligned} \tilde{y}_{ij} &= y_{ij} - \tilde{z}_{ij}^T \hat{\delta} \Phi \left(\frac{x_{1i} + \mathbf{x}_{2i}^T \hat{\psi}}{h_n} \right) + \tilde{z}_{ij}^T \hat{\delta} \Phi \left(\frac{x_{1i} + \mathbf{x}_{2i}^T \hat{\psi}}{h_n} \right) \frac{\mathbf{x}_{2i}^T}{h_n} \hat{\psi}, \\ \tilde{\mathbf{x}}_i &= \tilde{z}_{ij}^T \hat{\delta} \Phi \left(\frac{x_{1i} + \mathbf{x}_{2i}^T \hat{\psi}}{h_n} \right) \frac{\mathbf{x}_{2i}}{h_n}. \end{aligned}$$

Now, the estimation of β and ψ is essentially a penalized linear quantile regression problem, specific iterative algorithm is as follows:

Step 0 Initialize $(\hat{\beta}^{(0)}, \hat{\psi}^{(0)})$.

Step 1 Given $(\hat{\beta}^{(k-1)}, \hat{\psi}^{(k-1)})$, we have:

$$\hat{\delta}^{(k)} = \arg \min \sum_{i=1}^n \sum_{j=1}^m \rho_{\tau} \left\{ y_{ij} - \mathbf{z}_{ij}^T \hat{\beta}^{(k-1)} - \tilde{z}_{ij}^T \hat{\delta} \Phi \left(\frac{x_{1i} + \mathbf{x}_{2i}^T \hat{\psi}^{(k-1)}}{h_n} \right) \right\}.$$

Step 2 Given $\hat{\delta}^{(k)}$, the estimation $(\hat{\beta}^{(k)}, \hat{\psi}^{(k)})$ is derived by minimizing the following objective function:

$$\begin{aligned} \tilde{Q}(\beta, \psi) &= \sum_{i=1}^n \sum_{j=1}^m \rho_{\tau}(\tilde{y}_{ij} - \mathbf{z}_{ij}^T \beta - \tilde{\mathbf{x}}_i^T \psi) \\ &\quad + N \sum_{l=1}^p p'_{\lambda_1}(|\hat{\beta}_l^{(k-1)}|)|\beta_l| \\ &\quad + N \sum_{l=1}^d p'_{\lambda_2}(|\hat{\psi}_l^{(k-1)}|)|\psi_l|, \end{aligned}$$

and

$$\begin{aligned} \tilde{y}_{ij} &= y_{ij} - \tilde{z}_{ij}^T \hat{\delta}^{(k)} \Phi \left(\frac{x_{1i} + \mathbf{x}_{2i}^T \hat{\psi}^{(k-1)}}{h_n} \right) \\ &\quad + \tilde{z}_{ij}^T \hat{\delta}^{(k)} \Phi \left(\frac{x_{1i} + \mathbf{x}_{2i}^T \hat{\psi}^{(k-1)}}{h_n} \right) \frac{\mathbf{x}_{2i}^T}{h_n} \hat{\psi}^{(k-1)}, \end{aligned}$$

set $\tilde{\mathbf{x}}_i = \tilde{z}_{ij}^T \hat{\delta}^{(k)} \Phi \left(\frac{x_{1i} + \mathbf{x}_{2i}^T \hat{\psi}^{(k-1)}}{h_n} \right) \frac{\mathbf{x}_{2i}}{h_n}$, repeat step 1 and step

2 until the L2 norm of adjacent estimation for all parameters are less than the tolerance error.

Finally, according to the facts $\rho_{\tau}(cv) = c\rho_{\tau}(v)$ ($c > 0$) and $|\beta_l| = \rho_{\tau}(\beta_l) + \rho_{\tau}(-\beta_l)$, set $Np'_{\lambda_1}(|\hat{\beta}_l^{(k-1)}|) = c_l$, and $Np'_{\lambda_2}(|\hat{\psi}_l^{(k-1)}|) = d_l$, then, we rewrite the first and second penalty terms as follows:

$$\sum_{l=1}^p (\rho_{\tau}(c_l \beta_l) + \rho_{\tau}(-c_l \beta_l)), \sum_{l=1}^d (\rho_{\tau}(d_l \psi_l) + \rho_{\tau}(-d_l \psi_l)).$$

Furthermore, we construct the "unpenalized" linear quantile regression through incorporating additional pseudo observations. The augmented form of the defining variables is as follows:

$$\tilde{y}_l^+ = \begin{cases} \tilde{y}_l, & l=1, \dots, N \\ 0, & l=N+1, \dots, N+2p+2d \end{cases}$$

$$\tilde{z}_l^+ = \begin{cases} z_l, & l=1, \dots, N, \\ (0, \dots, 0, c_l, 0, \dots, 0), & l=N+1, \dots, N+p, \\ (0, \dots, 0, -c_l, 0, \dots, 0), & l=N+p+1, \dots, N+2p, \\ (0, \dots, 0), & l=N+2p+1, \dots, N+2p+2d, \end{cases}$$

$$\tilde{\mathbf{x}}_l^+ = \begin{cases} x_l, & l=1, \dots, N, \\ (0, \dots, 0), & l=N+1, \dots, N+2p, \\ (0, \dots, 0, d_l, 0, \dots, 0), & l=N+2p+1, \dots, N+2p+d, \\ (0, \dots, 0, -d_l, 0, \dots, 0), & l=N+2p+d+1, \dots, N+2p+2d, \end{cases}$$

then the objective function in Step 2 degenerates into linear quantile regression: $\sum_{l=1}^{N+2p+2d} \rho_{\tau}(\tilde{y}_l^+ - \tilde{z}_l^{+T} \beta - \tilde{\mathbf{x}}_l^{+T} \psi)$, and proceed with solving it.

In practice, BIC is frequently employed in variable selection problems for tuning parameter selection,

$$\begin{aligned} \text{BIC}(\lambda_1, \lambda_2) &= \log \left(\sum_{i=1}^n \sum_{j=1}^m \rho_{\tau} \left\{ y_{ij} - \mathbf{z}_{ij}^T \hat{\beta}_{\lambda_1} - \tilde{z}_{ij}^T \hat{\delta} \Phi \left(\frac{x_{1i} + \mathbf{x}_{2i}^T \hat{\psi}_{\lambda_2}}{h_n} \right) \right\} \right) \\ &\quad + \frac{\log(N)}{2N} d_{\lambda}, \end{aligned}$$

d_{λ} represents the total count of non-zero coefficients in both the slope and threshold components, while N denotes the total number of observations. Following Ruppert and Carroll^[22], we employ a two-step grid search algorithm to derive the optimal tuning parameters (λ_1, λ_2) .

