Journal of Economics and Administrative Sciences (JEAS)

Available online at http://jeasiq.uobaghdad.edu.iq

Semi parametric Estimators for Quantile Model via LASSO and SCAD with Missing Data

Aws Adnan Al-Tai
University of Baghdad, College of Administration and Economics, Baghdad, Iraq
Aws.adnan1201a@coadec.uobaghdad.edu.iq

Qutaiba N. Nayef Al-Kazaz
University of Baghdad, College of Administration and Economics, Baghdad, Iraq
dr.qutaiba@coadec.uobaghdad.edu.iq

Received: 13/9/2022 **Accepted: 28/9/2022** **Published: September / 2022**

Abstract

In this study, we made a comparison between LASSO & SCAD methods, which are two special methods for dealing with models in partial quantile regression. (Nadaraya & Watson Kernel) was used to estimate the non-parametric part ;in addition, the rule of thumb method was used to estimate the smoothing bandwidth (h). Penalty methods proved to be efficient in estimating the regression coefficients, but the SCAD method according to the mean squared error criterion (MSE) was the best after estimating the missing data using the mean imputation method.

Paper type: Research paper.

Keywords: Quantile regression, partial linear model, LASSO, SCAD, missing data, nearest neighbor.

1. Introduction

The semi-parametric model is one of the regression models that combines the characteristics of parametric regression and non-parametric regression in order to obtain the best curve for the data. Sherwood (2016) quantile with partial linear regression to estimate missing data, Alhamzawi et al (2019) suggested Bayesian Adaptive Lasso Quantile Regression, Tibshirani (1996) proposed the Lasso method, which is considered one of the most famous penal methods used in estimating and selecting a linear regression model simultaneously. The main objective of regression analysis is to reduce the observed data or summarize it to ensure its presentation, for the relationship between each of the explanatory and response variables, and to analyse the regression line and give a conceptual and an approximation to that relationship by drawing or displaying that relationship according to the direction of the approximation line. As the semi-parametric regression model is a model that combines parametric regression and nonparametric regression, one of the most famous semi-parametric models is the partial linear regression model and is symbolized by the symbol (PLM). In addition, it gives an easier explanation for the effect of each variable compared to a complete non-parametric regression, as well as better than the non-parametric model because it avoids the curse of dimensional problem that occurs when the number of explanatory variables is increased in the non-parametric model. Quantile Regression is one of the important regression methods that have the ability to investigate the relationship between the response variable and the explanatory variables and in the entire conditional distribution of the response variable by estimating the conditional percentiles $\left(\mathrm{Q}_{\tau}(\mathrm{Y}|\mathrm{X})\right), 0 < \tau < 1$ .It differs in the distribution of the response variable rather than being limited to estimating the conditional expectation (E(Y|X=x)) as in the normal mean regression. It is the observations of the explanatory variables.

$Q_{\tau}(Y|X) = \mathrm{X_i}\beta_{\tau}$, $0 < \tau < 1$.

2. Partial Quantile Linear Regression Model (PQLRM)

It is considered as one of the important regression models that have the ability to investigate the relationship between the response variable and the explanatory variables and in the full conditional distribution of the response variable by estimating the conditional function $\left(Q_{\tau}(Y|X)\right)$, $0 < \tau < 1$ the difference in the distribution of the response variable rather than being limited for estimating the conditional expectation $\left(\mathrm{E}(Y|X = x)\right)$ as in the normal mean regression. And since the partial quantile linear regression model is written according to the following formula:

$$Y_i = X_i^T \beta_{\tau,i} + g(T_i) + \varepsilon_i \ , \quad \mathrm{i} = 1, \dots, \mathrm{n} \qquad (1)$$

Where X and T are explanatory variables.

Y: The vector of the response variable of degree (n× 1)

g(.): Unknown smooth function of degree (n×1)

β: Unknown parameters vector of degree (p×1)

ε: is the error term

there is a difficulty in estimation process because there is no clear behavior of the variables in the non-parametric part of the model leads to use one of the smoothing methods to solution the noise in the non-parametric part in order to show its true behavior.

In this paper, (NW) will be used to build the model correctly, and in order to implement the estimation process, we will use penalty methods, such as the LASSO & SCAD method, which has the characteristics that qualify to perform the estimation process accurately, where the penalties perform the necessary approximation of variables of the model to extend the estimation process, taking into account all the variables of the model and with as little bias as possible (Zhao, 2015).

3. LASSO Method

Tibshirani 1996 Proposed the technique of the work of the (LASSO) method is based on reducing the sum of the squares of the residuals according to a constraint that represents the absolute sum of the coefficients, which is less than a certain constant. (LASSO) does the Shrinkage process, as it makes a penalty for the regression coefficients and makes some equal to zero. Also, Lasso's estimator for the parameter (β) according to the quantile regression model is written as follows: (Tibshirani, 1996)

$$\widehat{\beta}_{lasso} = \operatorname{argmin}_{\beta}\left\{\sum_{i=1}^{n}\rho_{\tau}\left(yi - X_i^T\beta_{\tau}\right)^2 + \lambda\sum_{j=1}^{p}\left|\beta_{\tau,j}\right|\right\} \qquad \dots(1)$$

Where: (λ) is called the Penalty Parameter or the Regularization Parameter, $\lambda\sum_{j=1}^{p}|\beta|$ called the Penalty Function.

