The Use MCP method in determining the most important factors affecting the incidence of viral hepatitis

Asaad Naser Hussein Mzedawee
asaad.nasir@qu.edu.iq
Al-Qadisiyah University, College of Administration and Economics

Corresponding Author : Asaad Naser Hussein Mzedawee

Abstract : One of the most prevalent illnesses in today's world is viral hepatitis, and there are several mathematical methods to determine the most important factors that lead to the infection with this disease. In this paper, the MCP method is used to select the most important factors (variables) that lead to this injury, where simulation is used to compare this method with some penalty methods (lasso, adaptive lasso and MCP). Using the MSE (mean square error) and was applied to the real data.

Keywords: Viral hepatitis, MCP, Variable selection

Introduction: Methods of selecting and diagnosing important variables are considered important processes in the study of medical, economic and social problems in order to determine which factors that have an impact on or relationship to the studied problem.

Among the classic methods that were used earlier are stepwise (1974), forward (1978), and backward, as well as some calibrations, including AIC, BIC .These methods were suffering from time consuming as well as accuracy issues, especially when studying many variables and when the of variables is greater than the sample size ($P > N$), so modern methods appeared to identify, select, and estimate important variables, including the Lasso method, and then the adaptive lasso method appeared, which relied on adding weight to the parnumberameter estimator . In this paper, three methods of selecting variables in the models were employed (lasso, adaptive lasso, MCP), where simulation methods were used to compare the methods by generating data close to the data of the results of viral hepatitis diseases. Where the standard (MMAD, SD) was used to compare these methods. And through the simulation results, that showed the advantage of the MCP method. The data of this paper were analyzed by the collected real data. The paper was divided as follows: the second paragraph included a review of the methods of selecting variables, while the third paragraph studied the simulation of the methods used. The fourth paragraph included the study of the real facts and finally the fifth paragraph included conclusions and recommendations.

Variable Selection

studies in for medical and social science comprise a sizable number of factors that are thought to be pertinent to the research. However, these variables' real effects are frequently sparse and dispersed. The concepts of impact Sparse guide regression analysis looks at a large number of possible variables. Variable selection strategies are commonly employed to identify significant regression models involving variables. The process of lowering the number of random variables while preserving as much information as possible is known as variable selection. It is one of the primary ways to lower the time, cost, and voltage dimensions.

Recent developments in data collection have led to the massive collection of multivariate data at a quick pace. The "Curse of Dimensionality" problem makes most statistical approaches difficult to apply to such big data sets. An linked mathematical space's exponential volume expansion that results from the addition of new dimensions is the cause of the "curse of dimensionality." The common statistical techniques for high dimensional data fail because of this issue. The variable selection it is one of the main solutions for the curse of dimensionality. Subset selection has grown in popularity as a study topic across numerous disciplines. Selecting a subset of significant variables to be used in the construction of a model is what it is. Then, implicit coefficients with a value of zero are assigned to any unimportant variables that have not been selected. For the purpose of doing variable selection, numerous methodologies have been devised. Several methods have been used four identify variables and important factors in linear models, Of prevailing methods to variables selection is the stepwise regression method. stepwise analysis is been initially proposed by Efroymson (1960), and is fully explained by Jennrich (1977a and 1977b), Draper and Smith (1981), and others. Through which the variables are determined sequentially from the regression models in order to determine the best fit model of all subsets models. It is possible to add or remove variables at different steps while using stepwise selection. It alternates between choosing which variable to add to the model while taking into account the variables that were previously removed, and choosing which variable to remove from the model. special methods of stepwise selection to find the best model, of which the forward, and backward elimination. The best model is chosen using a specific criterion for the selection processes discussed above. The Bayesian information criterion (BIC) and Akaike's information criterion (AI C) are common options. The Akaike information criterion (AI C) was put forth in 1974 by Akaike.(Akaike, 1974), under the name of "an information criterion", one of the most commonly used information criteria. AIC is a tool for evaluating an estimated statistical model's goodness of fit that is based on information theory. The AI C defined by

$$AIC = 2K - 2\ln(L) \qquad (1)$$

Here, we L have the likelihood function value for the estimated model and K the number of parameters in the statistical model. By adding a penalty term for the number of parameters in the model, the BIC also overcomes the over fitting issue. The formula for BIC defined by

$$BIC = -2\ln(L) + 2\ln(n)K \qquad (2)$$

Recent years have shown recent methods of selecting variables. These methods provide a tool through which we can develop the ability to interpret the model and the accuracy of the prediction by choosing the automatic variable.