2 Simulation

In order to assess the performance of our proposed estimation in finite samples, we consider the following data generation model:

$$y_{ij} = \mathbf{z}_{ij}^T \beta_0 + \tilde{z}_{ij}^T \delta_0 I \{ x_{1i} + \mathbf{x}_{2i}^T \psi_0 + 0.1 \Phi^{-1}(U_i) > 0 \} + (1 + 0.1 z_{2,ij}) \varepsilon_{ij},$$

covariates $\mathbf{z}_{ij} = (1, z_{1,ij}, \dots, z_{p-1,ij})^T$, and $z_{1,ij} \sim \exp(1)$, $z_{2,ij} \sim N(-1, 1)$, the remaining covariates $z_{3,ij}, \dots, z_{p-1,ij}$ are independent and conform to uniform distribution $U(0, 1)$; interaction variable $\tilde{z}_{ij} = (1, z_{1,ij})^T$, threshold variables $x_{1i} \sim N(0, 1)$, component $x_{2i}^* \sim N(1, 1)$ within $\mathbf{x}_{2i} = (1, x_{2i}^*, \dots, x_{di}^*)^T$, the remaining covariates $x_{3i}, \dots, x_{di}$ are independent and conform to uniform distribution $U(0, 1)$ too; setting $\beta_0 = (1, 1, 1, 0, \dots, 0_{p-3})^T$, $\delta_0 = (1, 1)^T$, and $\psi_0 = (-1, 1, 0, \dots, 0_{d-2})^T$. However, in our data generation model, the true value of $(\beta_{01}, \beta_{03}, \psi_{01})$ varies with the quantile. The random error term $\varepsilon_{ij} = \zeta + e_{ij}$, where ζ ensures $P(\varepsilon_{ij} < 0 | z_{ij}, x_i) = \tau$, we take into account two error

distributions:

Case 1 The random error vector conforms to multivariate normal distribution, that is, $\varepsilon_i = (\varepsilon_{i1}, \dots, \varepsilon_{im})^T \sim \text{MVN}(0, \Sigma)$, and the covariance matrix Σ follows AR-1 correlation structure with correlation coefficient $\rho = 0.3$, where the partial coefficients of the model are set to $(\beta_{01}, \beta_{03}, \psi_{01}) = (1 + \Phi^{-1}(\tau), 1 + 0.1\Phi^{-1}(\tau), 0.1\Phi^{-1}(\tau) - 1)$.

Case 2 The random error vector conforms to multivariate t distribution, that is, $\varepsilon_i = (\varepsilon_{i1}, \dots, \varepsilon_{im})^T \sim \text{MVT}_3(0, \Sigma)$, and the covariance matrix shares the same structure, where $(\beta_{01}, \beta_{03}, \psi_{01}) = (1 + qt(\tau, \text{df} = 3), 1 + 0.1 \times qt(\tau, \text{df} = 3), 0.1\Phi^{-1}(\tau) - 1)$, $qt(\tau, \text{df} = 3)$ is the τ quantile of the t distribution with degrees of freedom (df) 3.

We will assess the performance of our estimation across varying numbers of parameters ($p = d = 10$ or $p = d = 20$) or different quantiles ($\tau = 0.25, 0.5, 0.75$), each random experiment yields ($n = 400$ or 800) samples, with each sample being observed ($m = 4$) times, and the simulation is repeated 500 times. We assess the model's performance using four criterias: 1) True Negative (TN): This represents the average number of zero parameters that are correctly classified as zero by the model. A higher TN indicates better model performance; 2) False Negative (FN): The average number of non-zero parameters erroneously classify as zero. A lower FN indicates better model performance; 3) Correct (%): This represents the percentage of accurate true models

with non-zero parameters that are correctly identified. The closer this percentage is to 100, the more accurate the model estimation effect is; 4) MSE: It represents the mean square error of the parameter estimation, smaller MSE indicates that the estimated value $\hat{\theta}$ is closer to the true value θ_0 , and the better estimation effect of the model. By comparing the mean square error O.MSE of the oracle estimator with the mean square error P.MSE of our method, the P.MSE is closer to O.MSE means the better estimation effect of our approach.

Tables 1 and 2 present the variable selection outcomes of the single-index threshold quantile regression when the error terms follow a multivariate normal distribution or a multivariate T(3) distribution, respectively. In an ideal scenario, the component of TN for the linear segment is $p - 3$, and for the threshold segment it is $d - 2$. The oracle TN in scenario $p = d = 10$ and scenario $p = d = 20$ is 15 or 35, respectively. In all settings, our method can compress the true zero coefficients to zero and correctly identify all relevant covariates (FN=0), with P.MSE and the O.MSE is very close, which indicate that the model estimation is very effective.

Tables 3 and 4 show the estimators of the identified nonzero parameters by our method when the error terms follow a multivariate normal distribution or a multivariate T(3) distribution, respectively. The Biases for all non-zero parameters remain minimal across various parameter counts ($p = d = 10$ or $p = d = 20$) or quantiles ($\tau = 0.25, 0.5, 0.75$), and all metrics: Biases, SDs, and RMSEs decrease as the sample size increases.

Table 1 Variable selection in the case where $\varepsilon_i \sim \text{MVN}(0, \Sigma)$ and Σ follows an AR-1 structure

Parameter	Quantile	Model	$n = 400$				$n = 800$			
			TN	FN	Correct/%	MSE/ 10^{-2}	TN	FN	Correct/%	MSE/ 10^{-2}
$p = d = 10$	0.25	proposed	15	0	100	0.357	15	0	100	0.191
		oracle	15	0	100	0.357	15	0	100	0.191
	0.50	proposed	15	0	100	0.328	15	0	100	0.163
		oracle	15	0	100	0.332	15	0	100	0.160
	0.75	proposed	15	0	100	0.387	15	0	100	0.195
		oracle	15	0	100	0.366	15	0	100	0.193
$p = d = 20$	0.25	proposed	35	0	100	0.393	35	0	100	0.195
		oracle	35	0	100	0.357	35	0	100	0.191
	0.50	proposed	35	0	100	0.337	35	0	100	0.158
		oracle	35	0	100	0.332	35	0	100	0.160
	0.75	proposed	35	0	100	0.368	35	0	100	0.193
		oracle	35	0	100	0.366	35	0	100	0.193

Table 4 Estimation results of parameters in the case where $\varepsilon_i \sim \text{MVT}_3(\mathbf{0}, \Sigma)$ and Σ follows an AR-1 structure

