Model selection in binary regression using nonlocal prior with a practical application

Maytham Saeed Kadhim and Prof. Dr. Rahim Al-Hamzawi

Email: mohammedsaeedkadhim1999@gmail.com

Department of Statistics, College of Administration and Economics, University of Al-Qadisiyah, Iraq.

Abstract

The mater of determine the effective covariates in a binary regression model has got much attention via last year's, Bayesian variables selection approach using nonlocal priors have become more popular. It has a good properties for achieving variables selection and coefficients estimation . In this paper a new Bayesian approach for simplified shotgun stochastic search with screening has been proposed in binary quantile regression . Our model is depend on the inverse Laplace prior distributions for the binary quantile regression parameters. We compared our proposed model with other methods in same filed via simulation approach and real dataset . Our proposed model has accurate to performing comparison with other methods in estimating coefficient and selecting active variables .

Keywords: *Nonlocal Prior , variables selection, penalized regression, shotgun stochastic search.*

Introduction

Binary QR $(BiQR)$ is important special case of QR, which is widely used ingenetics,engineering farming, finance, medicine, and other fields of knowledge.Manski(1975) developed methods to estimate $BiQR$ models within, the traditional framework and Benoit and Van den Poel (2012) propose a Bayesian framework to $BiQR$. Kordas(2002) proposed binary QR for the aim of classification employing Qs.Thestandardbinary quantile regression $(BiQR)$ problem for the θth quantile can be defined as:

$$yi^{*}{}_{i} = \acute{x}_i\,\beta + \epsilon_i \ , \ \theta \in \ (0\,,1)\,, \qquad (1)$$

$$y_i = 1 \quad if \ \ y^{*}{}_{i} \ \ > 0\,, \qquad (2)$$

$$y_i = 0 \ other \ \ wies\,, \qquad (3)$$

where yi is the observed response of the subject determined by the latent unobserved response$y^{*}{}_{i}$ xi is the $p \times 1$ vector of regressors, β is the $p \times 1$,vectorofquantile coefficients to be estimated, and ϵ is the $n \times 1$ vector of errors whose distribution is restricted to have the θth quantile equal to zero. For an overview, we refer to Algamal et al. (2018); Alhamzawi (2015); Benoit et al. (2013); Benoit and Vanden Poel (2012); Bottai et al. (2010); Hashem et al. (2016); Ji et al. (2012);Li and Miu(2010); Rahman and Vossmeyer (2019); Wei et al. (2019).In a way similar to (Benoit et al., 2013; Benoit and Van den Poel, 2012),Binary QR $(BiQR)$ estimation may proceed by the solution to the following minimization problem:

$$min_{\beta} = \sum_{i=1}^{n} \rho\theta \left(y^{*}{}_{i} - g\,(\acute{x}_i\,\beta\,)\right), \qquad (4)$$

Where $g(\acute{x}_i\beta\,) = I\,\{\acute{x}_i\beta \ > 0\}$

A serious challenge in $BiQR$ lies in the identification of the active regressorsin regression. Here, we improve the prediction accuracy of $BiQR$ by proposing thereciprocal LASSO binary quantile regression $(rLASSO - BiQR)$ which has not beenproposed yet, that results from the following regularization problem:

$$min_{\beta} \sum_{i=1}^{n} \rho\theta \left(y^*_i - I \{\acute{x}_i\beta > 0\}\right) = + \lambda \sum_{j=1}^{p} \frac{1}{|\beta_j|} I\{\beta_j \neq 0\}. \quad (5)$$

In this paper, rather than minimizing (1.5), we solve the problem by constructing $SS - QR$ algorithm via a Gibbs sampler which involves constructing a Markov chain having the joint posterior for β as its stationary distribution . The our paper is organized five sections . Second section focused on Binary Quantile Regression with reciprocal LASSO penalty and Simplified Shotgun Stochastic Search with Screening in BiQR . Third section focused on simulation study via three examples. Real dataset has been introduced in section 4. Section 5 interested by conclusions and recommendations.