2-1. Lasso

Lasso is linear model estimation method proposed by Tibshirani (1996). The lasso estimate of β

$$\hat{\beta}_{\text{lasso}} = \arg\min_{\underline{\beta}} \{\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p}\beta_j x_{ij})^2 \} + \lambda\sum_{j=1}^{p}\left|\beta_j\right| \qquad (3)$$

Where $\sum_{j=1}^{p}\lambda\left|\beta_j\right|$ is called lasso penalty, and t , λ are user-defined tuning parameters that regulate the amount of shrinkage. Higher values of lead to a greater amount of shrinkage, whilst smaller values of t do the opposite.

2.2 Adaptive Lasso

A novel version of the Lasso (Tibshirani, 1996) was proposed by Zou in 2006. It was based on adaptive weights, which led to distinct penalizations for different penalty coefficients. The definition of the adaptive lasso is

$$\arg\min_{\underline{\beta}}\{\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda\sum_{j=1}^{p}w_j\left|\beta_j\right| \qquad (4)$$

Where $(w_1, w_2, ..., w_p)$ are the adaptive weights.

2.3 MCP Method

Another option for obtaining less biased regression coefficients in sparse models is the Minimax Concave Penalty (MCP). In order to solve the Lasso technique's inconsistent variable selection, Zhang (2010) suggested the MCP method, which uses the MCP penalty function to simultaneously estimate and pick linear regression variables. The following formula yields the MCP estimator:)Choon, C. L. (2012)

$$\hat{\beta}_j^{MCP} = \arg\min_{p}[\left\|y^* - X\underline{\beta}\right\|]^2 + \sum_{j=1}^{p}P_{\lambda,\gamma}^{MCP} \qquad (5)$$

Where: $\sum_{j=1}^{p}P_{\lambda,\gamma}^{MCP}$ the MCP penalty function

3. Simulation Study

This section, We execute MCMC simulation examples to illustrate The effectiveness of the suggested methods (LASSO , ALASSO & MCP).

1- Simulation study 1 (simple sparse case): β = (2,2,0,0,1,1,0,0) .
2- Simulation study 2 (sparse case): β =(1,1,0,0,0,1,1,0,0,0,1,1,0,0,0).
3- Simulation study 3 (dense case): $\beta = \left(\underbrace{0.95, 0.95, \ldots, 0.95}_{20}\right)$.

4- Simulation study 4 (very sparse case): $\beta = \left(\underbrace{1,1,0,\ldots,0}_{15}\right)$.

The data in the simulation examples are generated by

$$y_i^* = \chi_i^{'}\beta + \varepsilon_i \quad , \varepsilon_i \sim \mathrm{N}(0,\sigma^2), i = 1,2,....,n \qquad (6)$$

In each simulation study, we run 1000 replications. For each replication, we simulate 50 observations as a training set and 350 observations as a testing. Set the first 200 iteration as a burn-in. Approaches are compared using median of mean absolute deviation (MMAD) which can be calculated as

$$MMAD = medain(\sum_{i=1}^{200}\left|x_i^{'}\hat{\beta} - x_i^{'}\beta^{true}\right|$$

Table(1) : MMADs and SD for Simulation study 1.

Methods	σ^2	MMAD	SD
LASSO	3	22.88	7.135
ALASSO	3	1.722	0.51
MCP	3	**0.972**	0.409
LASSO	6	3.897	6.773
ALASSO	6	1.783	0.484
MCP	6	**1.282**	0.294
LASSO	9	2.891	3.521
ALASSO	9	1.827	0.46
MCP	9	**1.549**	0.286

We note from Table (1) that method (MCP) is better than the two methods (LASSO, ALASSO) because it recorded the lowest value of MMAD .

Table(2) : MMADs and SD for Simulation study 2.

Methods	σ^2	MMAD	SD
LASSO	3	32.88	3.634
ALASSO	3	0.952	1.842
MCP	3	**0.089**	1.704
LASSO	6	29.45	5.034
ALASSO	6	0.995	1.82
MCP	6	**0.302**	1.662
LASSO	9	28.15	9.64
ALASSO	9	1.067	1.793
MCP	9	**0.647**	1.564

We note from Table (2) that method (MCP) is better than the two methods (LASSO, ALASSO) because it recorded the lowest value of MMAD.

Table (3) : MMADs and SD for Simulation study 3.

Methods	σ^2	MMAD	SD
LASSO	3	31.4	2.154
ALASSO	3	0.528	3.322
MCP	3	**0.391**	3.184
LASSO	6	27.97	3.554
ALASSO	6	0.485	3.3
MCP	6	**1.178**	3.142
LASSO	9	26.67	8.16
ALASSO	9	0.413	3.273
MCP	9	**0.133**	3.044

We note from Table (3) that method (MCP) is better than the two methods (LASSO, ALASSO) because it recorded the lowest value of MMAD.

Table (4) : MMADs and SD for Simulation study 4.