Parameter		$p=d=10$						$p=d=20$					
		$n=400$			$n=800$			$n=400$			$n=800$		
		Bias	RMSE	SD	Bias	RMSE	SD	Bias	RMSE	SD	Bias	RMSE	SD
$\tau=0.25$	β_1	0.002 4	0.100 4	0.100 4	0.002 4	0.072 3	0.072 3	0.003 1	0.096 4	0.096 4	0.006 4	0.070 4	0.070 2
	β_2	-0.010 3	0.053 6	0.052 7	-0.004 5	0.039 1	0.038 9	-0.011 9	0.054 8	0.053 6	-0.008 4	0.039 3	0.038 5
	β_3	0.004 3	0.039 6	0.039 4	0.000 4	0.027 9	0.027 9	0.002 9	0.037 5	0.037 5	0.000 6	0.026 4	0.026 4
	δ_1	0.019 9	0.106 8	0.105 0	0.004 7	0.076 6	0.076 5	0.015 2	0.113 2	0.112 3	0.003 1	0.077 5	0.077 5
	δ_2	0.011 9	0.076 2	0.075 3	0.007 5	0.056 3	0.055 9	0.015 6	0.079 6	0.078 1	0.009 8	0.054 9	0.054 0
	ψ_0	0.015 6	0.0697	0.068 0	0.011 3	0.051 8	0.050 6	0.008 6	0.066 2	0.065 7	0.008 4	0.051 2	0.050 6
	ψ_1	0.000 4	0.056 7	0.056 7	0.005 1	0.039 0	0.038 7	0.006 6	0.053 8	0.053 5	0.007 8	0.041 5	0.040 8
$\tau=0.50$	β_1	-0.001 3	0.078 1	0.078 2	-0.001 0	0.050 9	0.051 0	-0.005 9	0.077 6	0.077 4	-0.002 0	0.051 3	0.051 5
	β_2	-0.006 2	0.041 7	0.041 3	-0.006 0	0.031 9	0.031 4	-0.007 2	0.041 8	0.041 2	-0.002 0	0.030 2	0.030 3
	β_3	0.001 8	0.031 8	0.041 3	0.000 2	0.022 7	0.022 7	0.000 2	0.030 7	0.030 7	0.002 3	0.021 4	0.021 3
	δ_1	0.012 6	0.090 7	0.090 0	0.004 4	0.060 4	0.060 3	0.021 2	0.092 1	0.089 8	0.001 5	0.064 3	0.064 5
	δ_2	0.013 0	0.063 1	0.061 9	0.009 7	0.046 5	0.045 5	0.011 1	0.058 2	0.057 1	0.008 8	0.047 5	0.046 8
	ψ_0	-0.001 1	0.060 1	0.060 2	0.000 1	0.044 1	0.044 2	-0.005 2	0.061 8	0.061 6	-0.003 4	0.043 8	0.043 8
	ψ_1	0.002 6	0.053 2	0.053 3	0.000 5	0.036 9	0.036 9	0.006 2	0.050 7	0.050 4	0.006 2	0.037 6	0.037 2
$\tau=0.75$	β_1	-0.017 5	0.095 4	0.093 9	-0.006 1	0.067 0	0.066 8	-0.017 3	0.089 7	0.088 2	-0.010 0	0.066 1	0.065 4
	β_2	-0.005 4	0.051 0	0.050 8	0.000 5	0.034 2	0.034 2	0.000 0	0.051 8	0.051 9	0.000 4	0.035 5	0.035 6
	β_3	-0.003 2	0.036 9	0.036 8	0.000 3	0.027 0	0.046 0	-0.003 2	0.038 2	0.038 2	-0.000 9	0.025 6	0.025 6
	δ_1	0.009 6	0.110 5	0.110 2	0.001 4	0.076 9	0.077 0	0.012 0	0.115 6	0.115 3	0.000 0	0.072 4	0.072 5
	δ_2	0.019 8	0.079 0	0.076 6	0.010 9	0.053 4	0.052 4	0.015 1	0.081 7	0.080 4	0.010 3	0.052 3	0.051 3
	ψ_0	-0.022 1	0.072 5	0.069 1	-0.019 2	0.049 9	0.046 0	-0.020 9	0.075 1	0.072 2	-0.017 3	0.050 0	0.047 0
	ψ_1	0.006 1	0.058 0	0.057 7	0.001 3	0.038 5	0.038 5	0.004 9	0.058 7	0.058 6	0.002 4	0.037 8	0.037 8

We further conduct sensitivity comparisons for two bandwidths $r = \log(N)/\sqrt{N}$ and $r^{\text{adj}} = 1/\sqrt[5]{N}$ under $p = d = 10$, $\tau = 0.5$, while keeping the remaining settings as the same as before. The RMSEs of the proposed estimators in Fig. 1 show that these two bandwidths do not exert significant effects on the performance of the proposed method

3 Real Data Analysis

We apply the proposed method to analyze the PA.3 trial dataset conducted by the Canadian Cancer Group. In this trial, patients with locally advanced or metastatic pancreatic cancer were randomly assigned to the erlotinib plus gemcitabine group or the gemcitabine monotherapy group. In the primary analysis, the survival rate of the erlotinib plus gemcitabine group was significantly improved compared with the gemcitabine monotherapy group^[26]. Shultz *et al*^[27] found that when each biomarker was evaluated individually, none of them was signifi-

cantly associated with erlotinib. However, He *et al*^[9] selected CA19-9 and AXL from 47 biomarkers to define the treatment-sensitive subgroup, and found that the erlotinib plus gemcitabine could prolong patient's survival time compared with gemcitabine monotherapy. Similarly, Wei *et al*^[10] selected CA19-9 and AXL from 6 biomarkers to define a treatment-sensitive subgroup and found that erlotinib plus gemcitabine improved patient's quality of life (QoL) scores compared with gemcitabine monotherapy. However, these studies only focused on selecting biomarkers for defining subgroups, not covariates. Moreover, none of these studies investigated whether distinct subgroup effects exist in patients across various QoL levels. Therefore, we use our method to select relevant covariates and biomarkers, and assess the differential therapeutic effect of erlotinib in subgroups defined by linear combinations of these biomarkers, as well as the differences in effects at different quantile levels of the outcomes.

A total of 377 patients were included in the analysis, of whom 196 received erlotinib treatment, and their quality of life was assessed using the global QoL score. For each patient in the trial, QoL scores were recorded at the baseline, as well as 4, 8, 12, and 16 weeks following the treatment. We focus on the changes in patient's QoL scores relative to the baseline (cQoL), and missing values have imputed using multiple imputation. The Q-Q plot of the imputed variable $y = \text{cQoL}$ and the Shapiro-Wilk normality test indicate that the data do not conform to a normal distribution. The calculated skewness is 0.139 3, indicating a right skewed distribution, so that quantile regression is considered to be more suitable than the mean regression. The covariates included in the analysis are: baseline QoL score (bQoL), time point

(Time), baseline age (Age), treatment index B (if the patient received erlotinib plus gemcitabine, $B = 1$; if the patient received gemcitabine alone, $B=0$), and pain level (Painmeas). The twelve biomarkers included in the analysis include proteins with lower overall survival and baseline expression levels associated with metastatic diseases, specifically: AXL: AXL receptor tyrosine kinase; CA 19-9: Carbohydrate antigen 19-9; IL8: Interleukin-8; CEA: Carcinoembryonic antigen; MUC-1: Mucin 1, cell surface associate; PDGFRA: Platelet-derived growth factor receptor; EGFR: Epidermal growth factor receptor; PDK1: Pyruvate dehydrogenase kinase 1; BMP-2: Bone morphogenetic protein 2; HER2: Erb-b2 receptor tyrosine kinase 2; PF4: Platelet factor 4; GAS6: Growth arrest-specific protein 6. We

