AL KUT JOURNAL OF ECONOMIC AND ADMINISTRATIVE SCIENCES
Publisher: College of Economics and Management - Wasit University

Comparison between some methods for Estimation and Variables Selection for Semi – Parametric Additive model with practical application

Hayder Raaid Talib[1] — University of Sumer, Dhi Qar Al-Rifai, Iraq — h.raid @ uos.edu.iq

Dr. Munaf Yousif Hmood[2] — University of Baghdad, Baghdad, Iraq — munaf.yousif@coadec.uobaghdad.edu.iq

Abstract:

The additive partial linear model (APLM) was used to estimate the effects of some linear and non-linear explanatory variables on the response variable represented in the market value of the Baghdad Soft Drinks Company. Four methods have been to choose and estimate the model, namely ALasso, SCAD, Elastic Net and Adaptive Elastic Net. The methods used were compared using comparison criteria represented by mean squares error and coefficient of determination. Through the results of the analysis, it has been noticed that the method Adaptive Elastic Net is more efficient than to other methods.

Key words: APLM, ALasso, SCAD, Elastic Net, Adaptive Elastic Net.

1- Introduction

In some applications of multiple regression, the number of explanatory variables is large, therefore, the analysis of these models becomes difficult. Therefore, for dealing with this problem, to reduce its dimensions by making some assumptions. and extending the single smoothing mode to a high-dimensional .

The basic idea of scatter plot smoothing can be extended to higher dimensions in a straightforward way. Theoretically, the regression smoothing for a d-dimensional predictor can be performed as in the case of a one-dimensional predictor. The local averaging procedure will still give asymptotically consistent approximations to the regression surface. However, there are two major problems with this approach to multiple regression smoothing. First, the regression function $m(x)$ is a high dimensional surface and since its form cannot be displayed for $(d > 2)$, it does not provide a geometrical description of the regression relationship between X and Y. Second, the basic element of nonparametric smoothing, averaging over neighborhoods, will often be applied to a relatively little set of points since even samples of size $(n \geq 1000)$ are surprisingly sparsely distributed in the higher dimensional Euclidean space. [5]

If we had $(n = 1000)$ points uniformly distributed over the ten-dimensional unit cube $[0, 1]^{10}$. What is our chance of catching some points in a neighborhood of reasonable size? An average over a neighborhood of diameter (0.3)(in each coordinate) result in a volume of $[0.3^{10} \approx 5.9 * 10^{6}]$ for the corresponding ten-dimensional cube. Hence, the expected number of observations in this cube will be $[5.9 * 10^{3}]$ and not much averaging can be expected. If, on the other hand, one fixes the count $(k = 10)$ of observations over which to average, the diameter of the typical (marginal) neighborhood will be larger than (0.63) which means that the average is extended over at least two thirds of the range along each coordinate. [5]

A first view of the sparsity of high dimensional data could lead one to the conclusion that one is simply in a hopeless situation - there is just not enough clay to make the bricks! This first view, however, is, as many first views are, a little bit too rough. Assume, for example, that the ten-dimensional regression surface is only a function of $[X_1]$, the first coordinate of X, and constant in all other coordinates. In this case, the ten-dimensional surface collapses down to a one-dimensional problem. A similar conclusion holds if the regression surface is a function only of certain linear combinations of the coordinates of the predictor variable.

The basic idea of additive models is to take advantage of the fact that a regression surface may be of a simple, additive structure.

When we talk about reduction dimensionality, it means that there are higher dimensions. These dimensions are known as variables. The more variables there are, the more difficult to deal with them. So, we're going to have a problem known as the curse of dimensions which occurs as a result of increase the number of explanatory variables dealt with the greater the number of dimensions, the more difficult it is to predict the dependent variable.

Not all of these variables may be effective (significant), in addition to what has been mentioned, there is a reduction in the procedures of the computer solution.

2- Additive Partial Linear Model (APLM)

The additive partial linear model (APLM) is one of the semi-parametric models, that received special attention recently because it combines both the parametric and non-parametric components. that is, the response variable depends on some explanatory variables in a linear and another non-linear manner. Hence, this model can put clearer explanations and is preferred over nonparametric, parametric models or nonparametric additive models, so it can be considered as a special case of generalized additive nonparametric regression models. [8][13]

This model is more flexible compared with linear model and more efficient compared with nonparametric regression model because it reduces a problem known as the curse of dimensionality or multidimensionality. [2]

The APLM model can be written as:

$$Y = X^T\beta + \sum_{k=1}^{K} g_k(Z_k) + \varepsilon \qquad . \quad . \quad . \qquad (1)$$

where $\{(X_i, Z_i, Y_i)\}_{i=1}^{n}$ denote to an independent and identical distributed random samples of size n, where

$X = (X_1, X_{2,\dots}, X_d)^T$ denote to the linear variables.

$Z = (Z_1, Z_{2,\dots}, Z_K)^T$ denote to the non-linear variables.

$(g_1, g_{2,\dots}, g_k)$ refers to unknown smoothing functions.

$B = (B1, B2, \dots, Bd)^T$ indicates a vector of unknown parameters.