2. Methods

2.1 Binary Quantile Regression with reciprocal LASSO penalty

In this section, we follow Kozumi and Kobayashi (2011) and use the following mixture representation:

$$\boldsymbol{\epsilon_i = (1 - 20)\acute{u}_i + \sqrt{2\, wizi}}, \quad \mathbf{(6)}$$

where wi and zi are mutually independent, $wi \sim Exp\left(\theta(1-\theta)\right) and\ \ zi \sim N(0\,,1)$. We use the same prior distributions in the previous section. Under each model $\boldsymbol{k}$,the sampling density for the observations is:

$$\boldsymbol{y_n \backslash \beta_k \sim N\, (X_k\beta_k + (1 - 2\theta)\, w, W_n)}, \quad \mathbf{(7)}$$

Where $W_n = diag\ (2w_1\,, \dots .\,, 2w_n)$. Again, we assume the inverse Laplace prior onthe regression coefficients. Then the full conditional distribution of $y^*$is given by:

$$\boldsymbol{y^*_i | y_i\,, \beta\,, \theta \sim N(\acute{x}_i\beta + (\,1 - 2\theta)w_i\ 2w_i) truncated\ at\ the\ left\ by\ 0\ \ if\ y_{i=1}} \quad \mathbf{(8)}$$

$$\boldsymbol{y^*_i | y_i\,, \beta\,, \theta \sim N(\acute{x}_i\beta + (\,1 - 2\theta)w_i\ 2w_i) truncated\ at\ the\ left\ by\ 0\ \ \ if\ y_{i=0}} \quad \mathbf{(9)}$$

2.2 Simplified Shotgun Stochastic Search with Screening in BiQR

Under model **k**, the marginal likelihood of the observations $m_k(y)^*$can be obtained by integrating out $\boldsymbol{\beta k}$, resulting in :

$$\boldsymbol{m_k(y)^* = (4\pi w_i\,)^{-1}\, Q_k\, \exp\{-R^*{}_k \backslash 4\}}\,, \quad \mathbf{(10)}$$

$$R^*{}_k = u' \,(I_n - WX_k (X_k W X_k)^{-1} X'{}_k W)(y^* - (1-2\theta)w) \qquad (11)$$

$$Q^*{}_k = \int \prod_{j=1}^{|k|} \frac{\lambda}{2\beta^2{}_{kj}} \exp\left\{- (\beta_k - \widehat{\beta}_k)' \, \Sigma_k^{*-1} \, (\beta_k - \widehat{\beta}_k)\backslash 4 - \sum_{j=1}^{|k|} \frac{\lambda}{|\beta_{k,j}|}\right\}, \qquad (12)$$

$$\widehat{\beta}_k = (X'{}_k W X_k)^{-1} X'{}_k W(y^* - (1-2\theta)w) \ \Sigma_k^* = (X'{}_k W X_k)^{-1}, \qquad (13).$$

For the hierarchical parameters models under Bayesian approach for the above of posterior distribution . MCMC algorithm is employed to sampling and updating the parameters of our model .

3. Simulation Study

In this part , we will used simulation study to investigating proposed model compared with other existing methods in the same filed; binary regression quantiles denoted by (Binary R Q)**.** Which it is introduced by (Manski, 1975)**.** Bayesian lasso binary quantile regression **denoted by ((**B L Binary QR)**).** Which it is introduced by (Benoit , et al., 2013)**.** Five quantile levels are used ($\tau = 0.15, \tau = 0.35, \tau = 0.55, \tau = 0.75 \; and \; \tau = 0.90$) **.** For each simulation examples the random error is distributed according normal distribution with mean equal zero and variance equal σ^2. In each simulation examples , the our algorithm run by 11000 iterations and the first 3000 iterations were excluded as burn in. The comparison methods are used three criterions are relative mean square error, denoted by (RMSE), Median of mean absolute deviations denoted by (MMAD). Mean absolute error denoted by (MAE)**.** All simulation results were done by using R package.

First Example

In first simulation example, we will used very sparse case as following model:

$y_i = \max\{c = 0, y_i{}^*\}, \quad i = 1,2, \ldots ,100,$

$$, \qquad y_i{}^* = x_i{}'\beta_\tau + \varepsilon_i \qquad 0 < \tau < 1$$

$$\varepsilon_i \sim N(0, \sigma^2).$$

The Nine explanatory variables variables from the standard Uniform (0,1) have been are generated. The true parameters are $\boldsymbol{\beta} = (1,0,0,0,0,0,0,0,0)^t$.

