Journal of Data Acquisition and Processing
ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.7733306

DIABETES DATA ANALYSIS USING ROBUST SPARSE MINIMUM AVERAGE VARIANCE

Sanaa Jabbar Tuama* and Prof. Dr. Ali J. Kadhim Alkenani
Department of Statistics, College of Administration and Economics, University of Al-Qadisiyah, Al Diwaniyah, Iraq, *E-mail: dmin.stat21.3@qu.edu.iq

Abstract

When the number of variables is large, regression analysis is a difficult process in the sense that increasing the number of variables contributes to increasing the complexity of the model and this may lead to the problem of the dimensionality curse, and this problem prompted researchers to work to reduce these high dimensions of the data, and there are two ways to reduce the dimensions, which is the method of choosing Variables (V.S) and Variables extractions. Under the assumptions of the SDR (Sufficient dimension reduction) theory, the researchers also worked on proposing methods to reduce dimensions, including merging SDR methods with regularization methods, and among these methods (SMAVE-AdEN (Alkenani and Rahman) (2020), which is a method for selecting a variable under assumptions The SDR theory, and the SMAVE-AdEN method is a result of the combination of Adaptive elastic net method and (MAVE) (Minimum average variance estimator) method for estimating minimum average variance.

to estimate the average minimum variance. This method is effective when the variables are highly correlated, but the SMAVE-AdEN method is not robust

and is a sensitive method that is affected when there are outliers in the data, because it uses the least squares criterion.

Here we propose a robust method for variable selection under SDR assumptions called (RSMAVE-AdEN). It is not affected by outliers found in both the explanatory and response variables. The efficiency of the proposed method was verified by studying and analyzing real data of diabetic patients.

Keywords: dimensionality reduction, RSMAVE-AdEN, sparsely robust adaptive elastic network, MAVE method.

1- Introduction

The study of regression when there are a large number of variables and a large sample size is a complex process, as it increases the complexity of the regression model, which prompted researchers to use a process to choose the variable because some explanatory variables are not essential in their effect on the dependent variable, as well as the existence of an internal correlation between the variables with each other. This leads to the emergence of the problem of the curse of linear dimensions, and therefore its effect is not significant, which calls for the exclusion of non-important variables and the selection of important variables to increase the accuracy of the model prediction.

This problem led the researchers to work on reducing the high dimensions of the data, as Cook suggested in (1998) the (Sufficient dimension reduction) method, this method to address the issue of high-dimensional data analysis. One of these methods is the MAVE method (Xia et al., 2002). Several methods have been proposed to combine SDR methods with regularization methods, and these methods are able to deal with high-dimensional data, which are based on the principle of minimizing the sum of squares of error.

By adding a certain penalty threshold to the parameters and decreasing some coefficients and making others equal to zero, it gives a (sparse model) that includes some variables and is interpretable. For example, researchers Alkenani and Rahman (2020) proposed the SMAVE-EN method, in which the method MAVE (minimum mean variance estimator) proposed by researcher Xia et al. (2002) with the EN elastic network proposed by Zou and Hastie (2005). This method is able to handle high data,And the strong correlation between the variables.

Researchers Alkenani and Rahman (2020) proposed the SMAVE-AdEN method, in which the MAVE method (minimum average variance estimator) is combined with AdEN Adaptive Elastic Net) proposed by researchers Zou and Helen (2009) to produce the SMAVE-AdEN method, and this method It has accurate estimates when the variables are highly correlated as the parameters and variable selection are estimated simultaneously. However, it is not accurate when there are outliers in the data.

In this article, we proposed a robust method (RSMAVE-AdEN), which can estimate parameters and define variables simultaneously, and is not affected by the presence of outliers in explanatory variables and response variables. The proposed method is verified by studying and analyzing real data of diabetic patients.