ε is a random error of the model and is independent of each(X, Z).

with Zero mean and variance equal to σ^2 and $E\big((\varepsilon|X, Z)\big) = 0$

To ensure that the nonparametric functions are defined, we assume that:

$E[g_k(Z_k)] = 0 \qquad \text{for} \quad k = 1,2,\dots,K$

APLM has the simplicity of a general linear model (GLM) and the flexibility of a general additive model (GAM). It combines both parametric and nonparametric components. [6]

In many cases, the nonparametric components are often the troublesome parameters and the main interest is the linear part, therefore, our aim is to select the variables in the parametric part. [7]

The natural extension of the APLM model is the Generalized Additive Partial Linear Model (GAPLM). [6][10]

$$E(Y_i \mid X_i, Z_i) = G\left\{X_i^T\beta + \sum_{k=1}^{K} g_k(Z_{k,i})\right\} \quad . \ . \ . \quad (2)$$

where $G(.)$ is a unknown link function.

We note that the model (1) is a special case of the model GAPLM in (2) , so when (K = 1) we get a model (1) which called Partial Linear Model: [6] [8][13]

$$Y = X^T\beta + g(Z) + \varepsilon \quad . \ . \ . \quad (3)$$

3- <u>Model Selection Methods</u>

The selection of variables is essential in the statistical analysis of high-dimensional data. In most studies, researchers are usually interested in finding the relationship between the dependent variable and a set of the studied explanatory variables. As well as in diagnosing the most important explanatory variables without others. By using an appropriate selection method and under appropriate conditions, consistent models can be built patterns that are easy to interpret and avoid over-prediction and estimation. There are in general two methods to select the model in the case of the additive partial linear model (APLM): [1]

3-1 <u>Classical Methods</u>

These methods are unstable and they take a long time from the computational point of view, especially when the number of explanatory variables is large in the model. Also, this selection process leads to an increase in unexplained random errors in the steps of selecting variables, and among these methods are The stepwise regression, The back-word elimination, The forward selection procedure, Akaike information criterion (AIC), Bayesian information criterion (BIC) and Schwarz's Bayesian criterion (SBC). [1][3][4]

3-2 Regularization Methods

Regularization methods are used for model selection in the case of linear models; therefore, the problem of selection regression models has received extensive studies from researchers. To overcome such problems, the penalized methods have gained great importance in recent times because of their speed in selecting the explanatory variables and the stability of the selection results compared with the classical selection methods as well as used to solve complex model problems. Generally, performance is closely related to more complex models because high complexity models have low bias and high variance. While models with low complexity tend to have a high bias with low variance. So, regularization methods are often used to control the complexity of the model by setting penalty functions with more complex models. That is, the regular methods help to reduce the overfitting problem. The aim behind adding the penalty constraint is to control the complexity of the model and to provide a criterion for choosing a variable by introducing some restrictions on the parameters. These restrictions force some parameters to have their values equal to zero, which lead to improve the model prediction accuracy and providing an easily interpretable model. These methods are: [2][6][7]

3-2-1 Adaptive Least Absolute Shrinkage and Selection Operator (A Lasso)

Adaptive Lasso was proposed by researcher Zou (2012) [13]. this method corrects the problems in the Lasso procedure considering that the lasso method is biased, especially for large coefficients, as well as the lack of oracle properties therefore, it was processed through the assumptions of different weights for different coefficients in the penalty function. Therefore, this method has been addressed and thus will reduce the amount of bias and the accuracy of selecting variables V.S, The resulting estimate from adaptive lasso is unbiased, consistent, and validates oracle properties.

In this method, the estimation and selection of explanatory variables are carried out in two steps:

the first step: the use of the Lasso method for a preliminary estimate, as well as working to reduce the dimensions of the model.

the second step: the use of the adaptive lasso method to select and estimate the components for the additive partial linear model (APLM)[10]

using the spline approximation of the nonparametric components, the penalized -least squares criterion with adaptive lasso functions can be written.

$$L_n(\beta_n, \gamma_n, \lambda_n) = \sum_{i=1}^{n}\left[Y_i - \sum_{k=1}^{p} X_{ik}\beta_{nk} - \sum_{j=1}^{J}\sum_{w=1}^{m_n} \gamma_{jw} B_w(T_{ij})\right]^2 + \lambda_n \sum_{k=1}^{p} w_{nk} \|\beta_{nk}\|_2 \quad . . . (4)$$

p is the number of linear explanatory variables in the model and (k=1,2,. .,p).

n is the sample size (i=1,2,...,n).

$B_w(T_{ij})$ is the B-Spline basis function.

γ vector of spline parameters.

λ represents the adjustment parameter or penalty parameter of the adaptive lasso and is non-negative.

w_{nk} weights estimated.

let (Z_{ij}) be the basis functions of the i^{th}observation of the j^{th}nonparametric function defined by (g_j)

$Z_{ij} = \left(B_1(T_{ij}), \ldots, B_{mn}(T_{ij})\right)'$ and $Z_j = (Z_{1j}, \ldots, Z_{nj})'$

Z_j represents the "design" matrix of degree$(n \times m_n)$ for j^{th} of the nonparametric functions g_j

$Z = (Z_1, \ldots, Z_j)$ represent to the total "design" matrix of degree $(n \times J_{m_n})$, therefore, equation (4) can be written as:

$$L_n(\beta_n, \gamma_n, \lambda_n) = \|Y - X\beta_n - Z\gamma_n\|_2^2 + \lambda_n \sum_{k=1}^{p} w_{nk} \|\beta_{nk}\|_2 \quad . . . (5)$$

$$\left(\hat{\beta}_n, \hat{\gamma}_n\right) = \arg \min_{(\beta_n, \gamma_n)} L_n(\beta_n, \gamma_n, \lambda_n) \quad . . . (6)$$

with constraint

$$\sum_{i=1}^{n} \sum_{w=1}^{m_n} \gamma_{jw} B_w\left(T_{ij}\right) = 0 \quad ; \quad 1 \leq j \leq J \quad . . . (7)$$