4. SCAD Method

Fan and Li (2001) proposed the oracle properties of SCAD besides to variables selection (Smoothly Clipped Absolute Deviation), and (Fan and Li, 2001) predicted that the LASSO penalty does not have Oracle properties؛The SCAD estimator for the parameter (β) can be written according to the quantile regression model as:

$$\widehat{\beta}_{scad} = \min_{\beta}\left\{\sum_{i=1}^{n}\rho_{\tau}\left(Y_i - X_i^T\beta_{\tau}\right)^2 + \sum_{j=1}^{p}P_{\lambda}\left(\left|\beta_{\tau,j}\right|\right)\right\} \qquad \dots(2)$$

Where: $\lambda \geq 0$, $\sum_{j=1}^{p}P_{\lambda}\left(\left|\beta_{\tau,j}\right|\right)$ called the Penalty Function

5. Penalty Parameter

The basic step in Penalized Regression is the selection of the penalty parameter or the so-called regularization parameter, which is symbolized by the symbol (λ), and it is the parameter that controls the amount of reduction of the parameters and the selection of the subset of the variables included in the final model. (Craven, and Wahba, 1978)

In our article, the generalized Cross Validation method was used to estimate the value of the penalty parameter, which is written according to the following formula:

$$V_{(\lambda)} = \frac{1}{n}\sum_{k=1}^{n}\left(\sum_{j=1}^{n}a_{kj}y_j - y_k\right)^2 \Big/ \left(1 - \frac{1}{n}\sum_{k=1}^{n}a_{kk}\right)^2 \qquad \dots(3)$$

Where: $\frac{\partial g_{n,\lambda}^{[k]}(t)}{\partial y_k} = a_{kk}$, $g_{n,\lambda}^{[k]}(t) = \sum_{j=1}^{n} a_{kj} y_j$

6. Nadaraya-Watson Estimator

This estimator is characterized by being used in non-parametric regression functions, but in this research, it will be employed to work as an estimator for the semi-parametric regression model.

As for the kernel function used with the (N.W) estimator, it has several properties, including:

1- $\int k(v)dv = 1$

2- $\int vk(v)dv = 0$

3- $\int v^z k(v)dv = 0$, $\forall\ z = 1, 3, \dots, k-1$

And where (k) represents the degree of the kernel function, it has been confirmed in most applications that these conditions are fulfilled when (z = 2), that is the kernel functions are of the second order, which are recognized either through derivation or integration. (Hmood and Mohamed, 2014)

(N.W) estimator in semi-parametric regression functions can be summarized according to the following formula:

$$\hat{g}_{N.W}(t) = \frac{\sum_{i=1}^{n} K_{ht}(t_i - t) y_i}{\sum_{i=1}^{n} K_{ht}(t_i - t)} \quad \dots (4)$$

The estimator of the kernel $\left(\hat{g}_{N.W}(t)\right)$ **can be written using a function of weights, which is equal to:**

$$W_{ht}(t, T_i) = \frac{K_{ht}(t_i - t)}{\sum_{i=1}^{n} K_{ht}(t_i - t)} \quad \dots (5)$$

Where it can be written in the following form:

$$\hat{g}_{N.W}(t) = \sum_{i=1}^{n} W_{ht}(t, T_i) y_i \quad \dots (6)$$

Whereas $\sum_{i=1}^{n} W_{ht}(T_i, t) = 1$

Thus, the estimator of (N.W) in the semi-parametric regression functions is according to the following formula:

$$g_n(T, \beta) = \sum_{i=1}^{n} W_{ni}(t)(Y_i - X_i'\beta) \quad \dots (7)$$

Where: $(\{W_{ni}(t)\}_{i=1}^{n})$ **denote to a series of weights and these weight functions can be normal if the following condition is met:** $\sum_{i=1}^{n} W_{ni}(t) = 1$

Where: $W_{ni} > 0$

As for the shape of the weights, it is determined by the weight function, which also represents the kernel function: $K\left(\frac{T_i - t}{h}\right)$**. And this function gets to its maximum when (Ti) approaches (t) and decreases when** (Ti) **gets away from (t). And the weight function can be defined as follows:**

$$W_{ni}(t) = \frac{K\left(\frac{t_i - t}{h_n}\right)}{\sum_{j=1}^{n} K\left(\frac{t_i - t}{h_n}\right)} \quad \dots (8)$$

7. Choosing the Smoothing Parameter

It is also called the bandwidth parameter and is denoted by (h). The choice of the kernel function and the smoothing parameter are necessary in estimating the regression function; in this research, the rule of thumb method was used to estimate the value of Smooth.

7.1 Rule Of Thumb (ROT)

This method is one of the important methods for selecting bandwidth parameter; the bandwidth parameter is obtained by reducing Weighted Mean Integrated squared error of the regression function estimator. (Hmood, and Stadtmüller, 2013)

$$WMISE = \int_{-\infty}^{+\infty} \left[Bias^2\left(\hat{m}(t)\right) + var\left(\hat{m}(t)\right)\right] w(t)dt$$

$$Bias\left(\hat{m}(t)\right) = \frac{h}{2} m''(t) \int_{-\infty}^{+\infty} u^2 K(u)du + O_p(h)$$

$$var\left(\hat{m}(t)\right) = \frac{\sigma^2(t)}{nhf(t)} \int_{-\infty}^{+\infty} k^2(u)du + O_p(nh)^{-1}$$

Whereas $W \geq 0$ represents a weight function and this decreases results in:-

$$h_{opt} = C(k) \left[\frac{\int \sigma^2(t)w(t) \Big/ f(t)dt}{\int \left(m''(t)\right)^2 w(t)dt} \right]^{1/5} .n^{-1/5} \quad \ldots (9)$$

Whereas:

C(K): Represents a specific constant value that depends on the type of kernel function used.