2- Methods for selecting the variable

The regression model may include a large number of explanatory variables and it is not possible to know which of the explanatory variables can influence the dependent variable (Hesterberg et al., 2008). Variable selection methods are divided into two types: Classical methods and Regularization methods).We would like to show that these methods have disadvantages such as instability, high variance and time consumption, i.e. the process of selecting the variable and estimating the parameters is not done simultaneously. Consequently, the resulting model has poor prediction accuracy (Breiman, 1996).The regularization methods have higher stability compared to classical methods, and also V.S variable selection and parameter estimation are performed simultaneously (Alkenani and Yu, 2013).We mention here some remedial procedures: for example [the Lasso method (Tibshirani, 1996],[Elastic Network Method (EN) Zou and Hastie,2005], [Adaptive Lasso Method (ALasso) by the researcher Zou, 2006], (The adaptive elastic net method (AdEN) [Zou and Zhang .2009].And other ways.

3 . Adaptive Elastic Net Method(AdEN)

This method was introduced by the researchers (Zon and Helen) in (2009) the AdEN method, as it deals with interlinear relationships, the adaptive elastic network estimator is defined by the equation shown below:

$$\widehat{\boldsymbol{\theta}}\ \textbf{(AdEN)=arg min}\sum_{i=1}^{n}(y_i - \boldsymbol{\theta}^T \boldsymbol{xi})^2 + \lambda_1 \sum_{k=_1}^{p} \boldsymbol{\theta}_k^2 + \lambda_2 \sum_{k=1}^{p} \boldsymbol{w}_k^* |\boldsymbol{\theta}_k| \quad \textbf{(1)}$$

4. SMAVE-AdEN

The two researchers presented in a year(2020) (Alkenani and Rahman) The (SMAVE-AdEN) method, which is a combination of two methods, MAVE (Xia, 2002) and the Adaptive Elastic Net method (Zou and Zhang, 2009) This method is expressed by the following equation:

$$\sum_{j=1}^{n}\sum_{i=1}^{n}[y_i - \{a_j + b_j^T\theta^T(x_i - x_j)\}]^2 W + \lambda_1\|\theta_m\|_2^2 + w_k^*\lambda_2\|\theta_m\|_1 \quad ,(2)$$

Despite its good advantages, it is affected by outliers, and we proposed a robust SMAVE-AdEN method, by replacing the least squares with Tukey biweight, expressed by the symbol RSMAVE-AdEN, and defined by the following equation:

$$\sum_{j=1}^{n}\sum_{i=1}^{n} p[y_i - \{a_j + b_j^T\theta^T(x_i - x_j) b_j^T\theta^T(x_i - x_j) w_{ij} + \lambda_1\|\theta_m\|_2^2 + w_k^*\lambda_2\|\theta_m\|_1,$$

Since when the derivative of the loss function is characterized as (aredescending derivative), the loss function is immune to outliers in x and y (Rousseeuw and Yohai, 1984). Tukey's criterion has this characteristic (Tukey, 1960).Tukey's criterion is expressed in the following equation:

$$p_c(U) = \begin{cases} \left(\frac{C^2}{6}\right)\{1 - [1 - (\frac{U}{C})^2]^3 \\ \frac{C^2}{6} \qquad\qquad if\,|u| > C \end{cases}$$

5. The applied side

Here, in the applied aspect, we analyze and study the relationship between the factors affecting diabetes in the blood, represented by the dependent variable Y and the independent variables (age, weight, ..etc.) in order to identify the performance of the proposed method RSMAVE-AdEN and the studied methods RSMAVE-EN, SMAVE-AdEN using the data Some patients in Al-Diwaniyah Teaching Hospital. We compared the proposed method RSMAVE-AdEN with other methods, namely the SMAVE-AdEN method (Alkenani and Rahman, 2020) and the RSMAVE-EN method (Alkenani and Aljobori, 2021). To verify the efficiency of the proposed method and its capabilities in selecting variables, we will rely on the standard of mean squared error (MSE) and average number of zero coefficients (Ave0's).

The researcher wrote a code in R language to calculate the results of the proposed method RSMAVE-AdEN.

6. Study sample and description of the study data

Data was collected for a sample consisting of (216) people with diabetes. The data collection mechanism was based on a form and direct interviews with patients in Al-Diwaniyah Teaching Hospital / Diabetes and Eye Center. A form was organized to collect data, and then the data of this form were tabulated, and Table No. (1) Explains the names of the explanatory variables and the dependent variable.