These centering restrictions are to ensure unique identification of the $g_j'^s$ which are analogs of the restriction $E\left[g_j\left(Z_j\right)\right] = 0 \,;\, 1 \leq j \leq J$

the values of (β_n, γ_n) that minimum (L_n) and satisfy the following condition:

$$Z'Z\,\gamma_n = Z'\,(Y - X\beta_n) \quad . . . (8)$$

assuming that the number of non-parametric compounds is not too large.

$$P_Z = Z\,(Z'Z)^{-1} Z' \quad . . . (9)$$

P_Z is the projection matrix for the column space of Z.

$(I - P_Z)$ are an idempotent matrix and the eigenvalues for $(I - P_Z)$ are 0 or 1 and the order $(I - P_Z)$ is (J_{mn}),therefore, equation (5) can be written as:

$$L_n(\beta_n, \lambda_n) = \|(I - P_Z)(Y - X\beta_n)\|_2 + \lambda_n \sum_{k=1}^{p} w_{nk} \|\beta_{nk}\|_2 \quad . . . (10)$$

the estimation and selection of variables in this way are carried out in two stages:

Step 1: Compute the estimator of the Lasso. let

$$L_{n1}(\beta_n; \lambda_{n1}) = \|(I - P_Z)(Y - X\beta_n)\|_2 + \lambda_{n1} \textstyle\sum_{k=1}^{p} \|\beta_{nk}\|_2 \quad . . . (11)$$

Whereas the objective function (11) is a special case of (10) when

$(w_{nk} = 1 \,;\, 1 \leq K \leq P)$ and thus, Lasso's estimations are:

$$\tilde{\beta}_n \equiv \tilde{\beta}_n(\lambda_{n1}) = \arg\min_{\beta_n} L_{n1}(\beta_n, \lambda_{n1}) \quad . . . (12)$$

and from equation (8) we get:

$$\tilde{\gamma}_n = (Z'Z)^{-1} Z' \left(Y - X\tilde{\beta}_n\right) \quad . . . (13)$$

Step 2: using the estimators of Lasso in the first step $(\tilde{\beta}_n)$ to

obtain the weights of the adaptive lasso function, since:

$$w_{nk} = \begin{cases} \|\tilde{\beta}_n\|_2^{-1} & \text{if } \|\tilde{\beta}_n\|_2 > 0 \\ \infty & \text{if } \|\tilde{\beta}_n\|_2 = 0 \end{cases} \quad . \quad . \quad . \qquad (14)$$

The adaptive lasso objective function is as follows:

$$L_{n2}(\beta_n;\lambda_{n2}) = \|(I - P_Z)(Y - X\beta_n)\|^2 + \lambda_{n2}\sum_{k=1}^{p} w_{nk}\|\beta_{nk}\|_2 \quad . \quad . \quad . \qquad (15)$$

thus, the adaptive lasso estimator is: [14]

$$\hat{\beta}_n \equiv \hat{\beta}_n(\lambda_{n2}) = \arg\min_{\beta_n} L_{n2}(\beta_n, \lambda_{n2}) \quad . \quad . \quad . \qquad (16)$$

$$\hat{\gamma}_n = (Z'Z)^{-1}Z'\left(Y - X\hat{\beta}_n\right) \quad . \quad . \quad . \qquad (17)$$

$$\hat{g}_{nj}(t) = \sum_{w=1}^{m_n} B_w(t)\,\hat{\gamma}_{jw} \;;\, j = 1,\ldots,J \quad . \quad . \quad . \qquad (18)$$

3-2-2 Smoothly clipped absolute deviation (SCAD)

Fan and Li (2001) [8] suggested the penalty SCAD method and prove that SCAD possibilities achieve the oracle properties, since SCAD is given a penalty function defined within the interval $[0,\infty)$, the penalty likelihood of nonparametric and semiparametric models is used to balance between model complexity and estimation accuracy. where the penalty objective function is defined as follows: [5]

$$L_p(\beta,\gamma) = \frac{1}{2}\sum_{i=1}^{n}\left[Y_i - \{\gamma^T B(Z_i) + X_i^T\beta\}\right]^2 + n\sum_{j=1}^{d} p\lambda_j(|\beta_j|) \quad . \quad . \quad . (19)$$

Since $p\lambda_j(.)$ are a penalty function and λ_j is the tuning parameter (which can be selected in several data-driven methods). it should be noted that the penalty functions and tuning parameters are not necessarily the same for all parameters.

$$p\lambda_j(|\beta_j|) = 0.5\ \lambda_j^2\ I\{|\beta_j| \neq 0\} \ \text{ as } \ I\{|\beta_j| \neq 0\}$$

This method develops criteria for selecting classical variables such as BIC, AIC, and RIC for APLM as follows:

$$L(\gamma,\beta) + \frac{n}{2}\sum_{j=1}^{d}\lambda_j^2 \; I\{|\beta_j| \neq 0\} \qquad . \quad . \quad . \qquad (20)$$

also, $\sum_{j=1}^{d} I\{|\beta_j| \neq 0\}$ is equal to the size of the specified models or the number of explanatory variables selected in the model. the RIC, BIC, AIC take $\left[\lambda_j = \sqrt{2/_{n\sigma}} \; , \; \sqrt{\log(n)/_{n\sigma}} \; , \; \sqrt{\log(d)/_{n\sigma}}\right]$ Respectively.