$m''(t)$: is the second derivative of the regression function & f(t): represents the probability density function.

The rule of thumb is based on a quadratic polynomial model matching.

$$m(t) = \alpha_0 + \alpha_1 t + \cdots + \alpha_4 t^4 \quad \ldots\ldots (10)$$

Whereas: $\hat{\sigma}^2$ represents the sum of the squares of the residuals, which is estimated using the polynomial model in the equation (10). Then the estimated values are substituted into the equation (9). We get:

$$\hat{h}_{opt} = C(k) \left[\frac{\hat{\sigma}^2(t) \int w(t)dt}{\sum_{j=1}^{n} \left(\hat{m}''(t_j)\right)^2 w(t)} \right]^{1/5} .n^{-1/5} \quad \ldots (11)$$

8. Missing Data

Sometimes, especially in the applied field, the information to be studied may not be available due to a number of reasons, including the loss of that information for insufficient (unknown) reasons and the loss may be done on purpose, data loss may bias this data, and this affects the quality of data Liang, et al (2004).

Missing Data Imputation: this method replaces the missing values in a data set with other possible values, and it has several benefits, including that the treatment of these missing data does not always depend on a specific method, and this allows researchers to choose a calculation method that is more appropriate with their applications.

Since each process of loss has a specific pattern (mechanism) according to which the loss takes place according to a certain probability, and as a result, the (Little and Rubin) classified the mechanisms of missing data into three types:

1.Missing at Random: this type of loss is only related to the values of other variables, while it is independent of its missing data. In such a case, the loss is random. (MAR).

2.Missing Completely at Random: this type of loss is due to the fact that the loss is independent of the missing data itself as well as independent of the values of other variables in the sample, then it can be said that the data is completely and randomly lost (MCAR).

3.Not missing at random: the reason for this loss is generated as a result of the lost data itself; that is, the loss of data will be intentional and not random (Not MAR).

8.1. Partial Quantile Regression Model in the case of complete and incomplete data

We consider the case where some (Y) values in a sample size n may be missing, but X and T are observed completely.

That is, we obtain the following incompletely observations:

$$\delta_i = \begin{cases} 0 & \text{if } Y_i \text{ is missing} \\ 1 & \text{otherwise} \end{cases}$$

In practice, the MAR assumption is usually justified in the nature of experiments, especially when it is consider hat missing Y mainly depends on X. MCAR has a stronger assumption than MAR, where, MCAR is a special case of MAR.

And assuming that Y is missing at random Qin, et al (2007).

Let:

$$r = \sum_{i=1}^{n} \delta i \quad , \quad m = n - r$$

Where: (m, r) is defined as the response group and the non-response group or the loss response variable Y (respectively).

In addition to that, (S_m) represents the non-response state, (S_r) represents the response state.

Assuming that (K) is a symmetric probability density function and ((h = h_n) the bandwidth that is decreasing towards zero with increasing sample size (n→ ∞).

Where:

$$Y_i - X_i^T \beta = g(T_i) + \varepsilon_i \qquad i = 1, 2, \dots, r$$

Assuming that (β) values are defined, we have a kernel estimator $\hat{g}(t)$ for $g(t)$, based on the complete observations data:

$$\hat{g}(T_i) = \frac{\sum_{j=1}^{n} \delta_j K\left(\frac{T_i - t_j}{h}\right)(Y_j - X_j \beta)}{\sum_{j=1}^{n} \delta_j K\left(\frac{T_i - t_j}{h}\right) + n^{-2}} \qquad \dots (12)$$

Where K(.) is called the kernel function, which can be obtained from using the (Gaussian kernel) the standard normal density function and using $\hat{g}(t)$ instead of g(t) in equation (12) we get:

$$Y_i - X_i^T\beta \approx \frac{\sum_{j=1}^{n} \delta_j K\left(\frac{T_i - t_j}{h}\right)(Y_j - X_j\beta)}{\sum_{j=1}^{n} \delta_j K\left(\frac{T_i - t_j}{h}\right) + n^{-2}} \quad , i \in r \qquad (13)$$

Since the n^{-2} component is added to avoid the case that the denominator is zero. Using the transformations, we get:

$Z_i \approx U_i^T\beta \ , \quad i \in S_r$

where:

$$Z_i = Y_i - \frac{\sum_{j=1}^{n} \delta_j Y_j K\left(\frac{(T_i - t_j)}{h}\right)}{\sum_{j=1}^{n} \delta_j K\left(\frac{(T_i - t_j)}{h}\right) + n^{-2}}$$

$$U_i = X_i - \frac{\sum_{j=1}^{n} \delta_j X_j K\left(\frac{(T_i - t_j)}{h}\right)}{\sum_{j=1}^{n} \delta_j K\left(\frac{(T_i - t_j)}{h}\right) + n^{-2}} \quad , i \in S_r$$