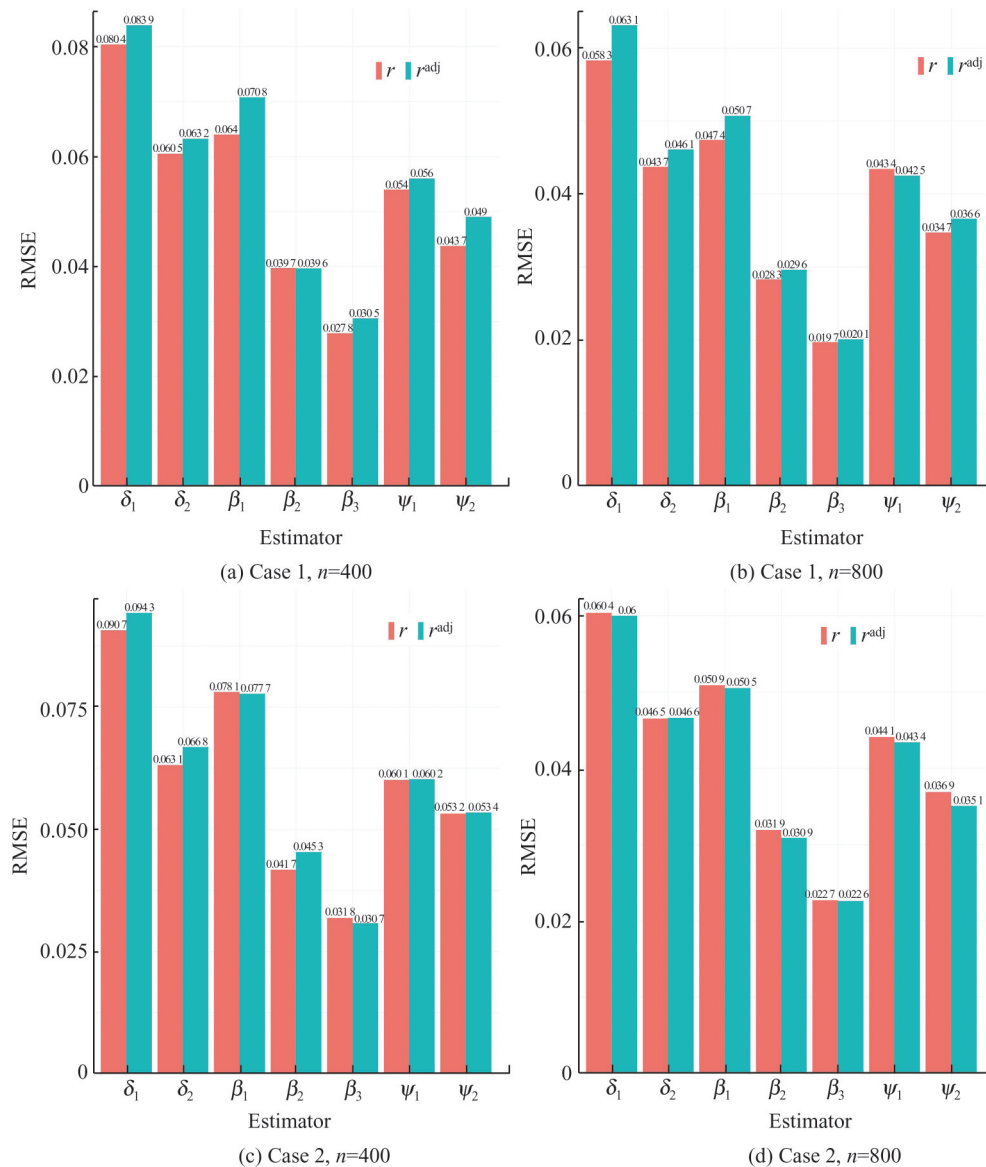


Fig. 1 The sensitivity diagram of parameter estimation under the bandwidth rates r and r^{adj} with $p=d=10$ and $\tau=0.5$

standardize the 12 biomarkers and impute missing values using the median.

It is worth noting that our model requires the development of significant biomarker x_1 and assumes its coefficient to be nonzero. With prior knowledge, we should primarily consider results from expert opinions or existing literature. In terms of analyzing the PA.3 dataset, existing literature has selected CA19-9 and AXL as significant biomarkers in mean regression^[10] and Cox regression^[9] analysis based on threshold models for y . Hence, we may consider utilizing CA19-9 or AXL as significant threshold variable for defining subgroups, with one designate as x_1 and the rest store in x_2 , we can also compare the differences in subgroup analysis when different biomarker is designated as x_1 . We construct the following single-index threshold quantile regression to fit the data:

$$\begin{aligned} \text{cQoL}_{ij} = & \beta_1 + \beta_2 \text{bQoL}_i + \beta_3 \text{Time}_{ij} + \beta_4 \text{Age}_i + \beta_5 B_i \\ & + \beta_6 \text{Painmeas}_i + \delta_1 I(x_{1i} + \mathbf{x}_{2i}^T \boldsymbol{\psi} > 0) \\ & + \delta_2 B_i I(x_{1i} + \mathbf{x}_{2i}^T \boldsymbol{\psi} > 0) + \varepsilon_{ij}, i = 1, \dots, 377, j = 1, \dots, 4. \end{aligned}$$

Set $y = \text{cQoL}$, $\mathbf{z} = (1, \text{bQoL}, \text{Time}, \text{Age}, B, \text{Painmeans})$, $\tilde{\mathbf{z}} = (1, B)$, $x_1 = \text{CA19-9 or AXL}$, $\mathbf{x} = (1, \text{CA19-9}, \text{AXL}, \text{IL8}, \text{CEA}, \text{MUC1}, \text{PDGFRA}, \text{EGFR}, \text{PDK1}, \text{IP10}, \text{HER2}, \text{PF4}, \text{GAS6})$, $\mathbf{x}_2$ represents the part of $\mathbf{x}$ minus x_1 , assuming that the error term satisfies $P(\varepsilon_{ij} < 0 | \mathbf{z}_{ij}, \mathbf{x}_i) = \tau$.

In order to conduct subgroup analysis of outcome variables at different levels as comprehensively as possible, we consider 7 quantile levels $\tau = 0.2, 0.3, \dots, 0.8$, and the same two-step grid search method as in the simulation study is used to select the tuning parameters λ_1 and λ_2 . Table 5 presents the point estimator, bootstrap standard error (BS-SE), 95% confidence interval (95%-CI), and p -value of each parameter obtained at the above quantile levels with fixing $x_1 = \text{AXL}$.

From Table 5, it can be seen that in terms of variable selection, our method sets the coefficients for observation time (Time) and treatment index variable (B) to 0 across all quantiles ranging from 0.2 to 0.8, additionally, it sets the coefficients for age (Age) to 0 at quantiles 0.4 to 0.8, while the effect at quantiles 0.2 to 0.3 have not statistically significant. Thus, we conclude that the effects of observation time (Time) and age (Age) on patient's QoL scores are not significant. However, the selection of treatment effect (B) as an unrelated variable may be due to the masking of the interaction effect between treatment and subgroups, resulting in an insignificant effect, which requires further confirmation in stratified analysis. In addition, CA19-9 is selected as an im-

portant biomarker at all quantiles, and the coefficients of the remaining 10 biomarkers are compressed to 0. The biomarkers we selected are consistent with the results of He *et al*^[9] and Wei *et al*^[10].

In terms of subgroup analysis, within the mid or low percentiles (0.2-0.6), the p -value of the interaction coefficient δ_2 between treatment and subgroup is less than 0.05, this demonstrates a significant difference in therapeutic effect between the two subgroups at the 0.05 significance level. However, within the high percentile range (0.7-0.8), the corresponding p -value is met with $p > 0.05$, this demonstrates that the difference in therapeutic effect between the two subgroups is not significant at the 0.05 significance level, the possible reason is that there is no treatment sensitive subgroup, and further exploration is required through stratified analysis.

Next, we define the biomarker positive subgroup and the biomarker negative subgroup according to linear combination of fixing biomarker AXL, intercept and selected biomarker CA19-9, corresponding coefficients (ψ_0, ψ_1) as shown in Table 5. Linear quantile regression has conducted on all covariates separately for the samples in two subgroups, the estimated therapeutic effect and their 95% confidence intervals at different quantiles present in Fig. 2, and the confidence intervals are derived from rank tests. Figure 2 reveals that there is significant heterogeneity in therapeutic effect between the two subgroups. Specifically, at all quantiles, within the biomarker positive subgroup, the therapeutic effect are all significant positive values, and the 95% confidence intervals mostly fall above the zero line, which indicate that erlotinib has a positive impact on improving QoL scores. Therefore, we believe that at fixed $x_1 = \text{AXL}$, the identified biomarker positive subgroup is the treatment sensitive subgroup, in which patients receiving treatment of erlotinib plus gemcitabine have significantly improved quality of life scores compared to treatment with gemcitabine alone. However, within the biomarker negative subgroup, the 95% confidence interval for therapeutic effect remains zero, which indicates that the results are not significant. Therefore, we believe that in this subgroup, patients do not benefit more from erlotinib plus gemcitabine treatment compared to treatment with gemcitabine alone.

