Available online at www.qu.edu.iq/journalcm
JOURNAL OF AL-QADISIYAH FOR COMPUTER SCIENCE AND MATHEMATICS
ISSN:2521-3504(online) ISSN:2074-0204(print)

Lasso Quantile Principal Component Regression

***Mohammed H. Al-Sharoot*[a]*, Fatimah K. Mohammed*[b] *, Hameedah N. Mayali*[c]**

[a] *Department of Statistics, College of Administration and Economics University of AL-Qadisiyah, IRAQ. Email: mohammed.Alsharoot@qu.edu.iq*

[b]*Education Directorate Babylon, Ministry of Eduucation, IRAQ .Email: stat.post14@qu.edu.iq*

[c]*Department of Statistics, College of Administration and Economics University of AL-Qadisiyah, IRAQ. Email: hameedah.naeem@qu.edu.iq*

ARTICLE INFO

Article history:
Received: 2 /12/2023
Rrevised form: 19 /12/2023
Accepted : 25 /12/23
Available online: 30 /12/23

Keywords: Quantile Regression , principal component , lasso, lasso quantile principal component.

Each keyword to start on a new line

ABSTRACT

The classical regression model is very sensitive to econometrics problems, one this econometrics problem is Multicollinearity, to overcome this problem ,we will use two solutions: Firstly via using principal component regression and second solution via using quantile regression. When mix between these methods together give as robust model against the Multicollinearity problem. The simulation scenario and real data using in this study.

MSC..

https:// 10.29304/jqcsm.2023.15.41356

1. Introduction

In empirical applications, regression is by far the most often used statistical methodology in a variety of filed , from the social and economic sciences to the ecological sciences. agricultural economics sciences and so on. Many regression models have been proposed , each model deal with different types of data, in order to provide as ease interpretation and the availability of tools and strategies suitable to and validity of assumptions. But sometime, these assumption is not satisfied, especially, with the era of big data, in other hand, the used of numerous independent variables give us a more accurate view on the dependent variable, but engenders redundant information deriving because of the correlation between independent variables . The Multicollinearity between independent variables is one of the main problems come with multiple linear regression, this problem is affects accuracy estimation of regression parameters , standard errors, explanation accuracy. there are several are the proposals to address the problem. such as partial least squares regression ,principal component regression and ridge regression.

∗Corresponding author Hameedah N. Mayali

Email addresses: hameedah.naeem@qu.edu.iq

Communicated by 'sub etitor'

The quantile regression models have a good properties compared with other regression models. quantile regression models are belong to a robust regression models family (Koenker and Geling, (2001). quantile regression models does not require any supposition about the random error distribution. But quantile regression models are not robust against the econometrics problems. To overcome these econometrics problems via principal component method associated with quantile regression model. Our contribution in this paper , we use lasso quantile principal component regression ,which it reduced the many components to few them, which it have high explanatory power .

The our paper is organized as follows: The Multicollinearity have been offered in section 2 . In section 3, Offers Lasso quantile principal component regression. In section 4 Offers Simulation approach with two simulation examples and real dataset. Brief the conclusions and recommendations have been presented in section 5.

2. General Concept of Multicollinearity

In the analysis of multiple linear regression, Multicollinearity has been a serious problem, the estimation methods may result in high variance in the estimates of the regression coefficients in the existence of Multicollinearity. When two or more independent variables are perfectly or highly correlated, a multiple regression model suffers from Multicollinearity, which it is a significant issue. There are many reasons for the occurrence of the Multicollinearity problem, The tendency of some economic variables to change due to the time factor, Using some time-lagging variables as explanatory variables in the model. It have been effecting on regression model , with the Multicollinearity problem ,the standard errors of parameters are high may be infinite and therefore, the confidence intervals are high to population parameters, because of the parameters variance also high and infinite . Also with the Multicollinearity problem, the ordinary least square estimation was unbiased estimated. Some parameters estimation are appear with different signs . Also with the Multicollinearity problem, increases the probability of making a type II error . Also with the Multicollinearity problem, the value of determination coefficient R^2 and is fake. Also with the Multicollinearity problem, the $t-$ test of regression coefficients are unclear. There are many methods to treatment this problem , Some of them is classical methods ,and others methods are depend on statistical theory.

