Estimation of Semi-Parametric Models Using Some Penal Methods, Multiple Index Model as an Applied Example

Abstract. In this study, some methods for estimating parameters and selecting the significant variable for the Semi-parametric Multiple Index Model (SMIM) were used at the same time. There are a large number of explanatory variables, and to avoid missing any of the important explanatory elements, special methods were used to select the significant variables. therefore, in this case, the use of parametric and non-parametric estimation methods produces poor performance, especially in the case of increasing the dimensions (curse of dimensionality). Accordingly, this research came to shed light on some semi-parametric methods for analyzing the multiple index model, which work on estimating the model and selecting the significant variables at the same time. These methods allow for dimensionality reduction, thus increasing the accuracy of the estimation while allowing greater flexibility and less risk in identifying errors.To achieve the goal of the study, some methods based on different penalty functions were used, which work on estimating and selecting variables for the semi-parametric multiple indexmodel at the same time. One of these methods is the (Sg MAVE-Lasso) and (Sg MAVE-MCP).
Key word: SMIM, MAVE,MAVE-Lasso, MAVE-MCP, Groupwise

INTRODUCTION

Most of the models used in regression analysis contain a set of unknown parameters, which are to be estimated and represent the parametric regression model, which often does not take into account the nonlinear effect of the explanatory variables.[1] On the other hand, there is another part that represents the nonparametric regression model, which is represented by the link function or the conditional distribution function g(.), and the non-parametric model depends on the estimation from the data directly and is not bound by the assumptions of the parametric model, so it has more flexibility, but the non-parametric model may not fulfill all the required assumptions and does not work completely, and also suffers from the problem of increasing the dimensions[2], or what is known as the curse dimensionality, i.e., increasing the number of explanatory variables,

and therefore it is difficult to know and choose the effective effective variables, which also suffers from the problem of self-correlation. (Autocorrelation). Therefore, there was a need for a model that exceeds these restrictions and assumptions and gives more accurate results, which is known as the semip-arametric regression model, which addresses the problem of increasing dimensions and the non-linear effect of variables. In other sciences, in which the explanatory variables are under one linear index (ßTX). Semi-parametric methods are used as a compromise between parametric (constrained) and non-parametric (flexible) models[3].

Related work

In (2012), the researcher (Huang et al) proposed a method to distinguish the linear and non-linear components in the semi-parametric models, as this approach identifies the parametric and non-parametric components of the semi-parametric model according to the data and assuming that the structure of the model is known, and the researcher explained that it provides flexible models for joint effects on The regression response variable combines the flexibility of non-parametric regression with the stability of linear regression.

In (2014), researchers (Xu Guo, Wangli Xu and Lixing Zhu) presented a study entitled "Multiple Index Models with Missing Covariates" in which they suggested the model formula as follows: $Y=g(\theta_0^T X) + \epsilon$ and they took into consideration Estimating the model by means of a weighted estimation equation by calling the inverse when the probability of choice is known in advance and is estimated in a parametric and non-parametric manner. The real choice. The research was applied to the clinical data of AIDS patients to find the most efficient method, and the research showed that the multiple index model is widely used in many statistical, economic, and other fields.

In (2021), the researchers (Lee & Wang) conducted a study in which they presented assumptions to reduce the sizes of the variables of the multiple model to reduce the problem of dimensions and thus choose the best model, and the study showed that the multiple model contributes to reducing the sizes of the explanatory variables to choose the effective variables more accurately one index model.

In (2021), (Chaohua Dong and others) presented an economic statistical study entitled (Multiple Index Model with Unstable Time Series Models, Theory and Application), where three types of time series were selected and a set of considerations were reached, namely :

First: The use of the multiple index model and the estimates associated with it allow uses and applications in various fields, especially economic and financial.

Second: The model building mechanism allows finding reliable estimates for data with extreme and high dimensional values and bypassing the spatio-temporal problem.
Third: The proposed models are applicable to several types of data, such as fixed and non-fixed time series data.
Fourth: The proposed estimation processes enjoy sobriety and can be used mathematically in the economic, financial and other fields.