The parameters of β can be estimated as according to the theory of the penalty linear quantile regression model as follows:

- In regard to adding the penalty limit for the SCAD method estimator:

$$\hat{\beta}_{scad,n} = \left(\sum_{i=1}^{n} \delta_i U_i U_i^T\right)^{-1} \left(\sum_{i=1}^{n} \delta_i U_i Z_i\right) + \sum_{j=1}^{p} P_\lambda\left(|\beta_{\tau,j}|\right) \qquad \dots (14)$$

- In regard to adding the penalty limit for the estimator of the LASSO method:

$$\hat{\beta}_{lasso,n} = \left(\sum_{i=1}^{n} \delta_i U_i U_i^T\right)^{-1} \left(\sum_{i=1}^{n} \delta_i U_i Z_i\right) + \lambda \sum_{j=1}^{p} |\beta_{\tau,j}| \qquad \dots (15)$$

And by substituting into the equation (12) we get:

$$\hat{g}_n(t_i) = \frac{\sum_{j=1}^{n} \delta_j K\left(\frac{T_i - t_j}{h}\right)(Y_j - X_j\hat{\beta}_n)}{\sum_{j=1}^{n} \delta_j K\left(\frac{T_i - t_j}{h}\right) + n^{-2}} \qquad \dots (16)$$

9. Mean Imputation in the Response Variable

This method is used to estimate the missing value in a series of data and compensate for it with the mean value of the studied variable, as this method helps to maintain the actual size of the data and reduce the discrepancy between observations and is characterized by its ease of use, and to reduce the value of deviations has a significant impact on its estimation, as we will obtain a biased value for the deviations, which also affects the value of covariance and correlations, where the mean value is compensated instead of the missing values, and if the missing values are few, the bias value will be relatively low (Jamshidian, and Mata, 2007).

10. Discussion of Results

10.1 Simulation

Simulations were performed using (1000) replicate, three sample sizes for each experience (n= 30, 50, 100), and as follows:

1- The explanatory variables of the parameter part (Xi) are generated in the following form (Hmood and Mohamed, 2014) :

$$X_1 = 2 \times \bar{X}_1 \times U$$

$$X_2 = 2 \times \bar{X}_2 \times U$$

whereas: $0 < U < 1$

Where the following values (5.2, 6.5) were used as initial values for the mean in the generation process

2- The explanatory non-parametric variables (ti) **are distributed normally with mean (0) and variance (1), where four values of variance of error (2, 5, 7.2, and 10.3) were used.**

3- The dependent variable is generated through the models used in simulation experiments through the use of regression functions for the explanatory variables of the parametric part and the non-parametric part with an error term added.

The following model was used:

$$g(t) = 3.2t^2 - 1$$

The kernel function used is a standard normal density function, Gaussian Kernel, as follows:

$$K(.) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right)$$

Where data missing data is random (MAR) according to the percentage of loss (10%, 20%, 30%), that is, if the cause of loss is related to the values of other variables only and is independent of the missing value.

Table No. (1) Simulation results when P=2 , G=2 $\sigma = 2$

MSE $\widehat{Y}$									
	NO MISS			MISS 10%		MISS 20%		MISS 30%	
tao	n	LASSO	SCAD	LASSO MI	SCAD MI	LASSO MI	SCAD MI	LASSO MI	SCAD MI
0.3	30	0.5417	0.5483	2.5978	2.5840	2.0687	2.0415	2.5441	2.4862
	50	3.2032	3.1879	2.2017	2.1923	3.1489	3.1409	4.2043	4.2061
	100	4.4994	4.4943	3.4726	3.4766	2.3091	2.3002	1.0777	1.0746

Table No. (2) Simulation results when P=2, G=2 $\sigma = 5$

MSE $\widehat{Y}$									
	NO MISS			MISS 10%		MISS 20%		MISS 30%	
tao	n	LASSO	SCAD	LASSO MI	SCAD MI	LASSO MI	SCAD MI	LASSO MI	SCAD MI
0.6	30	0.3015	0.3155	0.2842	0.4141	0.3234	0.3941	0.2869	0.3113
	50	3.3706	3.3391	3.4156	3.4301	3.5539	3.4054	3.3372	3.3441
	100	2.9428	3.0205	2.8648	2.6532	3.5513	3.1664	2.8156	3.1817

Table No. (3) Simulation results when P=2, G=2, $\sigma = 7.2$

MSE $\widehat{Y}$									
	NO MISS			MISS 10%		MISS 20%		MISS 30%	
tao	n	LASSO	SCAD	LASSO MI	SCAD MI	LASSO MI	SCAD MI	LASSO MI	SCAD MI
0.3	30	1.2820	1.2658	1.1117	1.2843	1.1889	1.1280	2.0232	1.2658
	50	0.5706	0.5191	0.5281	0.5685	0.5778	0.5285	1.8399	0.5191
	100	0.3879	0.3484	0.3039	0.6691	0.3744	0.3438	2.7406	0.3484