3. Lasso Quantile Principal Component Regression

3.1. Principal Component Regression

The multiple regression model is focus on estimation the relationship among with on dependent variable and a set of independent variables as according the formula(Eberly, L. E 2007):

$$y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon_i, \qquad (1)$$

where y_i , $(i = 1,2 \ldots \ldots n)$ is dependent variable ,$(X_1, X_{2,} \ldots X_k)$ are independent variables $(\beta_0, \beta_1, \beta_2 \ldots \ldots \beta_k)$ are unknown parameters , ε_i, $(i = 1,2 \ldots . . n)$ random error term is distributed as normal distribution with mean (0) and variance (σ^2) $\varepsilon_i \sim N(0, \sigma^2)$. The model showed in equation (1) may be suffered from many econometrics problem, the Multicollinearity one of these econometrics problem. To overcome this problem principal component regression model (P.C.Reg) have been used (Ergon, R. et al (2014)).

The P.C.Reg is focus on transformation of correlated independent variables to orthogonal compound (uncorrelated) as following

$$P.C = X\theta \qquad (2)$$

where(P. C) is matrix of principal component have degree (n. k) , that it full rank . θ is orthogonal matrix from standard characteristic vectors that corresponding the Eigen factor for information matrix (X^tX) ,that the matrix θ have degree (k. k) and its row γ_i (i = 1, … … . . n) and its column θ_j(j=1….,k). this matrix is necessary to making information matrix (X^tX) diagonal matrix , under the assumption that Eigen value for(X^tX)(sutter,j.M et al (1992).

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_q$$

We can reformulation the model in equation (1) via principal component (P.C) where the dependent variable (y_i) is become as function of orthogonal (P.C) instead of independent variables (X_s),it interconnected among themselves. We know the matrix θ is orthogonal ($\theta^t\theta = \theta\theta^t = 1$).

$$(P.C = X\theta) \ * \theta^t$$

$$(P.C)\theta^t = X\theta\theta^t$$

$$X = (P.C)\theta^t$$

$$y_i = (P.C)\theta^t\beta + \varepsilon_i,$$

Let $\theta^t\beta = \varphi$ the above model is become the following formula :

$$y_i = (P.C)\varphi + \varepsilon, \qquad (3\,)$$

The model in equation (3) is represented the relationship between dependent variable (y_i) and principal component (P.C). The model in equation (3) linked with quantile regression.

3.2. Lasso Quantile Principal Component Regression

Koenker and Bassett (1978) are introduced attractive regression model named quantile regression model , it is applied in many sciences such as Microarray study agricultural economics (Kostov and Davidova,(2013) , (Wang and He, (2007), ecological studies (Cade and Noon,(2003), and so on. The quantile regression is suitable with many applied study because, it is not requires any suppositions compared with classical regression model and it is very robust against outliers dataset. The model of quantile regression can be written as:

$$y_i = x_i^T\beta_\tau + \varepsilon_i, \qquad \tau \in (0,1), \qquad (4)$$

where y_i is dependent variable, where x_i^T is a $1 \times k$ of independent variables , β_τ is a $k \times 1$ of unknown parameters vector and τ is the quantile level.

when mix the model in equation (3) and equation (4) , we will obtaining the Quantile Principal Component Regression (P.C.Q.Reg) can be written as:

$$y_i = (P.C)\varphi_\tau + \varepsilon, \qquad (5\,)$$

where y_i is independent variable (P. C) is the principal component φ is unknown parameters of model in (5) , τ is quantile level $0 < \tau < 1$. That the coefficients of $\varphi_\tau$which are belong to P.C.Q.Reg model can be estimated by:

$$\min_{\boldsymbol{\varphi_\tau}} \ \sum_{i=1}^{n} \boldsymbol{\rho_\tau}(\mathbf{y_i} - (\mathbf{P.C})^{\mathbf{t}}\boldsymbol{\varphi_\tau}) \qquad (\mathbf{6})$$

Since check(loss) function show in equation (6) is not differentiable at (0) point ,see below figure. But (Koenker, (2005) explain the minimization of (6) via used a linear programming algorithm (Koenker and D'Orey, (1987).