Table 1: Show results of relative mean square error, denoted by (RMSE) , Median of mean absolute deviations (MMAD) and Mean absolute error MAE via averaged over 50 replications

	Comparison Methods		RMSE	MMAD	MAE
		Binary RQ	0.532 (0.334)	0.475 (0.383)	0.473 (0.374)
		B L Binary QR	0.526 (0.435)	0.517 (0.420)	0.427 (0.363)

Quantile levels	$\tau = 0.15$	B nonlocal BQR	0.352 (0.293)	0.246 (0.135)	0.237 (0.182)
	$\tau = 0.35$	Binary RQ	0.431 (0.371)	0.434 (0.273)	0.263 (0.223)
		B L Binary QR	0.345 (0.233)	0.371 (0.231)	0.447(0.333)
		B nonlocal BQR	0.234 (0.118)	0.264 (0.137)	0.227 (0.094)
	$\tau = 0.55$	Binary RQ	0.528 (0.394)	0.506 (0.427)	0.483 (0.384)
		B L Binary QR	0.496 (0.346)	0.537 (0.386)	0.383 (0.218)
		B nonlocal BQR	0.337 (0.248)	0.293 (0.163)	0.237 (0.173)
	$\tau = 0.75$	Binary RQ	0.531 (0.428)	0.592 (0.438)	0.571 (0.442)
		B L Binary QR	0.530 (0.471)	0.581(0.436)	0.427(0.337)
		B nonlocal BQR	0.382 (0.234)	0.333(0.219)	0.241(0.182)
	$\tau = 0.90$	Binary RQ	0.524 (0.461)	0.430 (0.320)	0.418 (0.371)
		B L Binary QR	0.487 (0.316)	0.464 (0.328)	0.438 (0.373)
		B nonlocal BQR	0.318 (0.186)	0.252 (0.173)	0.222 (0.162)

Note: In the parentheses are SDs of the MAD

From the results are showed in table 1. We see clearly the ,our proposed method is best compared with other existing methods via all quantile levels . The RMSE,MMAD,MAE and SD are generated by our proposed method (**B nonlocal BQR**) smallest from RMSE,MMAD,MAE and SD are generated by other methods . From these results ,we can judge **B nonlocal BQR** method is very efficient in variables selection an parameter estimation in binary quantile regression model .

Second example

In second simulation example, we will used sparse case as following model:

$$y_i = \max\{c = 0, {y_i}^*\}, \quad i = 1,2,\dots,100,$$

$$, \qquad \boldsymbol{{y_i}^* = {x_i}'\beta_\tau + \varepsilon_i} \qquad \boldsymbol{0 < \tau < 1}$$

$\varepsilon_i \sim N(0, \sigma^2)$.

The Nine explanatory variables variables from the standard Uniform (0,1) have been are generated. The true parameters are $\boldsymbol{\beta} = (1,0,0,3,0,0,1.5,0,0)^t$.

Table 2: Show results of relative mean square error, denoted by (RMSE) , Median of mean absolute deviations (MMAD) and Mean absolute error MAE via averaged over 50 replications

	Comparison Methods		RMSE	MMAD	MAE
Quantile levels	$\tau = 0.16$	Binary RQ	0.643 (044.6)	0.676 (04.84)	0.674 (0.476)
		B L Binary QR	0.636 (0.646)	0.617 (0.630)	0.637 (0.464)
		B nonlocal BQR	**0.363 (0.294)**	**0.266 (0.146)**	**0.247 (0.153)**
	$\tau = 0.46$	Binary RQ	0.641 (0.471)	0.646 (0.374)	0.364 (0.334)
		B L Binary QR	0.466 (0.344)	0.471 (0.341)	0.667(0.444)
		B nonlocal BQR	**0.246 (0.118)**	**0.266 (0.147)**	**0.237 (0.096)**
	$\tau = 0.66$	Binary RQ	0.638 (0.496)	0.606 (0.637)	0.684 (0.486)
		B L Binary QR	0.696 (0.466)	0.647 (0.486)	0.484 (0.318)
		B nonlocal BQR	**0.247 (0.168)**	**0.294 (0.164)**	**0.247 (0.174)**
	$\tau = 0.76$	Binary RQ	0.641 (0.638)	0.693 (0.648)	0.671 (0.663)
		B L Binary QR	0.640 (0.671)	0.681(0.646)	0.637(0.447)
		B nonlocal BQR	**0.383 (0.246)**	**0.344(0.119)**	**0.261(0.183)**
	$\tau = 0.90$	Binary RQ	0.636 (0.661)	0.640 (0.430)	0.618 (0.471)
		B L Binary QR	0.687 (0.416)	0.666 (0.438)	0.648 (0.474)
		B nonlocal BQR	**0.218 (0.186)**	**0.263 (0.174)**	**0.233 (0.163)**