The dependent variable (Y) crosses to blood sugar, and the explanatory variables are age (x1), weight (x2), height (x3), monthly income (x5), sex x6, number of family members (x7), cholesterol (x8), suffers from a problem psychological (X9), marital status (X10), exposure to trauma (X11), earning or employee (X12), nature of food (vegetarian, animal) X13, do you suffer from hyperactivity disorder (x14), blood type (X15), presence Other diseases (X16), heredity (diabetes in the family) (X17), smoking (x18), school level (X19), urea (X20), creatine (X21).

7.Test for the presence of outliers on real data

To perform this test, we use a criterion (μ±3σ) to determine outliers for each variable in real data (Lehmann ,2013).
We include here a drawing of the shape of the variables that contain outliers, as follows:

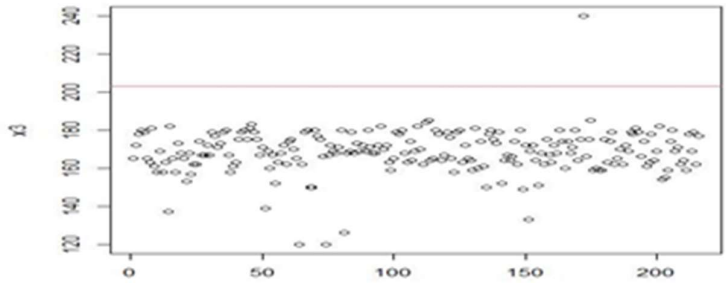

Figure (1) shows the outliers in the variable X3

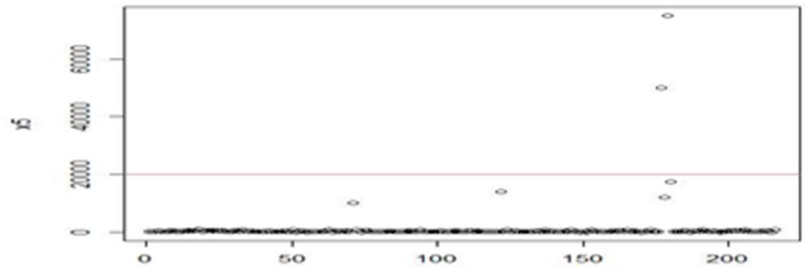

Figure (2) shows the outliers in the variable X5

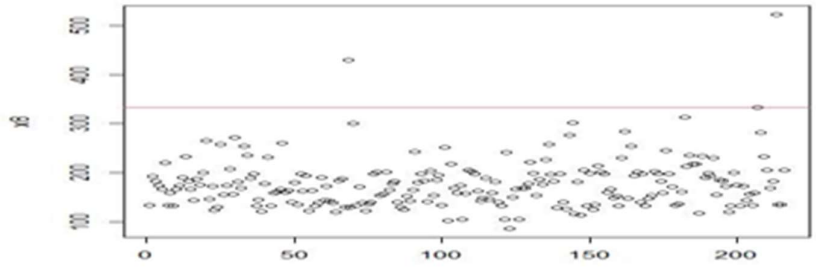

Figure 3 shows the outliers in the variable X8

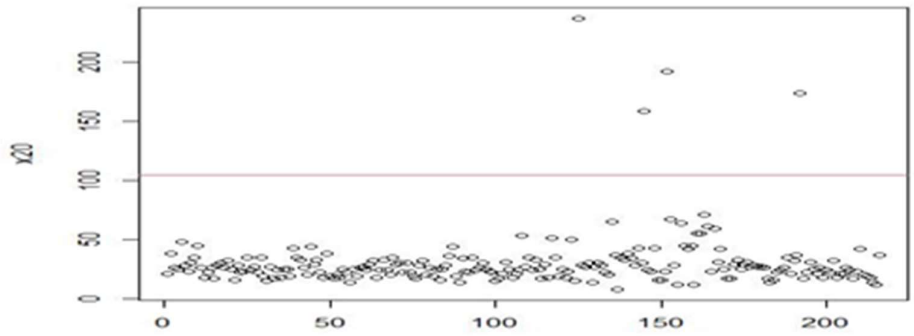

Figure 4 shows the outliers in the variable X20

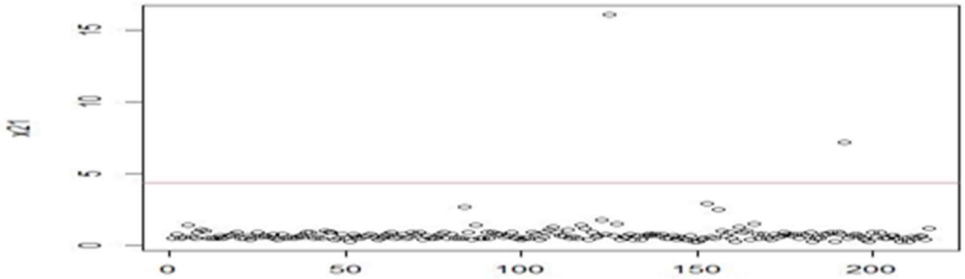

Figure (5) shows the outliers in the variable X21

The figures above showed that some of the variables had abnormal values such as x3, x5, x8, x20, and x21, so we do not need to pollute the data. Likewise, the variable y includes abnormal values, as shown in the following figure:

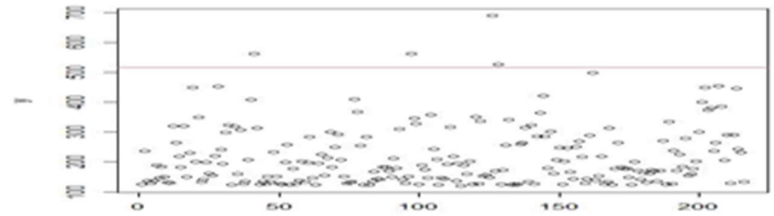

Figure (6) shows the outliers in the variable y

8. Real data results