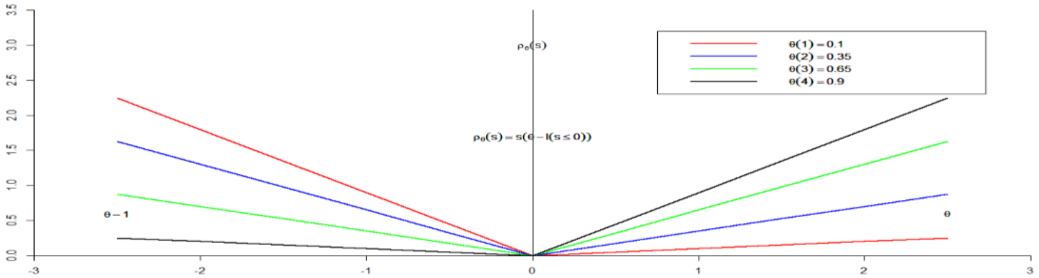

Fig. 1 - The check(loss) function at $\theta(1) = 0.10$ (red line), $\theta(2) = 0.35$ (blue line), $\theta(3) = 0.65$ (green line) and $\theta(4) = 0.65$ (black line)

One significant issue in building a regression models are the selection of the active independent variables. The variable selection procedure help to exclude unimportant independent variables this reflected on the forecasting accuracy . And it help to getting a good interpretation about important independent variables (Alhamzawi et al 2013). Newly, there has been big attention on case of sparse methods to shrink inefficient regression coefficients toward zero exactly. For example, Lasso (Tibshirani, (1996)). The lasso quantile principal component quantile (L.P.C.Q.Reg) model can be estimated by:

$$\min_{\varphi_\tau} \sum_{i=1}^{n} \rho_\tau (y_i - (P.C)^t \varphi_\tau) + \lambda \|\varphi\| \qquad (7)$$

Where $\lambda(\lambda \geq 0)$ is the shrinkage parameter. Lasso quantile principal components regression model can be estimated through building a good algorithm, after taking the mean for thousands.

4. Simulation study

The performance of the our proposed method lasso quantile principal component quantile referred to as (L.P.C.Q.Reg) is inspected by simulations scenario. L.P.C.Reg is compared with a set of existing methods.

First method that proposed by (Davino,C (2022)) "Handling Multicollinearity in quantile regression through the use of principal component regression" referred to as (Q.P.C Reg).

And second method is proposed by Kyung, M. (2021) "Bayesian analysis of quantile principal component regression model" referred to as (B.Q.P.C Reg).

We employed three quantile levels that are ($\tau = 0.15, \tau = 0.55$ and $\tau = 0.95$). For each simulation scenario, and also five distributions of random error term have been used . $\varepsilon_i \sim N(0,1)$, and$\varepsilon_i \sim$standard Laplace (0,1): and we will used five different correlation level ($\rho = 0.66, \rho = 0.70, \rho = 0.86, \rho = 0.90$ and $\rho = 0.95$)For each simulation, we run 11000 iterations and the first 1000 iterations were removed as burn in, we run one hundred simulations (r). Root mean square error (RMSE) have been used ,it colucleted by the following formulation , where

In our paper , we will used two examples of simulation scenario.RMSE=$\sqrt{\frac{(x^T\hat{\beta}\text{-}x^T\beta^{true})^2}{r}}$.

4.1 Simulation Example One

In this simulation scenario, we study the performance of the our proposed method with case of very sparse models. In particular, we consider the true models as in follows:

$y_i = x_{1i} + + \varepsilon_i$, i=1,2,....100

We simulate seven explanatory variables from a multivariate normal distribution with mean zero and $\text{cov}(x_i, x_j) = 0.5^{|i-j|}$. Thus the true coefficients of independent varibles, including the intercept term, are $\beta = (1,0,0,0,0,0,0)$

The results that listed in table -1 - we see the root mean square error (RMSE) is computed via our proposed method (L.P.C.Q.Reg) is much smaller than root mean square error (RMSE) is computed by other two methods (Q.P.C Reg) and (BQ.P.C Reg). This mean theour proposed method have a good performance compared with other method.

Table 1- Root Mean Square Error for the simulation studies for Simulation Example One.