Note: In the parentheses are SDs of the MAD

From the results are showed in table 2. We see clearly the ,our proposed method is best compared with other existing methods via all quantile levels . The RMSE,MMAD,MAE and SD are generated by our proposed method (**B nonlocal BQR**) smallest from RMSE,MMAD,MAE and SD are generated by other methods . From these results ,we can judge **B nonlocal BQR** method is very efficient in variables selection an parameter estimation in binary quantile regression model **.**

Third example

In third simulation example, we will used denste case as following model:

$y_i = \max\{c = 0, y_i{}^*\}$, $i = 1,2, \dots ,100$,

$$, \quad \boldsymbol{y_i{}^* = x_i{}'\beta_\tau + \varepsilon_i} \qquad \boldsymbol{0 < \tau < 1}$$

$\varepsilon_i \sim N(0, \sigma^2)$.

The Nine explanatory variables variables from the standard Uniform (0,1) have been are generated. The true parameters are $\boldsymbol{\beta}$ = (0.85, 0.85, 0.85, 0.85, 0.85, 0.85, 0.85, 0.85, 0.85)^t .

Table 3: Show results of relative mean square error, denoted by (RMSE) , Median of mean absolute deviations (MMAD) and Mean absolute error MAE via averaged over 50 replications

	Comparison Methods		RMSE	MMAD	MAE
Quantile levels	$\tau = 0.16$	Binary RQ	0.713 (0.337)	0.777 (0. 483)	0.773 (0.477)
		B L Binary QR	0.737 (0.487)	0.717 (0.430)	0.737 (0.573)
		B nonlocal BQR	**0.373 (0.193)**	**0.477 (0.107)**	**0.317 (0.093)**
	$\tau = 0.46$	Binary RQ	0.731 (0.571)	0.737 (0.473)	0.373 (0.233)
		B L Binary QR	0.677 (0.163)	0.671 (0.631)	0.567(0.336)
		B nonlocal BQR	**0.367 (0.168)**	**0.376 (0.137)**	**0.367 (0.097)**
	$\tau = 0.66$	Binary RQ	0.738 (0.467)	0.707 (0.537)	0.683 (0.587)
		B L Binary QR	0.697 (0.577)	0.837 (0.687)	0.583 (0.318)
		B nonlocal BQR	**0.317 (0.178)**	**0.393 (0.373)**	**0.337 (0.173)**
	$\tau = 0.76$	Binary RQ	0.731 (0.738)	0.793 (0.738)	0.671 (0.773)
		B L Binary QR	0.630 (0.771)	0.781(0.737)	0.737(0.337)
		B nonlocal BQR	**0.383 (0.137)**	**0.347(0.119)**	**0.351(0.133)**
	$\tau = 0.90$	Binary RQ	0.637 (0.471)	0.630 (0.530)	0.618 (0.471)
		B L Binary QR	0.687 (0.417)	0.767 (0.538)	0.768 (0.573)
		B nonlocal BQR	**0.318 (0.187)**	**0.373 (0.173)**	**0.366 (0.173)**

Note: In the parentheses are SDs of the MAD

From the results are showed in table 2. We see clearly the ,our proposed method is best compared with other existing methods via all quantile levels . The RMSE,MMAD,MAE and SD are generated by our proposed method (**B nonlocal BQR**) smallest from RMSE,MMAD,MAE and SD are generated by other methods . From these results ,we can judge **B nonlocal BQR** method is very efficient in variables selection an parameter estimation in binary quantile regression model .