Table No. (4) Simulation results when P=2, G=2, $\sigma = 10.3$

MSE $\widehat{Y}$									
	NO MISS			MISS 10%		MISS 20%		MISS 30%	
tao	n	LASSO	SCAD	LASSO MI	SCAD MI	LASSO MI	SCAD MI	LASSO MI	SCAD MI
0.6	30	1.9191	1.9437	2.0400	1.9831	2.0473	2.0853	1.9112	2.0338
	50	1.7242	1.6329	1.7211	1.6761	1.5390	1.7244	1.7189	1.6710
	100	3.5106	2.5296	2.6834	3.8340	2.5978	2.6341	3.6931	2.5721

*** No. of linear variates**
*** No. of non-parametric variates**
*** $\sigma = variance$**

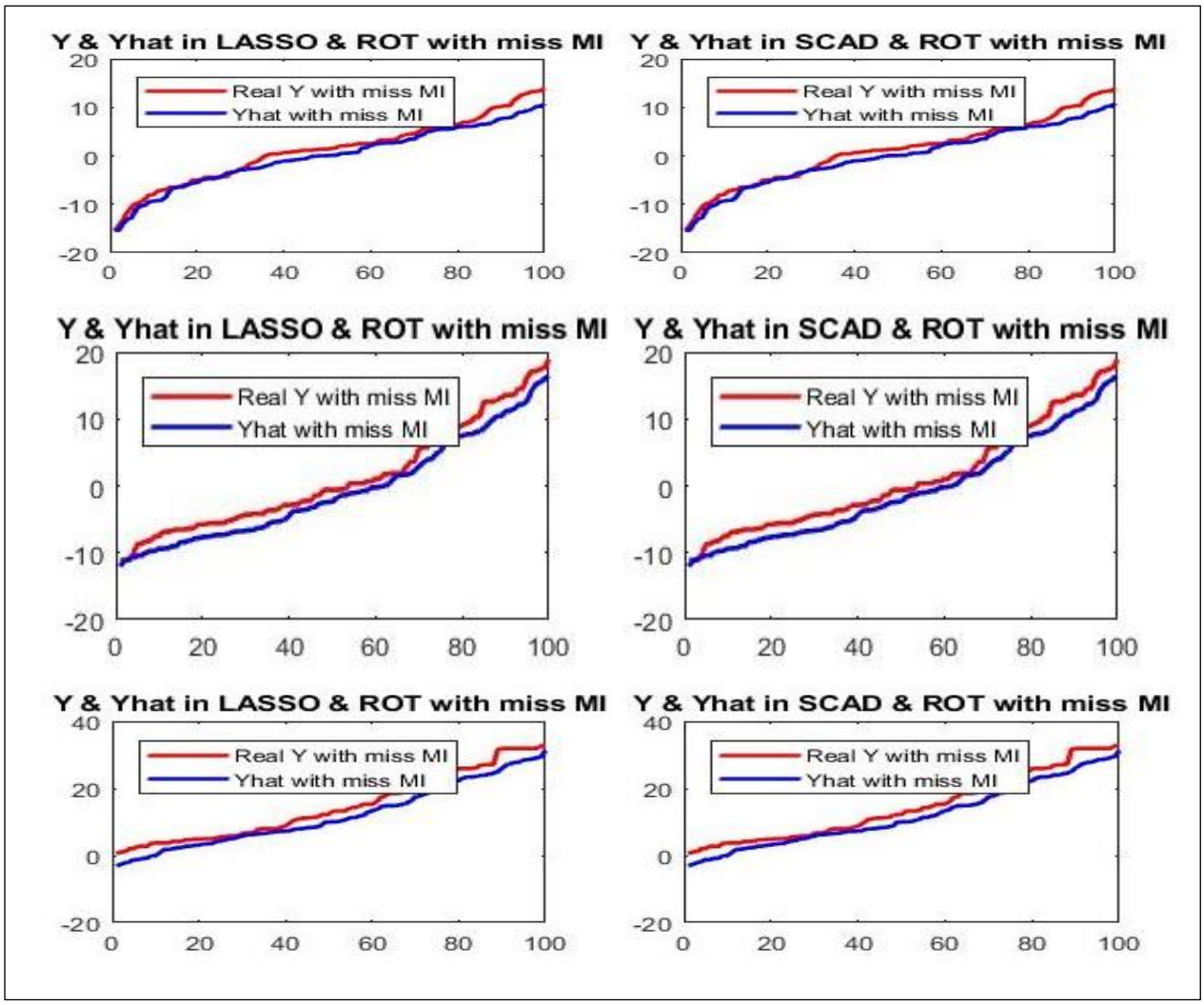

Figure1: shapes when n=100 $\tau = 0.3$ P=2 , G=2, $\sigma = 2$, miss 10%,20%&30%