	$\boldsymbol{\varepsilon_i \sim N(0,1)}$				
	$\boldsymbol{\tau = 0.15}$				
	$\boldsymbol{\rho = 0.66}$	$\boldsymbol{\rho = 0.70}$	$\boldsymbol{\rho = 0.86}$	$\boldsymbol{\rho = 0.90}$	$\boldsymbol{\rho = 0.95}$
Comparison Methods	RMSE	RMSE	RMSE	RMSE	RMSE
Q.P.C Reg	1.562	1.313	0.979	0.983	0.832
BQ.P.C Reg	1.313	1.259	0.892	0.884	0.742
L.P.C.Q.Reg	1.114	1.131	0.872	0.829	0.644
	$\tau = 0.55$				
Q.P.C Reg	1.403	1.389	0.986	0.884	0.784
BQ.P.C Reg	1.305	1.374	0.951	0.852	0.763
L.P.C.Q.Reg	1.323	1.223	0.905	0.844	0.705
	$\tau = 0.95$				
Q.P.C Reg	1.236	1.363	0.882	0.572	0.552
BQ.P.C Reg	1.234	1.114	0.779	0.565	0.532
L.P.C.Q.Reg	1.107	1.092	0.683	0.537	0.527
	$\varepsilon_i \sim$ standard Laplace (0,1)				
	$\tau = 0.15$				
Q.P.C Reg	0.987	0.923	0.822	0.641	0.575
BQ.P.C Reg	0. 863	0.912	0.801	0.626	0.565
L.P.C.Q.Reg	0.894	0.893	0.759	.0538	0.533
	$\tau = 0.55$				
Q.P.C Reg	0.752	0.735	0.581	0.530	0.552
BQ.P.C Reg	0.713	0.621	0.536	0.514	0.522
L.P.C.Q.Reg	0.652	0.635	0.519	0.473	0.482
	$\tau = 0.95$				
Q.P.C Reg	0.659	0.671	0.523	0.520	0.532
BQ.P.C Reg	0.623	0.633	0.492	0.513	0.491
L.P.C.Q.Reg	0.592	0.525	0.449	0.472	0.453

4.2 Simulation Example two

In this simulation scenario, we study the performance of the our proposed method with case of dense models. In particular, we consider the true models as in follows

$$y_i = 0.85x_{1i} + 0.85x_{2i}+0.85x_{3i} + 0.85x_{4i} + 0.85x_{5i} + 0.85x_{6i} + 0.85x_{7i} + \varepsilon_i, \qquad i=1,2,....100$$

We simulate seven explanatory variables from a multivariate normal distribution with mean zero and $\mathrm{cov}(x_i, x_j) = 0.5^{|i-j|}$. Thus the true coefficients of independent varibles, including the intercept term, are $\beta = (0.85,0.85,0.85,0.85,0.85,0.85,0.85)$

The results that listed in table -2 - we see the root mean square error (RMSE) is computed via our proposed method (L.P.C.Q.Reg) is much smaller than root mean square error (RMSE) is computed by other two methods (Q.P.C Reg) and (BQ.P.C Reg). This mean the our proposed method have a good performance compared with other method .

Table 2- Root Mean Square Error for the simulation studies for Simulation Example two .

			.$\boldsymbol{\varepsilon_i \sim N(0,1)}$		
			$\boldsymbol{\tau = 0.15}$		
	$\rho = 0.66$	$\rho = 0.70$	$\rho = 0.86$	$\rho = 0.90$	$\rho = 0.95$
Comparison Methods	RMSE	RMSE	RMSE	RMSE	RMSE
Q.P.C Reg	0.951	0.837	0.712	0.702	0.687
BQ.P.C Reg	0.877	0.829	0.713	0.687	0.653
L.P.C.Q.Reg	0.883	0.737	0.613	0.592	0.564
			$\tau = 0.55$		
Q.P.C Reg	0.964	0.896	0.834	.0773	0.755
BQ.P.C Reg	0.941	0.898	0.822	0.748	0.713
L.P.C.Q.Reg	0.865	0.764	0.631	0.683	0.654
			$\tau = 0.95$		
Q.P.C Reg	0.873	0.834	0.687	0.674	0.562
BQ.P.C Reg	0.788	0.822	0.619	0.613	0.586
L.P.C.Q.Reg	0.742	0.631	0.580	0.564	0.549
			$\varepsilon_i \sim$standard Laplace (0,1)		
			$\tau = 0.15$		
Q.P.C Reg	0.708	0.687	0.662	0. 642	0. 625
BQ.P.C Reg	0.696	0.619	0.647	0.635	0.610
L.P.C.Q.Reg	0.576	0.580	0.507	0.494	0.469
			$\tau = 0.55$		
Q.P.C Reg	0.704	0.662	0.517	0. 485	0. 457
BQ.P.C Reg	0.622	0.647	0.409	0.394	0.386
L.P.C.Q.Reg	0.424	0.507	0.399	0.381	0.368
			$\tau = 0.95$		
Q.P.C Reg	0.579	0.592	0.519	0.478	0.469
BQ.P.C Reg	0.441	0.372	0.503	0.384	0.351
L.P.C.Q.Reg	0.352	0.273	0.469	0.329	0.250