Real Dataset

Children Cancer Diseases

This dataset collected from Children's Specialist Hospital in Basrah city. This dataset contain one response variable (take chemotherapy dose or no) and 12 explanatory variables are age (x_1), Gender (x_2), The number of sisters and brothers (x_3), Weight (x_4), Height (x_5)**,** Body mass index (BMI) (x_6), Liver disease (x_7), Kidney disease(x_8), Family History (x_9), disease diagnosis(x_{10}), Father's age(x_{11}), mother's age(x_{12}), number of birth(x_{13}) , pregnancy duration for child (x_{14}) , Breastfeeding type (x_{15})

As like section, simulation study.in this section , we will compared two

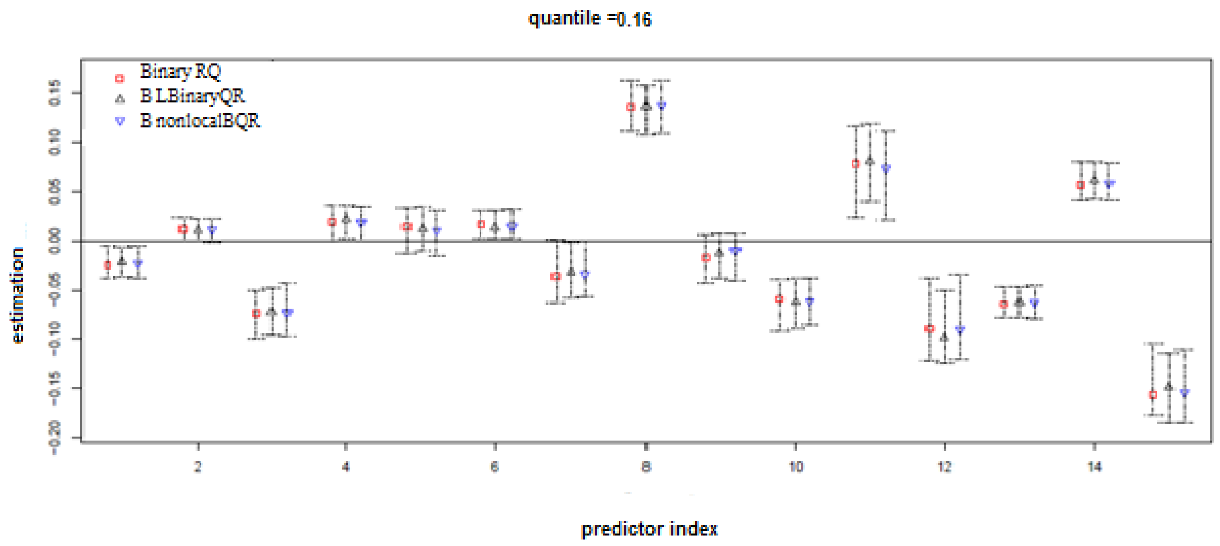
methods **(Binary RQ,** B L Binary QR**)** with our proposed method (**B nonlocal BQR**), the methods under this study are evaluated through three criterions (MSE,MAE and SD).

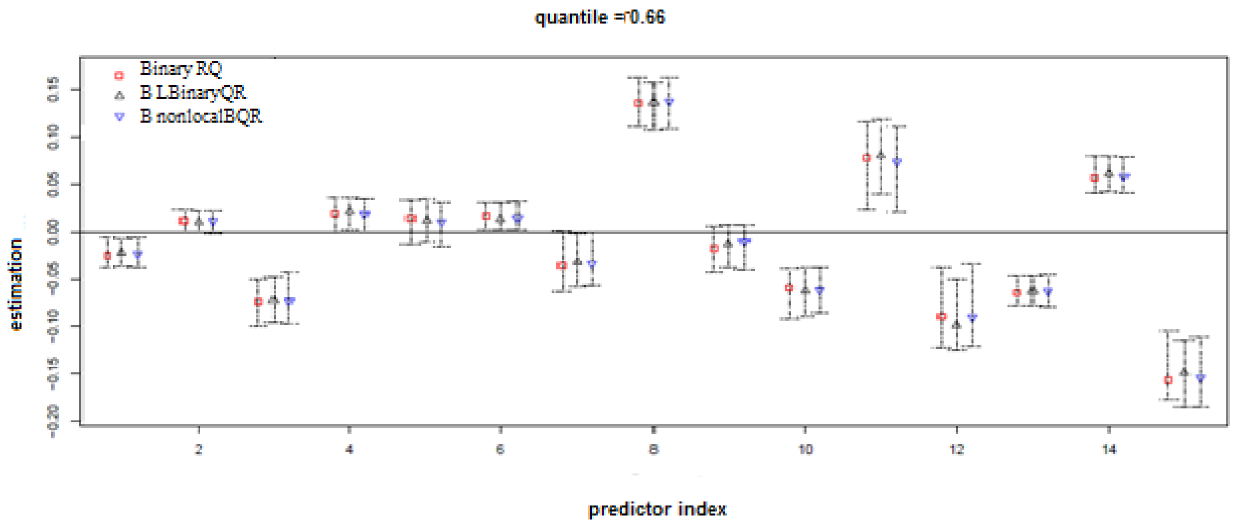
Table 4: MSEs ,MMAD,MAE and standard deviations (SD)for dataset of children cancer disease