Moreover, Table 6 presents the same four estimators at the corresponding quantile level with fixing $x_1 = \text{CA19-9}$. As can be seen from Table 6, the corresponding

Table 5 Estimation results of parameters when fixing $x_1 = \text{AXL}$

τ	Estimator	Estimate	BS-SE	95%-CI	p -value	τ	Estimator	Estimate	BS-SE	95%-CI	p -value
0.2	β_1	32.976	7.760	(17.766, 48.187)	<0.001	0.2	β_6	-0.112	0.050	(-0.212, -0.012)	0.026
0.3		27.897	3.686	(20.671, 35.123)	<0.001	0.3		-0.034	0.030	(-0.093, 0.024)	0.256
0.4		31.430	5.824	(20.014, 42.845)	<0.001	0.4		-0.060	0.036	(-0.130, 0.010)	0.096
0.5		36.214	3.805	(28.755, 43.673)	<0.001	0.5		-0.046	0.033	(-0.113, 0.019)	0.164
0.6		45.531	5.274	(35.194, 55.870)	<0.001	0.6		-0.093	0.035	(-0.163, -0.024)	0.007
0.7		50.220	4.606	(41.191, 59.249)	<0.001	0.7		-0.088	0.048	(-0.183, 0.007)	0.070
0.8		60.540	5.813	(49.145, 71.935)	<0.001	0.8		-0.125	0.047	(-0.218, -0.033)	0.007
0.2	β_2	-0.617	0.055	(-0.726, -0.509)	<0.001	0.2	δ_1	-3.781	4.058	(-11.736, 4.172)	0.351
0.3		-0.548	0.029	(-0.606, -0.491)	<0.001	0.3		-6.397	2.993	(-12.265, -0.529)	0.032
0.4		-0.568	0.066	(-0.698, -0.438)	<0.001	0.4		-5.892	5.067	(-15.824, 4.039)	0.244
0.5		-0.543	0.046	(-0.632, -0.450)	<0.001	0.5		-4.541	3.520	(-11.440, 2.358)	0.197
0.6		-0.613	0.056	(-0.723, -0.502)	<0.001	0.6		-5.031	2.871	(-10.660, 0.596)	0.079
0.7		-0.601	0.050	(-0.700, -0.502)	<0.001	0.7		-7.312	2.663	(-12.532, -2.092)	0.006
0.8		-0.657	0.056	(-0.767, -0.547)	<0.001	0.8		-10.748	2.964	(-16.558, -4.937)	<0.001
0.2	β_3	0	*	*	*	0.2	δ_2	10.080	4.811	(0.649, 19.511)	0.036
0.3		0	*	*	*	0.3		9.299	4.086	(1.290, 17.309)	0.028
0.4		0	*	*	*	0.4		12.118	5.896	(0.560, 23.675)	0.039
0.5		0	*	*	*	0.5		9.424	4.782	(0.049, 18.798)	0.048
0.6		0	*	*	*	0.6		7.773	3.709	(0.502, 15.043)	0.036
0.7		0	*	*	*	0.7		6.841	5.344	(-3.633, 17.315)	0.200
0.8		0	*	*	*	0.8		10.701	6.638	(-2.309, 23.712)	0.106
0.2	β_4	-0.136	0.096	(-0.325, 0.052)	0.157	0.2	ψ_0	-0.982	0.210	(-1.395, -0.568)	<0.001
0.3		-0.027	0.054	(-0.133, 0.078)	0.610	0.3		-0.665	0.148	(-0.956, -0.372)	<0.001
0.4		0	*	*	*	0.4		-0.885	0.113	(-1.107, -0.662)	<0.001
0.5		0	*	*	*	0.5		-1.093	0.127	(-1.342, -0.842)	<0.001
0.6		0	*	*	*	0.6		-0.869	0.116	(-1.907, -0.640)	<0.001
0.7		0	*	*	*	0.7		-0.911	0.122	(-1.151, -0.671)	<0.001
0.8		0	*	*	*	0.8		-0.848	0.147	(-1.136, -0.559)	<0.001
0.2	β_5	0	*	*	*	0.2	ψ_1	-1.217	0.200	(-1.609, -0.824)	<0.001
0.3		0	*	*	*	0.3		-1.032	0.197	(-1.418, -0.644)	<0.001
0.4		0	*	*	*	0.4		-0.824	0.252	(-1.319, -0.328)	0.001
0.5		0	*	*	*	0.5		-1.238	0.232	(-1.694, -0.781)	<0.001
0.6		0	*	*	*	0.6		-0.817	0.209	(-1.227, -0.407)	<0.001
0.7		0	*	*	*	0.7		-0.834	0.243	(-1.311, -0.357)	<0.001
0.8		0	*	*	*	0.8		-0.824	0.207	(-1.230, -0.417)	<0.001

Note: *: The previous coefficient estimates are zero, indicating that the covariate has not been selected into the model, so the later estimates, BS-SE, 95%-CI, and p -value have no estimates.

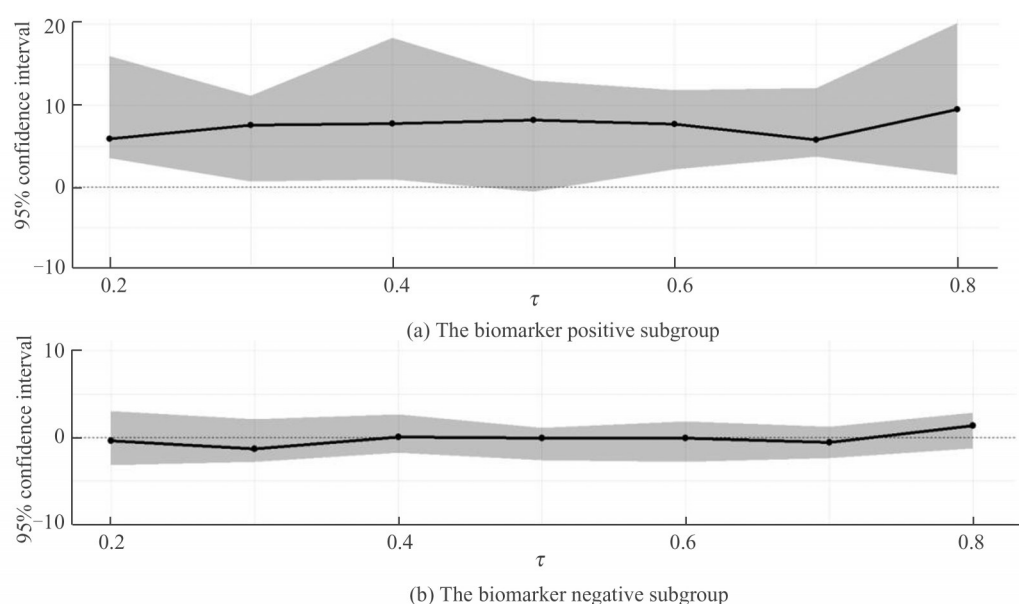


Fig. 2 The 95% confidence intervals of therapeutic effect when fixing $x_1 = \text{AXL}$

main results of variable selection and subgroup analysis are similar to those in Table 5 with fixing $x_1 = \text{AXL}$.