Semi-parametric regression models

Recently, researchers have relied on a new statistical method that integrates parametric regression functions with non-parametric regression functions at the same time. It is called the semi-parametric regression model, which works on regression analysis as a statistical method to study the explanatory variables. The term semi-parametric was proposed or developed by researcher (Oakes) in (1981)[4], which refers to two categories of models, the first category includes an unknown function of data that must be estimated in addition to the parameters (parametric part), and the second category does not include parameters but two or more unknown functions that can be estimated (the nonparametric part), Inaccurate application of parametric models may result in biased estimates and false inferences. On the other hand, nonparametric models provide great flexibility to know the shape of the function and have recently gained wide use in many fields to avoid the pitfalls of parametric models, but they suffer from the problem of increasing the dimensions that were referred to previously. When there are a large number of explanatory variables, where the data becomes random or scattered (sparsity), and thus the estimate is unreliable, and here nonparametric models are rarely used to achieve accurate results, Therefore, the above reasons and problems prompted researchers to turn to more reliable modern methods, which are represented by semi-parametric methods or models, which produce correct inferences in the event that conditions or hypotheses are not met, or the data is characterized by non-linearity.[1]

Semi parametric multi index model(SMIM)

The expansion of statistical applications and their entry into various fields and sciences prompted researchers to search for a more comprehensive model or method than the single index model, in line with the large number of explanatory variables, their correlation, and their effectiveness. (Li) was the first researcher to write about the multiple index model in the Journal of the American Statistical Association in 1991 as follows:[5]

$$Y= m(XT\beta 1 , \dots , XT\beta k \ , \varepsilon)$$

.Where (ß1, …, ßk) this is known as the unknown predictive vectors to be estimated.

m(.) : the unknown link function.
ε : random error.

According to the researcher's opinion, this model describes the state of the response variable (Y) depending on the P dimensions of the explanatory variable X through multiple indicators (XTß1 , … , XTßk) , after that, research and studies followed, and the model was used in many economic, banking and other fields, where researchers (H.Ichiura & Lang-fei) from the University of Michigan Department of Economics and Mathematics (1993) published research on the multiple model entitled semi-parametric estimation of the multiple index model in the form the following:[6]

$$Yi = f(X_i) + \varepsilon i = m(\beta_1' X_i, \beta'_2 X_i, \dots, \beta'_m X_i) + \varepsilon i = m(\theta X_i)$$

Properties of semi-parametric multiple index model

The ideal Semi-parametric Multiple Index (SMIM) model has a number of advantages, as follows: [17]

1. The multiple index model allows modeling the correlations between explanatory variables where the formula: Y= g (XTß1 , … ,)Working on reduce error variance.
2. The estimators resulting from the multiple model can achieve near-perfect statistical convergence even when the response variable is affected by large explanatory variables.
3. The multiple model is as accurate as the parametric model in estimating the parameter vector (β) and as accurate as the non-parametric model in estimating the link function.
4. It helps to reduce the sizes of the explanatory variables, which helps to reduce the problem of increasing the dimensions, and it chooses the effective variables more accurately than the other models, so it has a wide range of applications.
5. Consistency in the selection of variables is important as the correct model includes the important index set that represents non-zero (significant) parameters and also includes the unimportant index set that represents zero (non-significant) parameters.

Penalty function

The main idea that the penalty function operates on is that it prevents the emergence of a problem over the application (over fitting). And it means that the model with all its input variables (effective and others) may be below the desired ideal level, meaning that the dependent variable (response) depends only on a few important explanatory variables (effective), and thus produces inefficient estimators that do not have the smallest possible variance [12]. Therefore, we always strive to get rid of this problem by deleting the explanatory variables that are associated with other ineffective or unimportant variables,

thus getting rid of the problem of multilinearity. Therefore, the result of the penalty function is compensated by a large reduction of the sum of squared errors. That is, we try to make (SSE) as small as possible, but at the same time, the penalty limit will push us towards obtaining very large parameters, as most of the parameters are not zero with a large penalty limit, and therefore the estimate for any unimportant or ineffective parameter is zero for the penalty least squares estimates for the effect of the variable and then the automatic selection of the appropriate model. It is necessary to know that the penalty function depends mainly on the penalty parameter, also known as the (tuning parameter). and denoted by the symbol (λ), and the different penalty functions lead to different penalties for choosing variables[16]. In the following, we explain the penalty functions used in this research:

1. Lasso Penalty function:

It was suggested by the researcher (Tibshirani) (1996) and it is also known as the (L1) penalty function. It means (Least absolute shrinkage and selection operator penalty function) and takes the following formula: [15]

$$P_{Lasso}(\beta) = \lambda \left|\beta_j\right|$$

2. MCP Penalty function:

It was suggested by a researcher (Zhang) (2010) that it means (Minimax concave Penalty function) and takes the following formula:[14]

$$P_{MCP}(\beta) \begin{cases} \lambda |\beta| - \dfrac{\beta^2}{2a} & \text{if } |\beta| \le a\lambda \\ \dfrac{a\lambda^2}{2} & \text{Otherwise} \end{cases}$$