Comparison Methods		RMSE	MMAD	MAE
$\tau = 0.16$	Binary RQ	0.412 (0.334)	0.424 (0.343)	0.483 (0.374)
	B L Binary QR	0.435 (0.364)	0.426 (0.340)	0.466 (0.333)
	B nonlocal BQR	**0.242 (0.173)**	**0.245 (0.184)**	**0.236 (0.192)**
$\tau = 0.46$	Binary RQ	0.481 (0.341)	0.474 (0.243)	0.243 (0.233)
	B L Binary QR	0.364 (0.293)	0.311 (0.221)	0.486(0.363)
	B nonlocal BQR	**0.283 (0.126)**	**0.248 (0.182)**	**0.282 (0.154)**
$\tau = 0.66$	Binary RQ	0.453 (0.353)	0.441 (0.326)	0.543 (0.363)
	B L Binary QR	0.439 (0.314)	0.488 (0.325)	0.352 (0.204)
	B nonlocal BQR	**0.286 (0.144)**	**0.262 (0.139)**	**0.229 (0.159)**
$\tau = 0.76$	Binary RQ	0.431 (0.424)	0.452 (0.434)	0.461 (0.442)
	B L Binary QR	0.430 (0.461)	0.441(0.435)	0.426(0.336)
	B nonlocal BQR	**0.242 (0.234)**	**0.233(0.215)**	**0.241(0.142)**
$\tau = 0.90$	Binary RQ	0.424 (0.451)	0.430 (0.320)	0.414 (0.361)
	B L Binary QR	0.446 (0.315)	0.454 (0.324)	0.434 (0.363)
	B nonlocal BQR	**0.214 (0.145)**	**0.242 (0.163)**	**0.222 (0.152)**

From the results are listed in Table 8, the MSE,MMAD,MAE and SD generated by our proposed method is much smaller than MSE,MMAD,MAE and SD generated by others methods **(**Binary RQ**,** B L Binary QR**)**. This means, our proposed method has performance better than **(**B L Binary QR**,** B L Binary QR**)** via all quantile level.

The our proposed method is exceled on other existing method through variables selection and parameters estimation ,this clear from the following figures.