By analyzing real data for diabetic patients, the results shown in Table (1) showed the value of the parameters of the regression model (Beat).

Variants	SMAVE-AdEN	RSMAVE-EN	RSMAVE-AdEN
X_1	0.125907	0.120561	0.026029
X_2	-0.298978	0.091451	-0.016476
X_3	0.107531	0.00000	0.00000
X_5	0.00000	0.207853	-0.977794
X_6	0.056054	-0.118131	0.000000
X_7	0.181579	0.2541433	0.00000
X_8	0.181579	0.2644151	-0.066873

X_9	0.203986	-0.136066	0.000000
X_{10}	-0.170008	0.176567	-0.026008
X_{11}	-0.143819	0.359058	-0.036109
X_{12}	0.080801	-0.439647	0.013790
X_{13}	0.306188	-0.269757	0.013790
X_{14}	-0.043024	0.324821	-0.174384
X_{15}	-0.308703	0.134192	-0.028113
X_{16}	0.024399	0.041981	-0.013252
X_{17}	-0.071182	0.205518	0.025234
X_{18}	0.347748	-0.280087	0.032861
X_{19}	0.243710	-0.025385	0.033309
X_{20}	0.603145	-0.208113	0.035343
X_{21}	0.00000	-0.213411	-0.016350

Table (1) shows that the number of non-significant variables is (4) for the proposed method RSMAVE- AdEN, (1) for the RSMAVE-EN method, and (2) for the method. SMAVE-ADEN.

Table 2 shows the number of zeros and MSE for real data analysis of diabetes

method	number of zeros	MSE
RSMAVE-AdEN	4	40.92771
RSMAVE-EN	1	45.30264
SMAVE-AdEN	2	58.02755

Table (2) shows the number of zeros and the mean squared error MSE for the analysis of real data for diabetes and for the three methods. We note that the proposed method RSMAVE-AdEN has better performance and is superior to other methods RSMAVE-EN and SMAVE-AdEN. The proposed method gave the lowest level of MSE and this is a good behavior.
Figure (7) shows the MSE value of the methods RSMAVE-AdEN, RSMAVE-EN, SMAVE-AdEN