Estimation methods and algorithms

A set of methods for estimating and selecting variables for semi-parametric models in general has appeared, including the semi-parametric multiple index model. We will discuss these methods that helped improve the accuracy of the model. This study's two methods are discussed below:

1. Shrinkage groupwise MAVE with lasso penalty function(S.G MAVE-Lasso)

The researcher (Tao Wang et al. 2015) proposed this method,[8][13] which combines the Lasso-MAVE method and the smart shrinkage groupwise (Shrinkage groupwise). The penalty function Lasso was proposed by the researcher (Tibshrani) in 1996, which works

on selecting transactions with the least absolute shrinkage (Least absolute shrinkage and operator selection method), and it is one of the proposed penal least squares methods (PLS). The method of reducing the groupwise with the least variance (with the Lasso function) reduces the sum of the squares of the residuals under the following constraint: (the sum of the absolute values of the coefficients is less than a certain constant, let it be (t), which represents the reduction parameter:[12]

$\widehat{B}$Lasso = argmin SSE subject to $\sum_{j=1}^{p}|\widehat{B}_{Lasso}| \leq t$, According to this method, the central averages of the groupwise can be obtained by minimizing the following objective function:

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\left\{ y^j - a^i - \sum_{\iota=1}^{g} b_\iota^{iT} B_\iota^T (V_\iota^j - V_\iota^i) \right\}^2 W_j^i$$

Where $(Y^j - V^i)$, $i=1, \dots , n$ is a random sample from (Y, V) , $a^i \in \mathbb{R}$, $b_1^i \in \mathbb{R}^{d1}$, … , $b_g^i \in \mathbb{R}^{dg}$, $i= 1,\dots,n$, B_ι is a matrix $\in \mathbb{R}^{p_\iota X\, d_i}$

If the value ($\widetilde{B}= \bigoplus_{\iota=1}^{g} \widetilde{B}_\iota$) represents the groupwise with the minimum variance (MAVE), then the model estimate for the smart reduction group is defined as follows:

$\widehat{B} = \bigoplus_{\iota=1}^{g}$ diag $(\widehat{\alpha_\iota})\, \widetilde{B}$, where($\widehat{\alpha_\iota}$) represents the shrinkage index vectors . therefore, in an equal manner, the contraction index vectors are reduced according to the following equation:

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\left[\left\{ y^j - \tilde{a}^i - \sum_{\iota=1}^{g} \tilde{b}_\iota^{iT} \widetilde{B}_\iota^T \text{diag}\,(V_\iota^j - V_\iota^i)\alpha_\iota \right\}^2 \widetilde{W}_j^i + \lambda_n \sum_{\iota=1}^{g}\sum_{s=1}^{p_\iota} |\alpha_{\iota s}| \right]$$

Based on the foregoing and previous constraints, the (Lasso)-(groupwise MAVE) method tends to form coefficients equal to zero and thus works to contract and thus produces interpretable models.

The algorithms

Estimating and selecting variables in this method is according to the following algorithm:[9]

step (0): The first step is to obtain an initial estimate for the parameter by using the ordinary least squares method (ols).

step (1): And it is installed $\hat{\beta}_{(0)} = \hat{\beta}$, the solution vectors are calculated for the position constants $(\hat{a} , \hat{b})$ according to the following formula:

$$(\hat{a} , \hat{b}) = \text{argmin} \sum_{i=1}^{n}\sum_{j=1}^{n}\left[Y_i - \left\{ a_j + b_j^T B^T (V_i - V_j) \right\} \right]^2 W_{ij}$$

$$W_{ij} = \frac{K_h \left\{ B^T \left(V^j - V^i \right) \right\}}{\sum_{j=1}^{n} K_h \left\{ B^T (V^j - V^i) \right\}}$$

step (2): In this step, the parameter vector β is estimated and set up ($\hat{a}$, $\hat{b}$) using the following formula:

$$\hat{\beta}^{G.MAVE-LASSO} = \text{argmin}\sum_{i=1}^{n}\sum_{j=1}^{n}\left[Y_i - \{a_j + b_j^T B^T(V_i - V_j)\}\right]^2 W_{ij} + \lambda \sum_{j=1}^{p}\left|\hat{B}_J\right|$$

Where the $\lambda \sum_{j=1}^{p}\left|\hat{B}_J\right|$ is the penalty function LASSO.

step (3): The two steps (1) and (2) are repeated until the convergence rates with the parameter vector β are obtained.