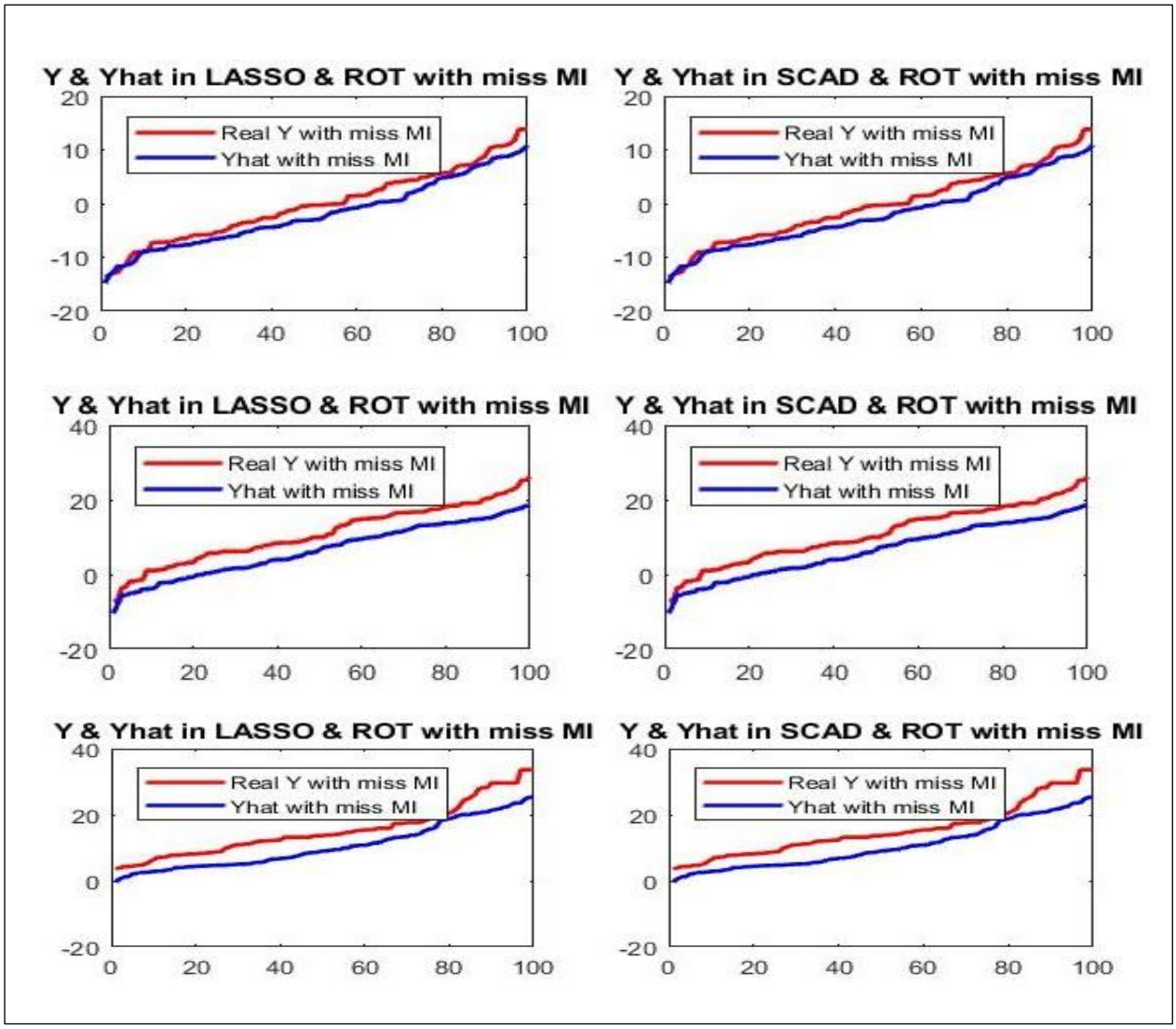

Figure2:Shapes when n=100 $\tau = 0.6$ P=2 , G=2, $\sigma = 5$ miss 10%,20%&30%

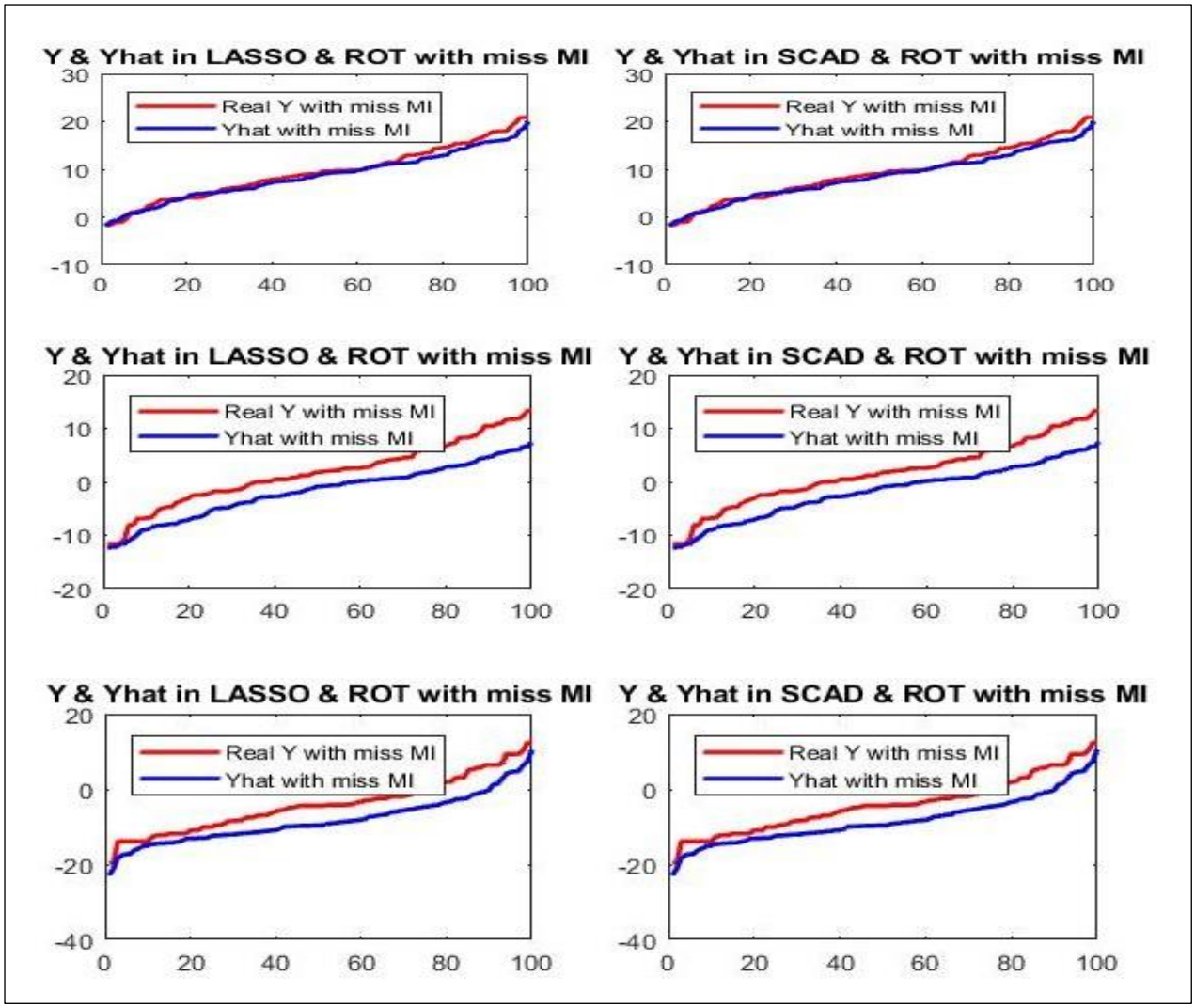

Figure3:Shapes when n=100 $\tau = 0.3$ P=2 , G=2, $\sigma = 7.2$, miss 10%,20%&30%