It was noted that the Bridge regression is equivalent to the penalty of L_q, as it was $P\lambda_j(|\beta_j|) = q^{-1}\lambda\,|\beta_j|^q$ and Lasso corresponds to L_1 and SCAD corresponds to the following functions:

$$p\lambda(|\beta_j|) = \begin{cases} \lambda\,|\beta| & \text{if } 0 \le |\beta| < \lambda \\ \dfrac{(a^2-1)\lambda^2 - (|\beta| - a\lambda)^2}{2\,(a-1)} & \text{if } \lambda \le |\beta| < a\lambda \\ \dfrac{(a+1)\,\lambda^2}{2} & \text{if } |\beta| \ge a\,\lambda \end{cases} \quad . \; . \; . \; (21)$$

where the parameter (a) can be estimated by the cross-validation criterion and (Fan and Li) suggested using (a = 3.7) as the optimal value, the SCAD method is a improvement of LASSO in terms of model bias and the Ridge regression in the case $(q > 1)$ in terms of stability, in addition, it has oracle properties. [11]

$$\text{let } \beta_0 = [\beta_{10}, \ldots, \beta_{d0}]^T = [\beta_{10}^T, \beta_{20}^T]^T \qquad . \quad . \quad . \qquad (22)$$

real values of (β) without generality loss, and assuming $(\beta_{10})$are all nonzero components and (β_{20}) are the parameters whose values are zero and let S be the length of (β_{10})(the number of parameters for β_{10}).

$$a_n = \text{Max}_{1\le j\le d}\{|p'\lambda_j(|\beta_{j0}|)|;\ \beta_{j0} \neq 0\} \qquad . \quad . \quad . \qquad (23)$$

$$b_n = \text{Max}_{1\le j\le d}\{|p''\lambda_j(|\beta_{j0}|)|;\ \beta_{j0} \neq 0\} \qquad . \quad . \quad . \qquad (24)$$

$K_n = \{p'\lambda_1(|\beta_{10}|)sgn(\beta_{10}), .., p'\lambda_s(|\beta_{s0}|)sgn(\beta_{s0})\}^T$ . . . (25)

$\Sigma_\lambda = diag\{p''\lambda_1(|\beta_{10}|), .., p''\lambda_s(|\beta_{s0}|)\}$ . . . (26)

Theorem: Assuming $[b_n \to 0]$ and $\left[a_n = o(n^{-1/2})\right]$ under the following regularity conditions:

let r be a positive integer and $\{V \in (0, 1]\}$ so that$[P = r + V > 1.5]$ and assuming H is the collection of functions g on [0,1] whose the r-th derivative , $g^{(r)}$ exists and satisfies the Lipschitz condition of order r:

$$\left|g^{(r)}(Z') - g^{(r)}(Z)\right| < C\,|Z' - Z|^r \text{ for } 0 < Z';\ Z \le 1$$

Since (C) is a positive constant.

[1] Each component function $g_{0k} \in H \quad \forall \quad k = 1, \ldots, K$

[2] The distribution of (Z) is absolutely continuous and its probability density function f is bounded away from zero and infinity on $[0, 1]^K$.

[3] The random vector X satisfies that for any vector $W \in R^d$

$$C\,\|W\|^2 \le W^T E\left(\left(X^{\otimes 2}\middle|Z = z\right)\right) W \le C\,\|W\|^2$$

[4] The number of entering knots satisfies the following inequality $\left[n^{1/(2p)} << J_n << n^{1/3}\right]$ that is, in the case of (P = 2), the number of internal knots is equal to $\left[J_n \sim n^{1/4} \log n\right]$.

[5] $\Gamma(Z)$ A projection function has an additive form, that is:

$$\Gamma(Z) = \Gamma_1(Z_1) + \Gamma_2(Z_2) + \ . \ . \ . + \Gamma_K(Z_K)$$

where: $E\,[\Gamma_K(Z_K)] = 0$ and $E\,[\Gamma_K(Z_K)]^2 < \infty$

$\Gamma_K \in H \ \forall\, k = 1, \ldots, K$; under the conditions above:

1) As the probability approaches one, there is a local estimator $(\hat{\beta})$ that is smaller than the term $L_P(\beta, \gamma)$ It shall be: $\left\|\hat{\beta} - \beta\right\| = O_P\left(n^{-1/2}\right)$

2) Additionally, if the $\left\{\lambda_j \to 0\ ;\ n^{1/2}\,\lambda_j \to \infty\right\}$ as well:

$$\lim_{n \to \infty} \inf \ \lim_{u \to 0^+} \inf \ p'\,\lambda_j(u)/\lambda_j > 0$$