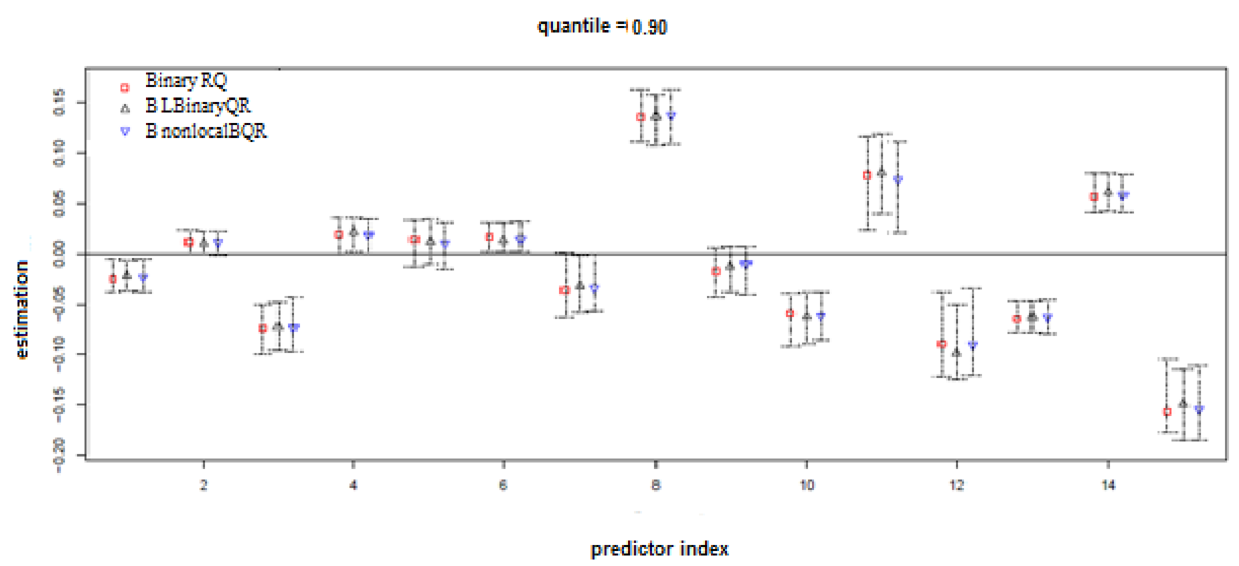


Figure -1 show plot of confidence interval for three methods via quantile levels (0.16,0.66 and 0.90)

From the figure 1 we see our proposed method has good performance to parameter estimation and variable selection. Also we see, there are six variables not effect from dataset are (variable 2,variable 4, variable 5, variable 6, variable 7and variable 8). Because it is possible that the estimations of these six variables are equal to zero.

5. Conclusion and Discussion

In this our paper, we suggest a new hierarchical prior of Bayesian reciprocal LASSO binary quantile regression by employing inverse Laplace prior distributions density.

The our algorithm for Bayesian reciprocal LASSO binary quantile regression with efficient and tractable to full posterior distributions. Simulation examples show that the our Gibbs sampler is easy and e effective for parameters estimation and variables selection in binary quantile regression under a variety of example. Also we conclude the proposed method is very active with real Dataset.

We can extend the Suggested to another study such as, Bayesian reciprocal LASSO Tobit regression, Bayesian reciprocal LASSO composite Tobit quantile regression and Bayesian reciprocal LASSO composite binary quantile regression.

Bibliography

[1] Algamal, Z. Y., R. Alhamzawi, and H. T. M. Ali (2018). Gene selection for microarray gene expression classification using bayesian lasso quantile regression. Computers in biology and medicine 97, 145{152.

[2] Alhamzawi, R. (2015). Model selection in quantile regression models. Journal of Applied Statistics 42(2), 445{458.

[3] Benoit, D. F., R. Alhamzawi, and K. Yu (2013). Bayesian lasso binary quantile regression. Computational Statistics 28(6), 2861{2873.

[4] Benoit, D. F. and D. Van den Poel (2012). Binary quantile regression: a bayesian approach based on the asymmetric laplace distribution. Journal of Applied Econometrics 27(7), 1174{1188.

[5] Bottai, M., B. Cai, and R. E. McKeown (2010). Logistic quantile regression for bounded outcomes. Statistics in medicine 29(2), 309{317.

[6] Hashem, H., V. Vinciotti, R. Alhamzawi, and K. Yu (2016). Quantile regression with group lasso for classification. Advances in Data Analysis and Classification 10(3),375{390.

[7] Ji, Y., N. Lin, and B. Zhang (2012). Model selection in binary and tobit quantile regression using the gibbs sampler. Computational Statistics & Data Analysis 56(4),827{839.

[8] Kordas, G. (2002). Credit scoring using binary quantile regression. In Statistical data analysis based on the l1-norm and related methods, pp. 125{137. Springer.

[9] Kozumi, H. and G. Kobayashi (2011). Gibbs sampling methods for bayesian quantile regression. Journal of statistical computation and simulation 81 (11), 1565{1578.

[10] Li, M.-Y. L. and P. Miu (2010). A hybrid bankruptcy prediction model with dynamic loadings on accounting-ratio-based and market-based information: A binary quantile regression approach. Journal of Empirical Finance 17(4), 818{833.

[11] Manski, C. F. (1975). Maximum score estimation of the stochastic utility model of choice. Journal of econometrics 3(3), 205{228.

[12] Rahman, M. A. and A. Vossmeyer (2019). Estimation and applications of quantile regression for binary longitudinal data. In Topics in Identification, Limited Dependent Variables,

Partial Observability, Experimentation, and Flexible Modeling: Part B. Emerald Publishing Limited.

.