2. Shrinkage groupwise MAVE with MCP penalty function) (S.G MAVE-MCP):

This method deals with the estimation and selection of parameters with the least mean variance of the wise shrinkage group with the penalty function (MCP), and it combines the two methods (MCP-MAVE) and the (Shrinkage groupwise). The (MCP) function is the (Minimax concave penalty) of the functions. Concave penalties, which are characterized by quality in estimation and selection,[11] where the characteristics of Oracle are achieved as a result of the fact that the algorithms use concave penalties that enable them to converge at different optimal values. The penalty function (MCP) was proposed by researcher (Zhang) in (2010). When the two methods (MCP-MAVE) are combined, a concavity is produced in the penalty loss points at certain thresholds for variable selection and impartiality, and the logic behind the penalty function (MCP) can be understood through the derivative of the function:[10]

$$\acute{P}MCP(\beta) = \begin{cases} (\lambda - \frac{|\beta|}{a}) & \text{if } |\beta| \le a\lambda \\ 0 & \text{Otherwise} \end{cases}$$

The MCP-MAVE method for estimating and selecting variables was proposed by researchers (Alkenani and Yu) in (2013) as follows:[10]

$$\min_B \left(\sum_{i=1}^{n}\sum_{j=1}^{n}\left[Y_i - \{a_j + (X_i - X_j)^T B b_j\}\right]^2 W_{ij} + n\sum_{j=1}^{p} P_{MCP}(|\beta_j|) \right)$$

The following can be said about the penalty estimator under (MCP) of the general linear regression model:

$$\hat{\beta}_{MCP} = \text{argmin}\left\{ \sum_{i=1}^{n} (Y_i - \sum_{j=1}^{p} B_j X_{ij})^2 + \lambda|\beta| - (\beta^2/2a) \right\}$$

Thus, under the MCP-MAVE method, it is possible to estimate and reduce shrinkage vectors and improve the quality of the estimation as follows:

$$\left[\left\{ y^j - \tilde{a}^i - \sum_{\iota=1}^{g} \tilde{b}_\iota^{iT}\tilde{\tilde{B}}_\iota^T \text{diag}\,(V_\iota^j - V_\iota^i)\alpha_\iota \right\}^2 \tilde{W}_j^i\right] + \sum_{i=1}^{n}\sum_{j=1}^{n}\left[Y_i - \{a_j + (X_i - X_j)^T B b_j\}\right]^2 + n\sum_{j=1}^{p} P_{MCP}(|\beta_j|)$$

Where the above equation represents the estimation and selection of variables with the lowest rate of variance for the smart contraction group with the (MCP-MAVE) method. Therefore, the (MCP-MAVE) method is another alternative for estimating and selecting the coefficients with the shrinkage groupwise and thus obtaining the least biased regression coefficients in the scattered models.

The algorithms

Estimating and selecting variables in this method is according to the following algorithm: [9]

step (0): The first step is to obtain an initial estimate for the parameter by using the ordinary least squares method (ols).

step (1): and it is installed $\hat{\beta}_{(0)} = \hat{\beta}$, the solution vectors are calculated for the position constants ($\hat{a}$, $\hat{b}$) according to the following formula:

$$(\hat{a}, \hat{b}) = \operatorname{argmin} \sum_{i=1}^{n} \sum_{j=1}^{n} \left[Y_i - \{ a_j + b_j^T B^T (V_i - V_j) \} \right]^2 W_{ij}$$

$$W_{ij} = \frac{K_h \{ B^T (V^j - V^i) \}}{\sum_{j=1}^{n} K_h \{ B^T (V^j - V^i) \}}$$

step (2): In this step the parameter vector β is estimated and install ($\hat{a}$, $\hat{b}$) according to the following formula:

$$\hat{\beta}^{G.MAVE-MCP} = \operatorname{argmin} \sum_{i=1}^{n} \sum_{j=1}^{n} \left[Y_i - \{ a_j + b_j^T B^T (V_i - V_j) \} \right]^2 W_{ij} + n \sum_{j=1}^{p} P_{MCP}(|\beta_j|)$$

Where the $n \sum_{j=1}^{p} P_{MCP}(|\beta_j|)$ is the penalty function MCP .

step (3): The two steps (1) and (2) are repeated until the convergence rates with the parameter vector β are obtained.