As the probability approaches one, the consistent estimator of $(\hat{\beta})$ for roots (n) in [1] achieves:

a) $\hat{\beta}_2 = 0$

b) $(\hat{\beta}_1)$ It has an approximately normal distribution, that is:

$$\sqrt{n}\left\{E\left(\tilde{X}_1^{\otimes 2}\right) + \Sigma_\lambda\right\} + \left[\hat{\beta}_1 - \beta_{10} + \left\{E\left(\tilde{X}_1^{\otimes 2}\right) + \Sigma_\lambda\right\}^{-1} K_u\right] \xrightarrow{D} N\left(0, \Sigma_s\right) \quad ... (27)$$

and $\Sigma_s = Var\left(\varepsilon\, \tilde{X}_1\right)$

From the above theory, it was noted that the selection procedure for the variable in a SCAD method can effectively determine the significant compounds with accompanying estimators and those who achieve the Oracle property. [11][12]

3-2-3 Elastic Net (EN)

The Elastic Net (EN) method was proposed by Zou and Hastie (2005) [14]. They point out that the Lasso method has limitations:

- **i.** If $(p >> n)$ then in this case Lasso selects approximately n variables.
- **ii.** If there is a set of highly correlated variables, lasso will select only one variable from the set and ignore the rest.
- **iii.** In the case $(n >> p)$ with a strong correlation between the explanatory variables an empirical observation of lasso prediction performance is shown it is controlled by ridge regression.

Estimates for (EN) can be obtained by the following formula:

$$L(\lambda_1, \lambda_2, \beta) = |Y - \gamma^T B(Z) - X^T \beta|^2 + \lambda_2 ||\beta||^2 + \lambda_1 |\beta|_1$$

$$L(\lambda_1, \lambda_2, \beta) = \sum_{i=1}^{n} \left[Y_i - \gamma^T B(Z_i) - X_i^T \beta \right]^2 + \lambda_2 ||\beta||^2 + \lambda_1 |\beta|_1 \quad ... (28)$$

where: $||\beta||^2 = \sum_{j=1}^{p} \beta_j^2$ and $|\beta|_1 = \sum_{j=1}^{p} |\beta_j|$

λ_2, λ_1 are non-negative EN tuning parameters, such that:

$\lambda_1 \geq 0$ and $\lambda_2 \geq 0$

n is the sample size $(i = 1,2, \ldots, n)$.

β_j^2: represents the square of parameters related to the penalty function of the ridge method.

$|\beta_j|$: represents absolute parameters related to the Lasso penalty function.

$[\gamma]$ spline parameters vector.

$B(Z_i)$ B-Spline basis function

with the following conditions $[Z'Z\,\gamma_n = Z'\,(Y - X\beta_n)]$ and

$[P_Z = Z\,(Z'Z)^{-1}Z']$ where (P_Z) is the projection matrix for the column space of Z.

$(I - P_Z)$ are an idempotent matrix and the eigenvalues for $(I - P_Z)$ are 0 or 1. [6]

therefore, equation (28) can be rewritten as:

$$L(\lambda_1, \lambda_2, \beta) = \|(I - P_Z)(Y - X\beta_n)\|_2 + \lambda_2|\beta|^2 + \lambda_1|\beta|_1 \quad . \; . \; . \quad (29)$$

The estimator $(\hat{\beta})$ by (EN) method reduces the sum of squares of penalty errors shown in Equation (29):

$$\hat{\beta} \equiv \hat{\beta}\,(\lambda_1, \lambda_2) = \arg\min_{\beta} L(\lambda_1, \lambda_2, \beta) \quad . \; . \; . \quad (30)$$

$$\hat{\gamma}_n = (Z'Z)^{-1}Z'\left(Y - X\hat{\beta}\right) \quad . \; . \; . \quad (31)$$

$$\hat{g}_{nj}(t) = \sum_{w=1}^{m_n} B_w(t)\,\hat{\gamma}_{jw} \;\; ; \;\; j = 1, \ldots, J \quad . \; . \; . \quad (32)$$

The (EN) method combines the two methods of Lasso and Ridge, and they also prove that the (EN) method has the property of oracle property.[9]

3-2-4 Adaptive Elastic Net (AdEN)

This method was suggested by Zou and Zhang (2009)[15] that combines the features of Adaptive Lasso and Elastic Net as the Adaptive Lasso method in some cases is unstable, especially in the case of high-dimensional data, due to the emergence of the problem of collinearity, while the Elastic Net method lack to the Oracle feature. for these reasons, the two methods were combined to get a method that has good properties of the chosen model and is defined by (AdEN) that works on the penalty square error loss function using a combination of penalty function L_2 and adaptive penalty function L_1.

To estimate (AdEN), this method can be summarized according to the following steps:

1) Estimate the parameters of the Lasso $\hat{\beta}_{(\text{Elastic Net})}$ model.