Similarly, we also define the biomarker positive subgroup and the biomarker negative subgroup according to linear combination of fixing biomarker CA19-9. The corresponding estimated therapeutic effect and their 95% confidence intervals at different quantiles present in Fig. 3. Figure 3 shows that there is still significant heterogeneity in the therapeutic effect between the two subgroups. But unlike in the case of fixing $x_1 = \text{AXL}$, when $x_1 = \text{CA19-9}$ is fixed, the treatment sensitive subgroup with significant therapeutic effect is the identified biomarker negative subgroup, patients in this subgroup who receive treatment with erlotinib plus gemcitabine will significantly improve their overall quality of life score, however, within the biomarker positive subgroup, patients did not benefit more from the combination therapy of erlotinib and gemcitabine. This indicates that the 0-1 values of the subgroup index variables have reversed in both scenarios, that is, the interaction coefficient δ_2 between treatment and subgroups is negative when $x_1 = \text{CA19-9}$, and positive when $x_1 = \text{AXL}$. It also explains why the coefficient β_5 of the treatment term is selected as a significant correlation variable when $x_1 = \text{CA19-9}$, while it is compressed to 0 when $x_1 = \text{AXL}$.

In summary, combining the results of subgroup analysis for setting $x_1 = \text{AXL}$ and $x_1 = \text{CA19-9}$, we can see that our method can identify treatment sensitive sub-

groups. In this subgroup, patients receiving treatment of erlotinib plus gemcitabine are more beneficial in improving their QoL score compared with using gemcitabine alone, and erlotinib is effective for patients with different QoL levels.

Finally, when fixing $x_1 = \text{AXL}$ or $x_1 = \text{CA19-9}$ separately, we also want to compare whether the defined subgroups are consistent. We draw Fig. 4 to visualize the overlap of the subgroups divided in both cases, when fixing $x_1 = \text{AXL}$, the treatment sensitive subgroup corresponding to the biomarker positive subgroup is defined as group 1; the treatment insensitive subgroup corresponding to the biomarker negative subgroup is defined as group 2. Conversely, when fixing $x_1 = \text{CA19-9}$, the treatment sensitive subgroup corresponding to the biomarker negative subgroup is defined as group 1; the treatment insensitive subgroup corresponding to the biomarker positive subgroup is defined as group 2. Figure 4 shows the intersection and union of group 1 and group 2 in two different scenarios, the orange bars corresponding to 11 indicate that they are all assigned to group 1 in both scenarios, the blue bars corresponding to 22 signify that they are all assigned to group 2, 12 and 21 represent the number of individuals assigned to distinct subgroups in each scenario, with no intersection, and are uniformly displayed with gray bars.

In order to clearly display the proportion of patients divided into different subgroups in both scenarios, we set the highest count point of each subgraph to 100 and compare the height of the orange and gray bars. It can be

Table 6 Estimation results of parameters when fixing $x_1 = \text{CA19-9}$

τ	Estimator	Estimate	BS-SE	95%-CI	p -value	τ	Estimator	Estimate	BS-SE	95%-CI	p -value
0.2	β_1	11.338	6.179	(−0.772, 23.449)	0.066	0.2	β_6	−0.018	0.048	(−0.113, 0.075)	0.695
0.3		24.367	6.458	(11.708, 37.026)	<0.001	0.3		−0.036	0.038	(−0.111, 0.039)	0.348
0.4		25.858	6.439	(13.236, 38.480)	<0.001	0.4		−0.064	0.034	(−0.132, 0.002)	0.058
0.5		30.937	5.707	(19.750, 42.124)	<0.001	0.5		−0.491	0.038	(−0.124, 0.025)	0.199
0.6		39.404	5.892	(27.854, 50.954)	<0.001	0.6		−0.084	0.035	(−0.154, −0.014)	0.018
0.7		43.781	6.667	(30.713, 56.849)	<0.001	0.7		−0.099	0.045	(−0.189, −0.010)	0.029
0.8		49.821	6.653	(36.779, 62.862)	<0.001	0.8		−0.121	0.048	(−0.215, −0.025)	0.012
0.2	β_2	−0.527	0.079	(−0.682, −0.372)	<0.001	0.2	δ_1	6.998	4.101	(−1.040, 15.037)	0.087
0.3		−0.539	0.041	(−0.621, −0.457)	<0.001	0.3		7.841	3.904	(0.188, 15.495)	0.044
0.4		−0.572	0.056	(−0.683, −0.461)	<0.001	0.4		6.188	4.541	(−2.712, 15.089)	0.173
0.5		−0.545	0.050	(−0.645, −0.446)	<0.001	0.5		5.592	3.520	(−1.309, 12.492)	0.112
0.6		−0.602	0.054	(−0.709, −0.495)	<0.001	0.6		4.668	3.435	(−2.064, 11.401)	0.174
0.7		−0.609	0.053	(−0.714, −0.504)	<0.001	0.7		7.063	2.713	(1.745, 12.381)	0.009
0.8		−0.651	0.055	(−0.760, −0.542)	<0.001	0.8		8.933	3.668	(1.743, 16.123)	0.014
0.2	β_3	0	*	*	*	0.2	δ_2	−12.212	5.521	(−23.033, −1.391)	0.026
0.3		0	*	*	*	0.3		−12.164	4.814	(−31.599, −2.728)	0.015
0.4		0.089	0.170	(−0.245, 0.424)	0.601	0.4		−11.361	4.818	(−20.805, −1.918)	0.018
0.5		0	*	*	*	0.5		−11.061	5.090	(−21.038, −1.083)	0.029
0.6		0	*	*	*	0.6		−7.958	5.222	(−18.193, 2.277)	0.127
0.7		0.036	0.182	(−0.320, 0.393)	0.840	0.7		−6.388	5.132	(−16.447, 3.670)	0.213
0.8		0.089	0.291	(−0.481, 0.660)	0.758	0.8		−9.518	7.382	(−23.988, 4.950)	0.197
0.2	β_4	0	*	*	*	0.2	ψ_0	1.230	0.241	(0.755, 1.704)	<0.001
0.3		−0.088	0.073	(−0.233, 0.056)	0.232	0.3		0.628	0.191	(0.252, 1.004)	0.001
0.4		0.007	0.064	(−0.118, 0.133)	0.912	0.4		0.953	0.185	(0.589, 1.316)	<0.001
0.5		0	*	*	*	0.5		0.894	0.161	(0.578, 1.210)	<0.001
0.6		0.005	0.038	(−0.069, 0.079)	0.887	0.6		0.896	0.163	(0.575, 1.216)	<0.001
0.7		0	*	*	*	0.7		0.859	0.207	(0.451, 1.266)	<0.001
0.8		0	*	*	*	0.8		1.047	0.179	(0.695, 1.398)	<0.001
0.2	β_5	11.860	4.821	(2.441, 21.310)	0.013	0.2	ψ_1	−1.401	0.196	(−1.785, −1.015)	<0.001
0.3		10.775	4.572	(1.813, 19.737)	0.018	0.3		−0.973	0.225	(−1.415, −0.530)	<0.001
0.4		11.256	4.538	(2.361, 20.152)	0.013	0.4		−1.069	0.270	(−1.598, −0.539)	<0.001
0.5		10.947	4.756	(1.624, 20.270)	0.021	0.5		−0.770	0.261	(−1.282, −0.258)	0.003
0.6		8.000	4.752	(−1.315, 17.315)	0.092	0.6		−0.877	0.269	(−1.404, −0.349)	0.001
0.7		5.874	5.001	(−3.927, 15.676)	0.240	0.7		−0.952	0.206	(−1.355, −0.547)	<0.001
0.8		9.984	7.172	(−4.073, 24.043)	0.163	0.8		−1.260	0.234	(−1.720, −0.799)	<0.001

Note: *: The previous coefficient estimates are zero, indicating that the covariate has not been selected into the model, so the later estimates, BS-SE, 95%-CI, and p -value have no estimates.

found that the overlap of the treatment sensitive subgroups identified in both scenarios is very high, which in-

dicates that the application of our method for subgroup identification has good robustness in the selection of x_1 .