Methods	σ^2	MMAD	SD
LASSO	3	37.61	4.629
ALASSO	3	1.937	3.315
MCP	3	**1.858**	3.246
LASSO	6	30.912	8.132
ALASSO	6	1.939	3.288
MCP	6	**1.477**	3.152
LASSO	9	24.795	9.512
ALASSO	9	1.786	3.281
MCP	9	**1.297**	3.094

We note from Table (4) that method (MCP) is better than the two methods (LASSO, ALASSO) because it recorded the lowest value of MMAD .

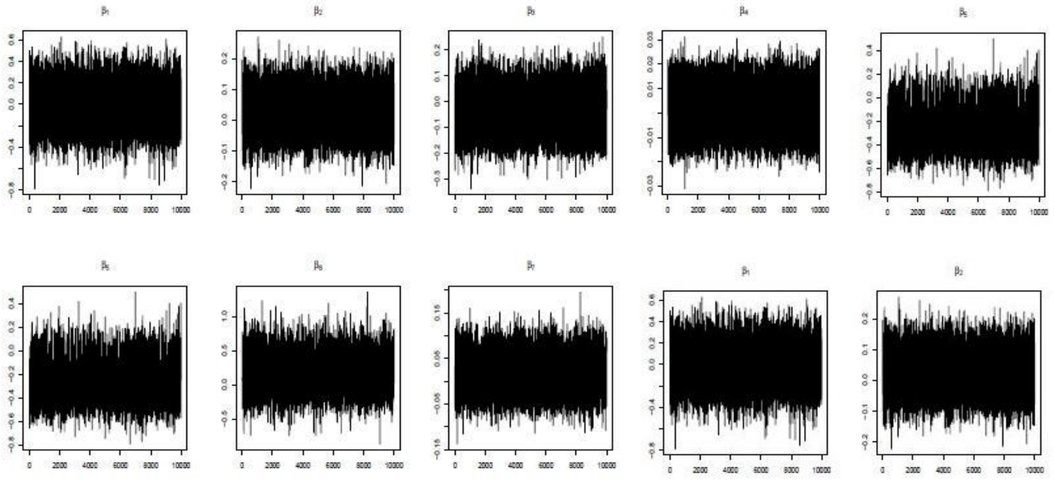
4. Real Data

Here, we describe the data pertaining to Hepatitis disease by outlining the most significant influencing factors and estimating their impact on the response variable (patient's death or recovery). Additionally, we aim to identify the optimal model using the MCP method, based on criteria such as MSE, MMAD, and RMSE. Real data regarding the research subject were collected from hospitals in Baghdad Governorate for the year 2021. A panel of doctors specializing in respiratory and blood diseases was consulted to determine the key factors affecting the disease. Data for the year 2021 were collected from a sample size of (250) patients via the dedicated "Sleeping Patient" file for each patient, simplifying the analysis process.

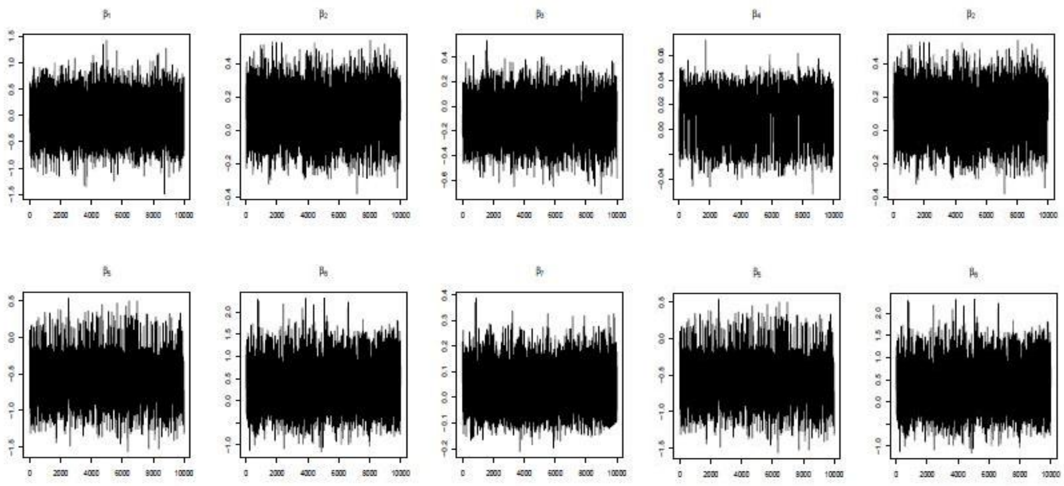
The Age is represented by variable X1.
The gender is represented by the variable X2.
The variable X_3 represents the General health condition .
The variable X_4 represents the Medical history.
The variable X_5 represents the Healthy behaviors.
The variable X_6 represents the Vaccination.
The variable X_7 represents the Environmental factors.
The variable X_8 represents the Exposure to harmful substances.
The variable X_9 represents the Travel.
The variable X_{10} represents the Economic status.