2) Estimate the weights $\widehat{W}_j$: $\widehat{W}_j = \left(|\hat{\beta}_{j(\text{Elastic Net})}|\right)^{-\theta}$

where θ is a value positive and defined as:

$\theta > \frac{2V}{1-V} \quad ; \quad 0 \leq V < 1$

3) Find the AdEN estimator with the following formula:

$$\hat{\beta}_{(\text{Lasso})} = \arg\min_{\beta} \sum_{i=1}^{n} \left[Y_i - \{\gamma^T B(Z_i) + X_i^T \beta\} \right]^2 + \lambda \sum_{t=1}^{p} |\beta_t| \quad . \; . \; . \; (33)$$

with the following conditions $[Z'Z\,\gamma_n = Z'\,(Y - X\beta_n)]$

$Z = (Z_1, \ldots, Z_j)$ represents total design matrix of degree $(n \times J_{mn})$

and $[P_Z = Z\,(Z'Z)^{-1}Z']$ where (P_Z) is the projection matrix for the column space of Z, and $[I - P_Z]$ are an idempotent matrix and the eigenvalues for $(I - P_Z)$ are 0 or 1. [10]

$$L_{n1}(\beta_n; \lambda_{n1}) = \|(I - P_Z)(Y - X\beta_n)\|^2 + \lambda_{n1} \sum_{k=1}^{p} \|\beta_{nk}\|_2 \quad . \; . \; . \; (34)$$

Estimator Lasso is:

$$\tilde{\beta}_n \equiv \tilde{\beta}_n(\lambda_{n1}) = \arg\min_{\beta_n} L_{n1}(\beta_n, \lambda_{n1}) \quad . \; . \; . \; (35)$$

using Lasso estimates in (35) to get the weights of the Adaptive Elastic Net (AdEN) penalty function, since:

$$w_{nk} = \begin{cases} \left(|\tilde{\beta}_n|\right)^{-\theta} & \text{if } \|\tilde{\beta}_n\|_2 > 0 \\ \infty & \text{if } \|\tilde{\beta}_n\|_2 = 0 \end{cases} \quad . \; . \; . \; (36)$$

the adaptive elastic net (AdEN)objective function is as follows:

$$\hat{\beta}(\text{AdEN}) = \left(1 + \frac{\lambda_2}{n}\right) \arg\min_{\beta} \|(I - P_Z)(Y - X\beta_n)\|_2 + \lambda_{n2}\|\beta\|^2 + \lambda_{n1}\sum_{j=1}^{p}\hat{W}_j \quad . . . \quad (37)$$

where: λ_2, λ_1 are the non-negative AdEN tuning parameters

$\lambda_1 \geq 0$ and $\lambda_2 \geq 0$; $\|\beta\|^2 = \sum_{j=1}^{p}\beta_j^2$

β_j^2: represents the square of parameters related to the penalty function of the ridge method.

$|\beta_j|$: represents absolute parameters related to the Lasso penalty function.

$[\gamma]$ spline parameters vector. $B(Z_i)$: B-Spline basis function.

thus, the adaptive elastic net (AdEN)estimator is:

$$\hat{\beta}_n \equiv \hat{\beta}\,(\lambda_{n1}, \lambda_{n2}) = \arg\min_{\beta} L(\lambda_{n1}, \lambda_{n2}, \beta) \quad . . . \quad (38)$$

$$\hat{\gamma}_n = (Z'Z)^{-1}Z'\left(Y - X\hat{\beta}_n\right) \quad . . . \quad (39)$$

$$\hat{g}_{nj}(t) = \sum_{w=1}^{m_n} B_w(t)\,\hat{\gamma}_{jw} \quad ; \quad j = 1, \ldots, J \quad . . . \quad (40)$$

4. Practical Part

This chapter includes the practical aspect in which the real accounting data are employed for several Iraqi companies. The factors that influence the market value of the name were identified and the data were obtained. The results of the practical application were obtained through programs written in the R language.

4.1 Description of the sample and study variables

The study sample consists of the Baghdad Soft Drinks Company and they are considered one of the Iraqi commercial companies listed in the Iraq Stock Exchange during the period from 2012 AD to 2019 AD, as its data is regularly presented through this market during the years of the study, as it was selected according to the following conditions, which are to be listed and traded Its shares during the above period and the possibility of obtaining the data or information necessary to complete the study, as for the research variables, they were 47 variables (Market value or stock market value, Dividend distribution, Financial leverage, Return on assets, Sales growth rate, Annual Growth Rate, Company Size, Operating cash flow, Liquidity ratios, The amount of investment spending, Issuance of new shares, Net Working Capital, Accounts Receivable Turnover, Average Collection Period, Inventory Turnover, Average Age of Inventory, Operating Cycle, Payable Turnover, Payable Turnover in Days, Fixed Asset Turnover, Total Asset Turnover, Debt Ratio, Leverage Ratio, Debt / Equity Ratio, Times Interest Earned, Return on Common Equity, Return on Total Assets, Return on Investment, Gross Profit Margin, Profit Margin, Earnings Per Share, Price / Earnings Ratio, Dividend Yield, Dividend Payout, Book Value Per Share)

4-2 Estimation

The Estimation and Variables Selection for Estimation and Variables Selection by using methods (ALasso, SCAD, EN, AdEN). based on a program written in R language, and the results were as follows:

Table 1. Results of the study: Estimation and variables selection, of the coefficients using the additive partial linear model (APLM).

j	ALasso		SCAD		EN		AdEN	
	β_j	SE	β_j	SE	β_j	SE	β_j	SE
1	0.8222	1.0742	0.7174	0.3266	0.7312	0.109	0.7437	0.9783
2	-0.2214	0.7243	Removed	-	Removed	-	-0.2325	0.6981
3	Removed	-	Removed	-	Removed	-	Removed	-
4	-0.1075	0.4545	Removed	-	Removed	-	-0.1150	0.4033
5	0.6344	1.6591	0.5371	0.319	0.5662	0.1135	0.6395	1.488