4.3 Real Dataset

In this section, the air Pollution Data within package in (bayesQR) package in R programing (Lindmark and Karlsson, (2011)) have been used to evaluating the performance of our proposed method compared with other method. the air Pollution Data was measured through the Public Roads Administration in Norway. It consists of one dependent variable referred to as (y_i)(the log (concentration of NO2 per hour)) and seven independent variables are: temperature referred to as (x_1), the temperature difference (x_2), the log (number of cars per hour) referred to as (x_3),), wind speed in meters per second (x_4), wind direction (x_5), day number referred to as (x_6) and time of day in hours referred to as (x_7), in this section we compare three methods : (Q.P.C Reg, BQ.P.C Reg, and our proposed approach). These methods are evaluated by depend on the Root mean squared error (RMSE) and standard deviations (SD). The result of (RMSE and SD) listed in table 3 are computed via three quantile levels ($\tau = 0.15, \tau = 0.55$ and $\tau = 0.95$) as ,it show in table 3

Table -3 Standard deviations (SD) and RMSEs for the air pollution data

Methods	τ = 0.15	τ = 0.55	τ = 0.95
			RMSE (SD)
	RMSE (SD)	RMSE (SD)	
Q.P.C Reg	0.471 (0.396)	0.466 (0.394)	0.435 (0.418)
BQ.P.C Reg	0.454 (0.423)	0.424 (0.363)	0.417 (0.384)
L.P.C.Q.Reg	0.395 (0.355)	0.375 (0.346)	0.368 (0.292)

From the rsults showed in the table 3, we see the the root mean square error (RMSE) is computed via our proposed method (L.P.C.Q.Reg) is much smaller than root mean square error (RMSE) is computed by other two methods (Q.P.C Reg) and (BQ.P.C Reg). This mean the our proposed method have a good performance compared with other method ,also with real data. We see the the Standard deviations (SD) is computed via our proposed method (L.P.C.Q.Reg) is much smaller than Standard deviations (SD) is computed by other two methods (Q.P.C Reg) and (BQ.P.C Reg). Also, This mean the our proposed method have a good performance compared with other method. From all via all criteria used in the current study , we concluded the our proposed method have a good performance compared with other methods . We can estimating and 95% intervals for the three quantile level 0.15, 0.55 and 0.95 by the following figures .