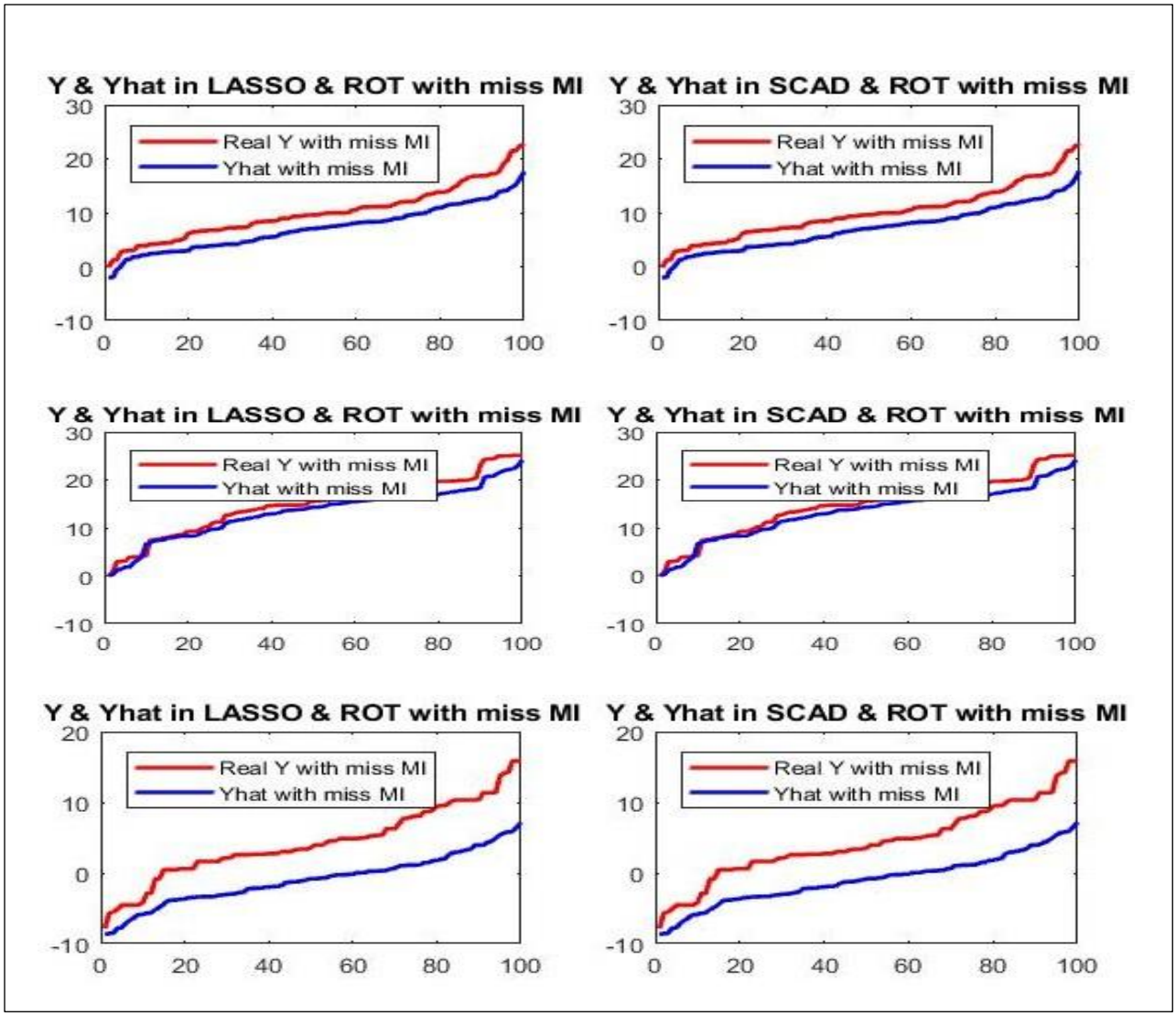

Figure4:Shape when n=100 $\tau = 0.6$ P=2 , G=2, $\sigma = 10.3$, miss 10%,20%&30%

11. Conclusions

The estimation process using LASSO and SCAD method is relatively simple, fast and very suitable for semi-polar models; on the other hand, adding the missing to the data had a significant effect on the amount of mean square error (MSE) in the estimated model, where it was noticed that the higher the variance, the lower the mean square error (MSE). We conclude from this comparison that there is a clear convergence between the estimation process by the two methods, without loss, with the SCAD method preferred when data loss occurs. A significant effect of the missing appeared on the LASSO method with an increase in the amount of least squares, knowing that the method that was used to estimate the missing data was the mean compensation method (MI) and the rule of thumb method to estimate the smoothing parameter.

References

1. Alhamzawi, R., Yu, K. and Benoit, D.F., 2012. Bayesian adaptive Lasso quantile regression. Statistical Modelling, 12(3), pp.279-297.
2. Craven, P. and Wahba, G., 1978. Smoothing noisy data with spline functions. Numerische mathematik, 31(4), pp.377-403.
3. Fan, J. and Li, R., 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American statistical Association, 96(456), pp.1348-1360
4. Hmood, M.Y. and Mohamed, M., 2014. A comparison of the Semiparametric estimators model using different smoothing methods. Journal of Economic and Administrative Sciences, 20(75), pp.376-394.
5. Hmood, M.Y. and Stadtmüller, U., 2013. A New Version of Local Linear Estimators. Chilean journal of statistics, 4(2), pp.61-74.
6. Jamshidian, M. and Mata, M., 2007. Advances in analysis of mean and covariance structure when data are incomplete. In Handbook of latent variable and related models (pp. 21-44). North-Holland.
7. Koenker, R. and Bassett Jr, G., 1978. Regression quantiles. Econometrica: journal of the Econometric Society, pp.33-50.