Simulation study

The simulation method was used to determine the best semi-parametric methods that were used in the research to estimate and select the variables for the Semi-parametric Multiple Index Model (SMIM). The format of the model was chosen as follows:

$$Y = g\left(\beta_1^T X \dots \beta_d^T X \right) + \epsilon_i \quad , \; i = 1,2, \dots , n ,$$

Also, two link functions $g\left(\beta_1^T X \dots \beta_d^T X \right)$ were used for models that fit most of the cases, which were used in published research dealing with the study of multiple models. In the basic stage, three experiments were studied, and they chose default values as shown in the table below, which shows the cases of the simulation study:

TABLE I. The cases of the simulation study

		$\sigma = 1$			$\sigma = 0.1$		
Experiment	**P**	n			n		
I	3	30	50	100	30	50	100
II	5	30	50	100	30	50	100
III	7	30	50	100	30	50	100

As each of the three experiments was done at two levels of standard deviation (σ = 1, 0.1), and different sample sizes (n = 30, 50, 100), and different dimensions of the explanatory variables (p = 3, 5, 7), The multiple model was estimated for all experiments using the estimation methods mentioned in the research and compared between them according to the mean squared error (MSE).

Conclusions

1. Through the repeated steps of simulation experiments based on the data in the previous table, it was noted that the first method (S.G MAVE-Lasso) in general gives lower (MSE) rates than the second method, and this shows that it is the best method for estimating the multiple model.
2. The first method of estimating with the penalty function (LASSO) gives sparse estimates for the estimated parameters ($\hat{\beta}$) , which means that explanatory variables with zero coefficients are removed from the model and the rest of the variables are kept , and process of estimating and selecting variables in this method is continuous, which allows obtaining more stable models.

Recommendations

1. We recommend the use of estimation methods that include the penalty function (LASSO), which are efficient and have important features for estimating multiple models.

2. We also recommend the use of other methods for estimating the multiple-index model with other penalty functions and other functions of the more complex model that may allow obtaining more stable models.

References

1. Horowitz, J. L., & Lee, S. (2002). Semiparametric methods in applied econometrics: Do the models fit the data?. Statistical Modelling, 2(1), 3-22.

2. Simonoff, J. S., & Tsai, C. L. (2002). Score tests for the single index model. Technometrics, 44(2), 142-151.

3. Zhu, L. P., Qian, L. Y., & Lin, J. G. (2011). Variable selection in a class of single-index models. Annals of the Institute of Statistical Mathematics, 63(6), 1277-1293

4. Powell, J. L. (1994). Estimation of semiparametric models. Handbook of econometrics, 4, 2443-2521.

5. Ichimura, H., & Lee, L. F. (1991, June). Semiparametric least squares estimation of multiple index models: single equation estimation. In Nonparametric and semiparametric methods in econometrics and statistics: Proceedings of the Fifth International Symposium in Economic Theory and Econometrics. Cambridge (pp. 3-49) .

6. Ichimura, H., & Lee, L. F. (1993). Semiparametric estimation of multiple index models. In Nonparametric and Semiparametric Methods in Econometrics and Statistics. Proceedings of the Fifth International Symposium in Economic Theory and Econometrics.

7. Hong, H. G., & Zhou, J. (2013). A multi-index model for quantile regression with ordinal data. Journal of Applied Statistics, 40(6), 1231-1245.

8. Wang, T., Xu, P., & Zhu, L. (2015). Variable selection and estimation for semi-parametric multiple-index models. Bernoulli, 242-275.

9. Wu, W., Hilafu, H., & Xue, Y. (2019). Simultaneous estimation for semi-parametric multi-index models. Journal of Statistical Computation and Simulation, 89(12), 2354-2372.

10. Al-Kenani, A. J. K. (2013). Some statistical methods for dimension reduction (Doctoral dissertation, Brunel University, School of Information Systems, Computing and Mathematics).

11. Jiang, H., Zheng, W., Luo, L., & Dong, Y. (2019). A two-stage minimax concave penalty based method in pruned AdaBoost ensemble. Applied Soft Computing, 83, 105674

12. Su, L., & Zhang, Y. (2013). Variable selection in nonparametric and semiparametric regression models.

13. Breheny, P., & Zeng, Y. (2018). Regularization paths for regression models with grouped covariates. R package version, 3-1

14. Zhang, C. H. (2010). Nearly unbiased variable selection under minimax concave penalty. The Annals of statistics, 38(2), 894-942.
15. Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58(1), 267-288
16. Huang, J., & Xie, H. (2007). Asymptotic oracle properties of SCAD-penalized least squares estimators. In Asymptotics: Particles, processes and inverse problems (pp. 149-166). Institute of Mathematical Statistics.
17. Gamarnik, D., & Gaudio, J. (2020). Estimation of monotone multi-index models. arXiv preprint arXiv:2006.02806.

.