6	0.7540	1.1552	0.7351	0.2756	0.7355	0.108	0.7691	1.0567
7	0.1954	0.7246	Removed	-	Removed	-	0.2005	0.6376
8	Removed	-	Removed	-	0.6861	0.1032	Removed	-
9	0.9042	2.5532	0.9974	0.2938	0.9683	0.1137	0.9250	2.1501
10	-0.4708	1.2129	Removed	-	-0.5603	0.1109	-0.4788	1.0784
11	0.4928	0.6454	0.5752	0.1842	0.5484	0.1074	0.5093	0.6104
12	Removed	-	-0.2435	0.032	Removed	-	Removed	-
13	0.5162	0.9385	Removed	-	0.5079	0.0996	0.5254	0.8206
14	-0.6566	1.4784	-0.6321	0.2173	-0.6917	0.1	-0.6828	1.3345
15	0.7264	0.6053	0.8182	0.0159	0.8091	0.0978	0.7582	0.555
16	0.7402	1.2349	0.7927	0.3058	0.7886	0.1089	0.7571	1.1108
17	Removed	-	Removed	-	Removed	-	-0.4809	2.3079
18	-0.3004	0.8353	Removed	-	Removed	-	-0.3224	0.7607

19	Removed	-	Removed	-	Removed	-	Removed	-
20	Removed	-	Removed	-	Removed	-	0.1943	0.9018
21	0.1607	0.3364	Removed	-	Removed	-	0.1608	0.3236
22	0.4846	1.2116	Removed	-	0.4846	0.1004	0.4952	1.071
23	Removed	-	Removed	-	Removed	-	Removed	-
24	Removed	-	Removed	-	Removed	-	Removed	-
25	-0.2136	0.8526	Removed	-	Removed	-	-0.2309	0.8042
26	Removed	-	Removed	-	Removed	-	Removed	-
27	Removed	-	Removed	-	Removed	-	Removed	-
28	3.5238	1.2278	4.1721	0.7086	3.5370	0.0453	3.6315	1.1012
29	Removed	-	0.5287	0.2903	0.5434	0.1163	Removed	-
30	Removed	-	Removed	-	Removed	-	Removed	-
31	Removed	-	Removed	-	Removed	-	Removed	-

32	Removed	-	0.5336	0.2903	0.5391	0.1164	Removed	-
33	Removed	-	0.5195	0.2903	0.5454	0.1163	Removed	-
34	Removed	-	0.5232	0.2903	0.5770	0.1162	Removed	-
35	Removed	-	0.5358	0.2903	0.5350	0.1164	Removed	-
36	Removed	-	0.5240	0.2903	0.5432	0.1164	Removed	-
37	Removed	-	Removed	-	Removed	-	Removed	-
38	Removed	-	Removed	-	Removed	-	Removed	-
39	Removed	-	0.5346	0.2903	0.5392	0.1164	Removed	-
40	Removed	-	0.0800	0.032	Removed	-	Removed	-
41	Removed	-	-0.2260	0.032	Removed	-	Removed	-
42	Removed	-	-0.1981	0.032	Removed	-	Removed	-
43	Removed	-	-0.3185	0.032	Removed	-	Removed	-
44	Removed	-	Removed	-	Removed	-	Removed	-

45	Removed	-	Removed	-	Removed	-	Removed	-
46	Removed	-	-0.2624	0.032	Removed	-	Removed	-
47	0.4678	0.7201	0.4644	0.2583	0.4853	0.1028	0.4738	0.6553

4-3 Comparison

A comparison was made between the estimation and variables selection based on each of the criteria mean squares error, prediction error and coefficient of determination and the results were as follows:

Table 2. The criteria represent a comparison between the methods.

Methods	MSE	R^2
ALasso	0.1296	86.62%
SCAD	0.1541	84.01%
EN	0.1383	85.73%
AdEN	0.0608	93.73%

Through the results of the above table and through the comparison criteria described, we note that the value of the mean squares error and prediction error is less for the Adaptive Elastic Net (AdEN) method than the value of the other methods. The values of the coefficient of determination for the Adaptive Elastic Net method are greater than the values of the coefficient of determination for the other methods.

It can also be illustrated by the drawing It can also be illustrated by the drawing:

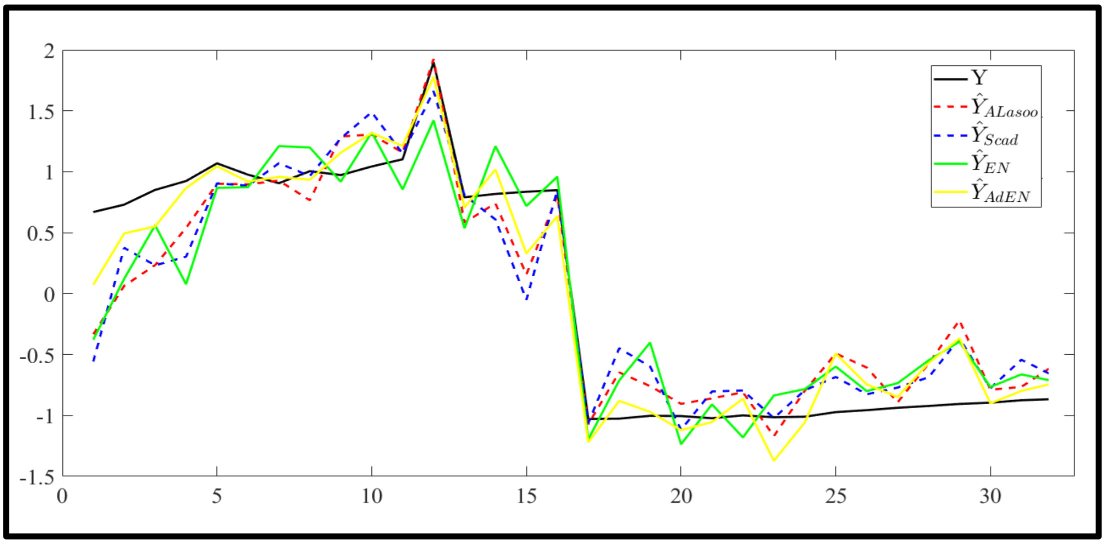

Figure 1. shows the real and estimated values in for ways for the (ALasso, SCAD, EN, AdEN) methods.

The above figure shows the amount of bias between the real values and the estimated values of the response variable values estimated by for ways for the (ALasso, SCAD, EN, AdEN) methods.

5-Conclusions

1. The additive partial linear model (APLM) to reduce a high-dimensional problem. It is considered more flexible than other models because it contains multiple linear components and multiple nonlinear components and additive.
2. The Adaptive Elastic Net (AdEN) is more efficient than the other methods, so it is preferred to use it to Estimation and Variables Selection for Semi – Parametric Additive model.
3. It is possible to apply this APLM in a number of fields, including health, educational, social, financial, economic, and others, which have high flexibility in their inclusion on more than one linear explanatory variable as well as more than one non-linear parametric function. These features make the model highly flexible in various applications.
4. Through the results of the analysis, it was noted that there are some variables with an important impact that must be focused and followed up by the shareholders and beneficiaries.

References

[1]. Akaike, Hirotugu (1973)" **Maximum likelihood identification of Gaussian autoregressive Moving average Models**" Biometrika, Vol.60, No.2. pp. 255 -265. http://www.jstor.org/stable/2334537

[2]. Hmood, M. Y. and Saleh, T. A. (2016), "**Comparing some Penalized methods in analyzing the semi-parametric single index model with practical application**", Journal of economics and administrative sciences, Vol. 90, No. 22, pp. 407-427.

[3]. Hmood, M. Y. and Kateh, M. M. (2014), "**Comparison of estimations of a semi-parametric model using different smoothing methods**", Journal of economics and administrative sciences, Vol. 75, No. 20, pp. 376-394.

[4]. Fathi, Iman Tariq (2012) "**Using information criteria methods and model diagnostic methods to choose the best multiple linear regression model with application to children with thalassemia patients in Mosul**" Journal of Education and Science - Volume 25 Issue 2, Pages 188-200.

[5]. Hammoudi, Nader Iskandar (2018) " **Using the AIC and Schwartz (SC) Information Standards in Comparing Nonlinear Growth Models for Different Fish Species**" University of Research Journal - Volume 40, Issue 3, Pages 45-66.

[6]. Hua Liang and Haobo Rne (2014) "**Generalized Partially linear Measurement Error Models**" Journal of Computational and Graphical statistics Vol. 14, No. 1 pp. 237-250. http://dx.doi.org/10.1198/106186005X37481

[7]. Hurdle, Wolfgang and Liang, Hua and Gao, Jiti (2000) "**partially Linear Models**", Institut fur Statistik und Okonometrie Humboldt – University at zu Berlin D-10178 Berlin, Germany https://mpra.ub.uni-muenchen.de/39562/

[8]. Jianqling Fan and Runze Li (2001) **"Variable Selection Via No concave Penalized Likelihood and its oracle Properties**" Journal of the American Statistical Association, Vol. 96, No. 456, pp. 1348–1360

[9]. Lian, H, Hua-Liang and David Ruppert (2013) "**Separation of covariates into Non-Parametric and Parametric Parts in High- Dimensional partially Linear Additive models**" Statistical Sinica Vol.25 No.2 pp.591-607

[10]. https://www.jstor.org/stable/24311036#:~:text=https%3A//www.jstor.org/stable/24311036

[11]. Wang, Q .and Yin, X (2008)" **A nonlinear Multi-Dimensional Variable Selection Method for high dimensional data: sparse MAVA**" computational statistics. and data Analysis Vol.52, No.9, PP. 4512 - 4520. https://doi.org/10.1016/j.csda.2008.03.003

[12]. Xinyu Zhang and Wendun wang (2019) " **Optimal Model Averaging Estimation for partially Linear Models** " Statistical Sinica Vol. 29, No. 2 pp. 693-718.

[13]. Xuewen Lu and Peter X. K, song (2015) "**Efficient Estimation of the Partly linear Additive Hazards Model with Current status data**" Scandinavian Journal of statistics, Vol. 42 , No.1, P.P. 306- 328.

[14]. Zhou, Yu and Zhu, Lipping (2013) "**Dimension reduction and predictor selection in Simi Parametric Models**" Biometrika, Vol.100, No.3, PP. 641 - 654.

[15]. Zou, H (2012) "**The adaptive Lasso and It's oracle properties**". Journal of the American statistical Association Vol.101, N0.476 1418- 1429. http://amstat.tandfonline.com/loi/uasa20

[16]. Zou, H. and Zhang, H.H. (2009)."**On the adaptive elastic net with diverging number of parameters**" Annals of statistics, Vol.37, No.4 , 1733-1751.