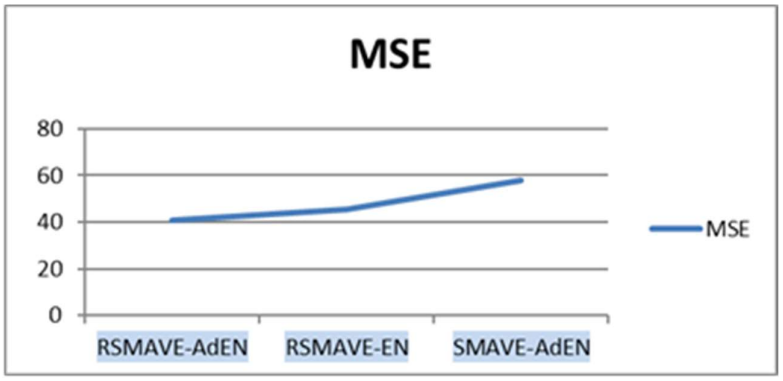

Figure (7)

Figure (7) shows the number of zero coefficients for the methods RSMAVE-AdEN, RSMAVE-EN, SMAVE-AdEN.

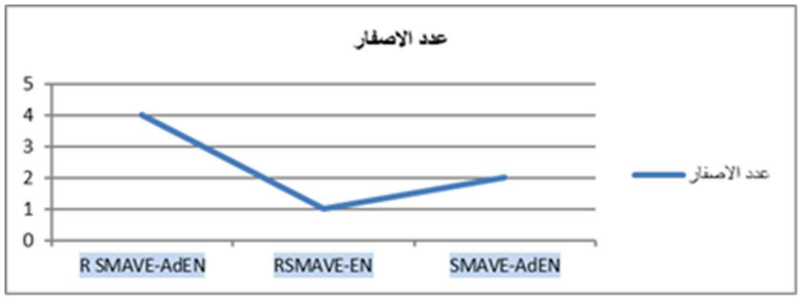

Figure (8)

9. Conclusions

1- The proposed method RSMAVE-AdEN in this research is a robust method for selecting variables and reducing dimensions at one time.

2- This method is efficient when the variables are highly correlated within the SDR settings.

3- The results showed that the proposed RSMAVE-AdEN method has good performance in variable selection and estimation accuracy even with the presence of outliers in the variable x and the response variable y.

References

- Alkenani, A. (2020). Robust variable selection in sliced inverse regression using Tukey biweight criterion and ball covariance. Journal of Physics Conference Series, 1664, 012034.
- Alkenani, A. and Aljobori, N. (2021). Robust sparse MAVE through elastic net penalty. International journal of Agricultural and Statistical Sciences, Vol.17, Supplement 1, 2039 - 2046.
- Alkenani, A. and Rahman, E. (2020). Sparse minimum average variance estimation via the adaptive elastic net when the predictors correlated, Journal of Physics Conference Series, 1591, 012041.
- Alkenani, A. and Rahman, E. (2021). Regularized MAVE through the elastic net with correlated predictors, Journal of Physics Conference Series, 1897, 012018.
- Alkenani, A. and Yu, K. (2013). Sparse MAVE with oracle penalties. Advances and Applications in Statistics 34, 85–105. Bellman, R. E. (1961). Adaptive Control Processes. Princeton University Press, Princeton, New Jersey.
- Breiman, L. (1996). Heuristics of instability and stabilization in model selection. The Annals of Statistics, 24(6), 2350–2383.
- Cook, R. (1998). Regression graphics: ideas for studying the regression through graphics. New York, Wily.
- Hesterberg, T., Choi, N. H., Meier, L., & Fraley, C. (2008). Least angle and ℓ1 penalized regression: A review. Statistics Surveys, 2, 61-93.
- Rousseeuw, P. and Yohai, V. (1984). Robust regression by means of s-estimators. In Robust and Nonlinear Time Series Analysis, pages 256-272.
- Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society, Series B, 58, 267–288.
- Tukey, J. W. (1960). A survey of sampling from contaminated distributions. Contributions to Probability and statistics, 2:448-485.

- Xia, Y. et al. (2002). An adaptive estimation of dimension reduction space. Journal of the Royal Statistical Society Series B 64, 363–410.
- Zou, H. (2006). The adaptive lasso and its oracle properties. Journal of the American Statistical Association, 101(476), 1418–1429.
- Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net, Journal of the Royal Statistical Society, Series B 67, 301–320.
- Zou, H., and Zhang, H. (2009). On the adaptive elastic-net with a diverging number of parameters.Annals of Statistics, 37(4), 1733