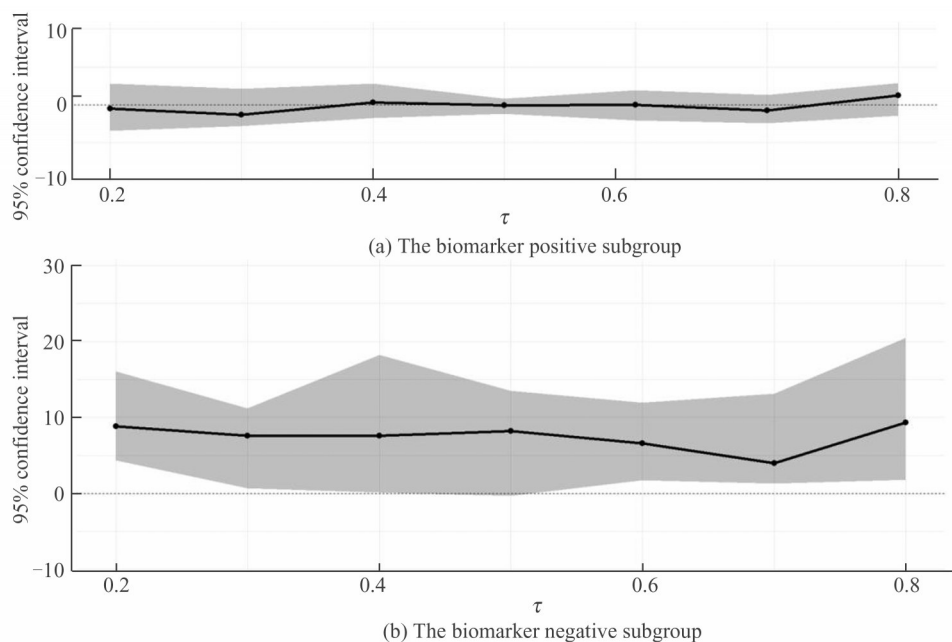


Fig. 3 The 95% confidence intervals of therapeutic effect when fixing $x_1 = \text{CA19-9}$

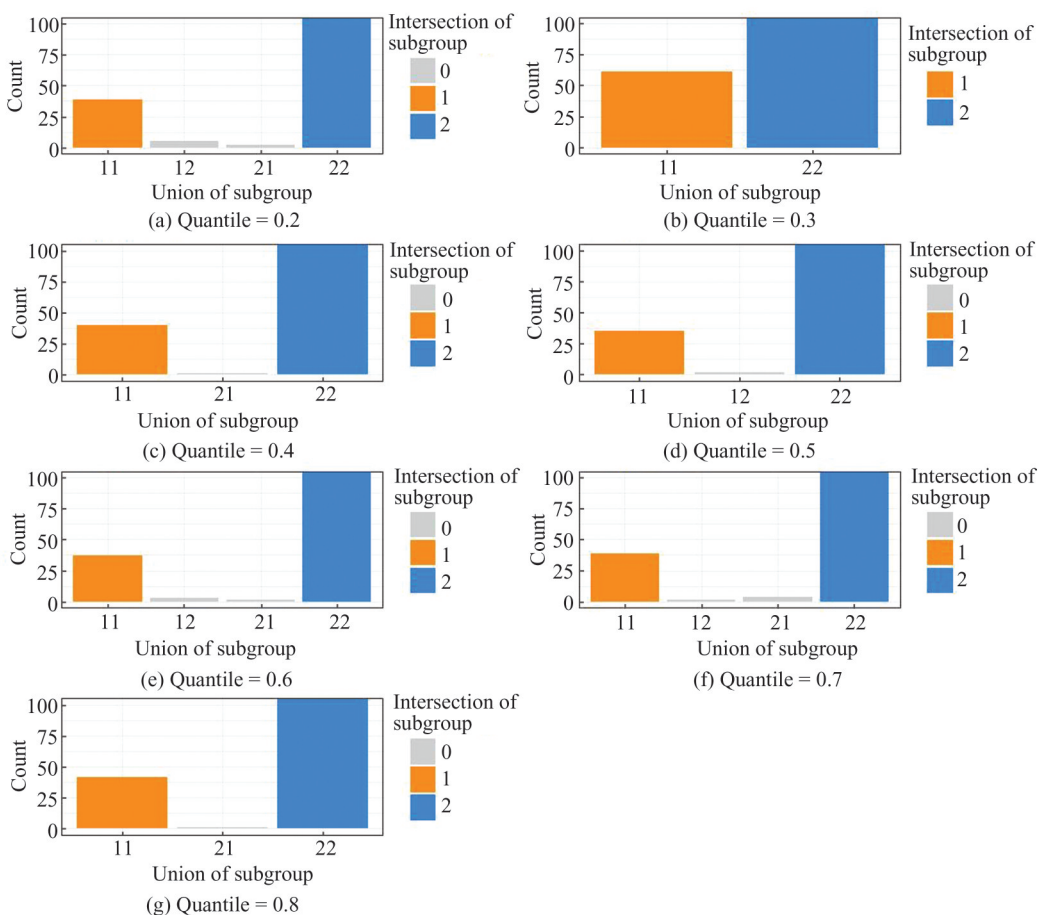


Fig. 4 Schematic diagram of subgroup overlap for two different settings

4 Conclusion

This article considers a single-index threshold quantile regression for subgroup analysis, and selects covariates and biomarkers for defining subgroups, an efficient algorithm is proposed to smooth and locally linearly approximate the subgroup index function and non-convex penalty, respectively. Based on pseudo observations, the corresponding estimation problem degenerates into linear quantile regression, and all unknown parameters are iteratively solved through a two-step estimation process. Numerical simulations demonstrate that the proposed algorithm performs well with a moderate number of variables. Analysis of real data from the PA.3 trial shows that it can distinguish patient between treatment sensitive and treatment insensitive subgroups.

However, our model only considers the case of one threshold, so patient can only be divided into two subgroups. Furthermore, multiple thresholds can be introduced to divide subjects into multiple subgroups with different covariate effects, for related studies, please refer to Li *et al.*^[28-29]. In addition, the recognizability constraint in our model requires specifying a threshold variable x_1 with non-zero coefficient. In practical applications, this variable can be determined through professional knowledge. However, when prior knowledge is not available, it may be achieved by extending the score type specification test method designed by Zhang *et al.*^[30] for selecting the non-zero threshold variables. In addition, considering the internal correlation of longitudinal data and conducting variable selection and parameter estimation under certain condition $p > n$ is our future research direction.