8. Liang, H., Wang, S., Robins, J.M. and Carroll, R.J., 2004. Estimation in partially linear models with missing covariates. Journal of the American Statistical Association, 99(466), pp.357-367.
9. Qin, Y., Zhang, S., Zhu, X., Zhang, J. and Zhang, C., 2007. Semi-parametric optimization for missing data imputation. Applied Intelligence, 27(1), pp.79-88.
10. Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58(1), pp.267-288.
11. Wellner, J.A., Klaassen, C.A. and Ritov, Y.A., 2006. Semiparametric models: a review of progress since BKRW (1993).
12. Zhao, P.X., 2015. Quantile Regression for Partially Linear Models with Missing Responses at Random. In Applied Mechanics and Materials (Vol. 727, pp. 1013-1016). Trans Tech Publications Ltd.
13. Sherwood, B., 2016. Variable selection for additive partial linear quantile regression with missing covariates. Journal of Multivariate Analysis, 152, pp.206-223.

المقدرات شبه المعلمية للأنموذج التجزيئي باستعمال LASSO و SCAD مع البيانات المفقودة

الباحث/ أوس عدنان الطائي
جامعة بغداد، كلية الإدارة والاقتصاد، بغداد، العراق
Aws.adnan1201a@coadec.uobaghdad.edu.iq

أ.د. قتيبة نبيل القزاز
جامعة بغداد، كلية الإدارة والاقتصاد، بغداد، العراق
dr.qutaiba@coadec.uobaghdad.edu.iq

Received: 13/9/2022 **Accepted: 28/9/2022** **Published: September / 2022**

مستخلص البحث

في هذه الدراسة، قمنا بإجراء مقارنة بين طريقتين LASSO وSCAD، وهما طريقتان خاصتان للتعامل مع النماذج في الانحدار التجزيئي الجزئي. تم استخدام (Nadarya & Watson Kernel) لتقدير الجزء اللامعلمي (non-parametric)، بالإضافة إلى ذلك، تم استخدام طريقة قاعدة الإبهام لتقدير المعلمة التمهيدية (h). أثبتت طرق الجزاء فعاليتها في تقدير معاملات الانحدار، لكن طريقة SCAD وفقًا لمعيار متوسط مربعات الخطأ (MSE)، كانت الأفضل بعد تقدير البيانات المفقودة باستخدام طريقة التعويض بالمتوسط (mean imputation).

نوع البحث: ورقة بحثية.

المصطلحات الرئيسة للبحث: الانحدار التجزيئي، انموذج الانحدار الجزئي، لاسو، سكاد، البيانات المفقودة، المجاور الأقرب.

*البحث مستل من رسالة ماجستير