Table(5) : The parameter estimations for the data.

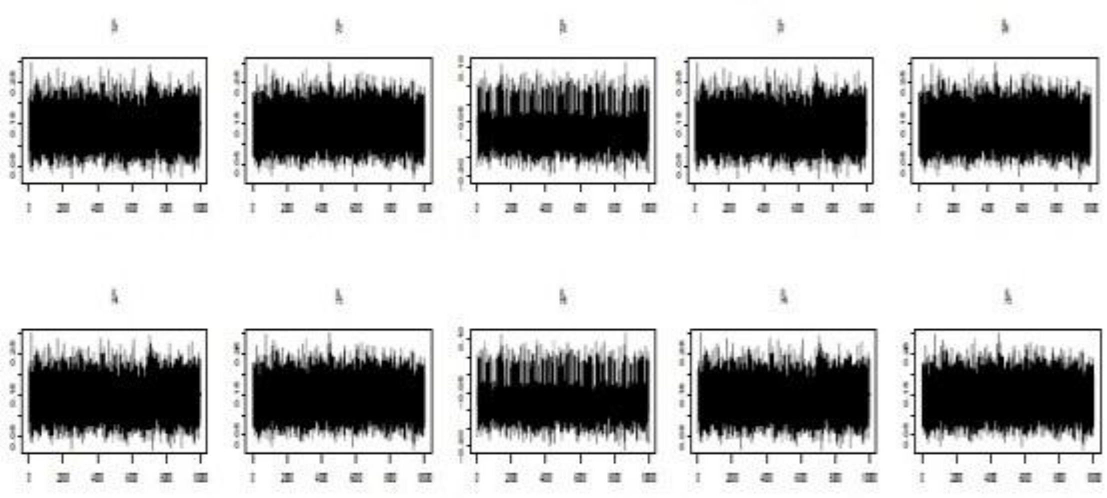
Coefficients	LASSO	ALASSO	MCP
$\hat{\beta}_1$	-2.5	2.2	0
$\hat{\beta}_2$	0	0	0
$\hat{\beta}_3$	-6.32	2.55	6.25
$\hat{\beta}_4$	-4	0	0.06
$\hat{\beta}_5$	0	0	0
$\hat{\beta}_6$	7.3	0.2	0.21
$\hat{\beta}_7$	-1.3	0	-0.06
$\hat{\beta}_8$	11.2	2.8	0
$\hat{\beta}_9$	0	0	-0.32
$\hat{\beta}_{10}$	2.2	3.8	0



Figure(1) : Trace plots based for real data using MCP method.



Figure(2) : Trace plots based for real data using ALASSO method.



Figure(3) : Trace plots based for real data using LASSO method.

5- Conclusions

Through this study of methods for selecting classical and modern variables, we conclude the following

1- Modern methods (Lasso, Adaptive lasso, MCP) have shown good performance in simulation experiments. Where, These methods provide a device through which we can get the capacity for interpret the models and the accuracy of the prediction by choosing the variables

2-By comparing these methods, a method(MCP) has shown better performance than the methods mentioned in the study, especially in the application of real data .

3. the study variables of the real data show that a variable (General health condition, Medical history. And Vaccination) had a direct effect on viral hepatitis disease .

4- We recommend using variable selection methods in analyzing medical data, especially chronic diseases.

6- Reference

[1] Akaike, H. (1974). A New Look at the Statistical Model Identification. IEEE Transactions on Automatic Control Vol. 19, No. 6, 716723.

[2] Choon, C. L. (2012). Minimax concave bridge penalty function for variable selection.

[3] Draper, N. and Smith, H. (1981) Applied Regression Analysis, 2d Edition, New York: John Wiley & Sons, Inc

[4] Efroymson, M. A. (1960). Multiple regression analysis. Mathematical methods for digital computers, 1, (pp. 191-203).New York: Wiley.

[5] Jennrich, R. I. (1977a). Stepwise regression. Statistical methods for digital computers(Vol. 3, pp. 58-75). New York: Wiley

[6] Jennrich, R. I. (1977b). Stepwise regression. Statistical methods for digital computers(Vol. 3, pp. 76-96). New York: Wiley.

[7] Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society, Series B 58, 267– 288

[8] Schwarz, G. (1978). Estimating the Dimension of a Model. The Annals of Statistics Vol. 6, No. 2, 461-464

[9] Zou, H. (2006). The adaptive lasso and its oracle properties. Journal of the American Statistical Association 101, 1418– 1429 .

[10] 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

[11] Zhang, C. H. (2010). Nearly unbiased variable selection under minimax concave penalty. Annals of Statistics 38, 894– 942.