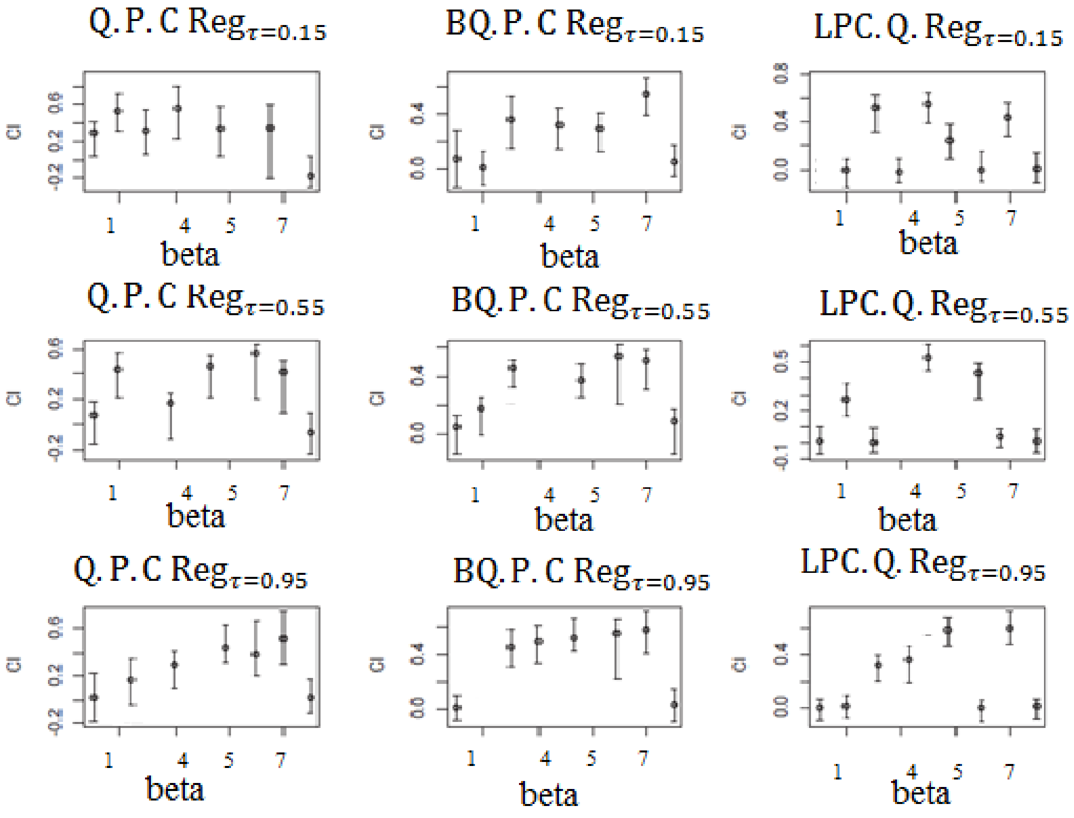

Figure.2-
Show the confidence interval for estimating at 95% intervals for the three quantile level 0.15, 0.55 and 0.95 and three methods under study

From the above figure we see our proposed method (L.P.C.QReg) is a good performance to coefficient estimation and variable selection compared with other method. In LPC. Q. Reg$_{\tau=0.15}$,we see the independent variables (x_1, x_3, x_5, x_9) are excluded from our model, because of it coefficients estimation closed from zero exactly. In LPC. Q. Reg$_{\tau=0.55}$,we see the independent variables (x_1, x_3, x_6, x_7) are excluded from our model, because of it coefficients estimation closed from zero exactly. In LPC. Q. Reg$_{\tau=0.55}$,we see the independent variables (x_1, x_2, x_5, x_7) are excluded from our model, because of it coefficients estimation closed from zero exactly.

5. Conclusions and Recommendations

5.1 Conclusions

In results showed from simulation examples and real data sets ,we see the our proposed methods (LPC. Q. Reg) is better than other method for variable selection and parameters estimation via all quantile level and all different errors. All methods (Q.P.C Reg,BQ.P.C Reg andL.P.C.Q.Reg) have a good performance with high correlation between independent variables , but our proposed method L.P.C.Q.Reg was very active with all correlation level between independent variables.

5.2 Recommendations

We recommended extension this study with another shrinkage method such as group lasso function , adaptive Lasso function and fused lasso function. And ,We recommended extension this study with binary quantile regression and with topit quantile regression model. And applied these proposed method on real data about specific phenomenon.

References

[1] Alhamzawi, R., “Tobit quantile regression with adaptive lasso penalty”, The 4th International Scientific Conference of Arab Statistics. 2013, 450 ISSN, pp. 1681-6870).

[2] Cade, B. S. and Noon, B. R. “A gentle introduction to quantile regression . 2003.

[3] Davino, C., Romano, R., & Vistocco, D. Handling multicollinearity in quantile regression through the use of principal component regression. METRON. 2022, 80(2), 153-174.

[4] Eberly, L. E, Multiple linear regression. Topics in Biostatistics. 2007. 165-187.

[5] Ergon, R., Granato, D., & Ares, G. Principal component regression (PCR) and partial least squares regression (PLSR). Mathematical and statistical methods in food science and technology Wiley Blackwell, Chichester. 2014. 121-42.

[6] Kim, B., Kanai, M. I., Oh, Y., Kyung, M., Kim, E. K., Jang, I. H., ... & Lee, W. J. Response of the microbiome–gut–brain axis in Drosophila to amino acid deficit. Nature. 2021. 593(7860), 570-574.

[7] Koenker, R. Quantile Regression. Cambridge Books. Cambridge University Press. 2005.

[8] Koenker, R. and G. J. Bassett. Regression quantiles. Econometrica. 1978. 46, 33–50.

[9] Koenker, R. and V. D’Orey. Algorithm AS 229: Computing regression quantiles. 1987.

[10] Koenker, R., & Geling, O. Reappraising medfly longevity: a quantile regression survival analysis. Journal of the American Statistical Association. 2001. 96(454), 458-468.

[11] Kostov, P., & Davidova, S. A quantile regression analysis of the effect of farmers’ attitudes and perceptions on market participation. Journal of Agricultural Economics. 2013. 64(1), 112-132.

[12] Kyung, M. Bayesian analysis of quantile principal component regression model. The Korean Data & Information Science Society. 2021. 32(4), 739-755.

[13] Sutter, J. M., Kalivas, J. H., & Lang, P. M. Which principal components to utilize for principal component regression. Journal of chemometrics. 1992. 6(4), 217-225.

[14] Tibshirani, R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological). 1996. 267-288.

[15] Wang, H., & He, X. Detecting differential expressions in GeneChip microarray studies: a quantile approach. Journal of the American Statistical Association, 2007. 102(477), 104-112.