References

- [1] Alosh M, Huque M F, Bretz F, *et al.* Tutorial on statistical considerations on subgroup analysis in confirmatory clinical trials[J]. *Statistics in Medicine*, 2017, **36**(8): 1334-1360.
- [2] Sachdev J C, Sandoval A C, Jahanzeb M. Update on precision medicine in breast cancer[J]. *Cancer Treatment and Research*, 2019, **178**: 45-80.
- [3] Su X, Tsai C L, Wang H, *et al.* Subgroup analysis via recursive partitioning[J]. *Journal of Machine Learning Research*, 2009, **10**(2): 141-158.
- [4] Lipkovich I, Dmitrienko A, Denne J, *et al.* Subgroup identification based on differential effect search: A recursive partitioning method for establishing response to treatment in patient subpopulations[J]. *Statistics in Medicine*, 2011, **30**(21): 2601-2621.
- [5] Zhang Z H, Seibold H, Vettore M V, *et al.* Subgroup identification in clinical trials: An overview of available methods and their implementations with R[J]. *Annals of Translational Medicine*, 2018, **6**(7): 122.
- [6] Jiang Z Y, Du C G, Jablensky A, *et al.* Analysis of schizophrenia data using a nonlinear threshold index logistic model[J]. *PLoS One*, 2014, **9**(10): e109454.
- [7] Fan A L, Song R, Lu W B. Change-plane analysis for subgroup detection and sample size calculation[J]. *Journal of the American Statistical Association*, 2017, **112**(518): 769-778.
- [8] Vander Weele T J, Luedtke A R, Vander Laan M J, *et al.* Selecting optimal subgroups for treatment using many covariates[J]. *Epidemiology*, 2019, **30**(3): 334-341.
- [9] He Y, Lin H Z, Tu D S. A single-index threshold Cox proportional hazard model for identifying a treatment-sensitive subset based on multiple biomarkers[J]. *Statistics in Medicine*, 2018, **37**(23): 3267-3279.
- [10] Wei K C, Zhu H C, Qin G Y, *et al.* Multiply robust subgroup analysis based on a single-index threshold linear marginal model for longitudinal data with dropouts[J]. *Statistics in Medicine*, 2022, **41**(15): 2822-2839.
- [11] Cai Y Z, Stander J. Quantile self-exciting threshold autoregressive time series models[J]. *Journal of Time Series Analysis*, 2008, **29**(1): 186-202.
- [12] Galvao Jr A F, Montes-Rojas G, Olmo J. Threshold quantile autoregressive models[J]. *Journal of Time Series Analysis*, 2011, **32**(3): 253-267.
- [13] Lee S, Seo M H, Shin Y. Testing for threshold effects in regression models[J]. *Journal of the American Statistical Association*, 2011, **106**(493): 220-231.
- [14] Su L J, Xu P. Common threshold in quantile regressions with an application to pricing for reputation[J]. *Econometric Reviews*, 2019, **38**(4): 417-450.
- [15] Zhang Y Y, Wang H J, Zhu Z Y. Single-index thresholding in quantile regression[J]. *Journal of the American Statistical Association*, 2022, **117**(540): 2222-2237.
- [16] Tibshirani R. Regression shrinkage and selection via the lasso[J]. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 1996, **58**(1): 267-288.
- [17] Tibshirani R, Saunders M, Rosset S, *et al.* Sparsity and smoothness via the fused lasso[J]. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 2005, **67**(1): 91-108.
- [18] Zou H. The adaptive lasso and its oracle properties[J]. *Journal of the American Statistical Association*, 2006, **101**(476): 1418-1429.
- [19] Friedman J, Hastie T, Tibshirani R. Sparse inverse covari-

- ance estimation with the graphical lasso[J]. *Biostatistics*, 2008, **9**(3): 432-441.
- [20] Fan J Q, Li R Z. Variable selection via nonconcave penalized likelihood and its oracle properties[J]. *Journal of the American Statistical Association*, 2001, **96**(456): 1348-1360.
- [21] Zhang Y, Lian H, Yu Y. Ultra-high dimensional single-index quantile regression[J]. *Journal of Machine Learning Research*, 2020, **21**(224): 1-25.
- [22] Ruppert D, Carroll R J. Theory & methods: Spatially-adaptive penalties for spline fitting[J]. *Australian & New Zealand Journal of Statistics*, 2000, **42**(2): 205-223.
- [23] Horowitz J L. A smoothed maximum score estimator for the binary response model[J]. *Econometrica*, 1992, **60**(3): 505.
- [24] Seo M H, Linton O. A smoothed least squares estimator for threshold regression models[J]. *Journal of Econometrics*, 2007, **141**(2): 704-735.
- [25] Zou H, Li R Z. One-step sparse estimates in nonconcave penalized likelihood models[J]. *Annals of Statistics*, 2008, **36**(4): 1509-1533.
- [26] Moore M J, Goldstein D, Hamm J, *et al.* Erlotinib plus gemcitabine compared with gemcitabine alone in patients with advanced pancreatic cancer: A phase III trial of the National Cancer Institute of Canada Clinical Trials Group[J]. *Journal of Clinical Oncology*, 2007, **25**(15): 1960-1966.
- [27] Shultz D B, Pai J, Chiu W, *et al.* A novel biomarker panel examining response to gemcitabine with or without erlotinib for pancreatic cancer therapy in NCIC clinical trials group PA.3[J]. *PLoS One*, 2016, **11**(1): e0147995.
- [28] Li J L, Jin B S. Multi-threshold accelerated failure time model[J]. *The Annals of Statistics*, 2018, **46**(6A): 2657-2682.
- [29] Li J L, Li Y G, Jin B S, *et al.* Multithreshold change plane model: Estimation theory and applications in subgroup identification[J]. *Statistics in Medicine*, 2021, **40**(15): 3440-3459.
- [30] Zhang L W, Wang H J, Zhu Z Y. Testing for change points due to a covariate threshold in quantile regression[J]. *Statistica Sinica*, 2014, **24**(4): 1859-1877.

基于变量选择的单指标阈值惩罚分位数回归模型的亚组分析

祁辉^{1,2}, 薛雅心³

1. 中南财经政法大学 统计与数学学院, 湖北 武汉 430073

2. 三明学院 信息工程学院, 福建 三明 365004

3. 复旦大学 公共卫生学院生物统计教研室, 上海 200032

摘要: 在临床研究中, 亚组分析可以帮助识别对特定治疗反应较好或较差的患者群体, 提高治疗效果和安全性, 在精准医疗中具有重要意义。为此, 本文考虑了含有多个协变量和生物标志物的纵向数据下的亚组分析方法, 并基于多个生物标志物的线性组合是否超过某一阈值来划分亚组, 根据亚组与暴露变量间的交互作用来评估亚组间的异质效应。利用分位数回归方法更好地刻画响应变量的全局分布, 并施加稀疏性惩罚实现协变量和生物标志物的变量选择。模拟研究验证了本文提出的估计方法在变量选择和参数估计方面的有效性, 最后, 将此方法应用到了PA.3试验的数据分析当中, 进一步说明了本文方法的实用性。

关键词: 纵向数据; 亚组分析; 阈值模型; 分位数回归; 变量选择

□