Variable selection methods applied to the mathematics scores of Indonesian students based on convex penalized likelihood

V M Santi\textsuperscript{1,}\textsuperscript{*}, K A Notodiputro\textsuperscript{2} and B Sartono\textsuperscript{2}

\textsuperscript{1}Statistics Study Program, Faculty of Mathematics and Natural Sciences, State University of Jakarta, Indonesia
\textsuperscript{2}Department of Statistics, Faculty of Mathematics and Natural Sciences, Bogor Agricultural University, Bogor, Indonesia

* vmsanti@unj.ac.id

Abstract. Variable selection is an important topic in linear regression analysis. In practice, a large number of predictors usually are introduced at the initial stage of modeling to attenuate possible modeling biases. Stepwise deletion and subset selection are usually used which can be computationally expensive and ignore stochastic errors in the variable selection process. In addition, the best subset selection of variables suffers from several disadvantages, the most severe of which is its lack of stability. In this article, penalized likelihood approaches are proposed to handle these kinds of problems. The proposed methods select variables and estimate coefficients simultaneously. Some of penalty functions are used to produce sparse solutions. Based on the RMSE and Generalized Information Criterion (GIC) criteria, it was found that the factors affecting Indonesian mathematics scores, where LASSO produces 11 important variables for the model while SCAD has 6 variables which mean that the LASSO model is more complex than SCAD. The MCP produces a simpler model with 5 important variables but has excessive biased. The results also showed that the SCAD penalty function had the best performance compared to LASSO, Ridge and MCP. Ridge penalty has a worst performance based on all criteria.

1. Introduction

The selection of variables is one of the important things in linear regression analysis. In general, statisticians usually use the stepwise and subset selection methods to increase predictability and to select significant variables. Although both of these methods are practically used, this selection procedure ignores the stochastic error inherited at the variable selection stage. Therefore, the theoretical nature of this method becomes difficult to understand. In addition, the best subset selection method has several disadvantages, one of which is the lack of model stability [1]. In an effort to automatically and simultaneously select variables, the approach is usually used via penalized least squares with the penalty function. The penalty function must be singular at origin to produce a sparse solution, to meet certain conditions to produce a continuous model (for stability from model selection), and be limited by a constant to produce estimates that are almost unbiased for large coefficients.

Many methods have been proposed and their properties have been studied, for example, LASSO [2] and the Ridge Regression proposed in [3] are a member of the least squares identifiable. Although both of these penalty functions are related to Lq but did not meet all the requirements of the required
properties above [4,5], SCAD [6], Elastic-Net [7], adaptive LASSO [8], COSSO [9], SICA [10], MCP [5], truncated L1 [11] and SELO [12].

Unlike the traditional variable selection procedure, the nature of sampling in the penalized likelihood can be determined. Predictors of possible negligence work equally well if the sub model is correctly known before. This means that the penalty function of Smoothly Clipped Absolute Deviation (SCAD) [6] has a better performance and is able to suppress excessive bias and outperform the maximum probability estimator. This is very analogous to the phenomenon of super efficiency in the example of Hodges [13]. The MCP provides the convexity of the penalized loss in sparse regions to the greatest extent given certain thresholds for variable selection and unbiasedness [5].

Possible penalized methods also can be applied easily to high-dimensional nonparametric modeling. In a series of works by Stone and its collaborators [14], traditional variable selection approaches are modified to select useful spline subbases. The approach to penalized likelihood in [15], and [16] and the references therein, is based on quadratic penalties. Comparison of variable selection methods and estimation of regression parameters via penalized likelihood with several penalty functions applied in analyzing factors that influence the average value of mathematics students in Indonesia based on the 2015 PISA data.

2. Material and methods

2.1. Material
Based on PISA data in 2015, there were 6,513 students in Indonesia measured using questionnaire instruments. The sample size that observed was 232 schools in Indonesia. The number of independent variables used is 200 based on the number of items in the PISA questionnaire. The observation of the response variable in this study is the average student math score aggregates as the average mathematical value of students in the school.

2.2. Methods
In this study, several penalty functions were used via convex penalized likelihood namely ridge regression, LASSO, MCP and SCAD which were used to select independent variables while estimating the regression model parameters regarding to the factors that influence the average mathematics scores of students in Indonesia based on PISA year data 2015. In the SCAD penalty function the regularization parameter value $a = 3.7$ is based on [6]. The criteria for Cross Validation (CV) and Generalized Cross Validation (GCV) are used to determine the optimum lambda value. To see the goodness of the model we use the criteria for Root Mean Square Error (RMSE) and Generalized Information Criterion (GIC).

3. Result and discussion
In the initial stage of selecting variables while determining the model parameter coefficients, try with ridge regression first. Based on the ridge regression results, the BIC value of 1108.910 and RMSE is 45.89411 with all the independent variables entering the model. The resulting ridge regression model has excessive bias and high complexity because ridge regression performs a penalized for all parameter coefficients but does not reach zero or does not produce a sparse solution (see Figure 2d).

| Approachment | Lambda (λ) Optimum |
|--------------|--------------------|
| SCAD         | 0.030065426        |
| LASSO        | 0.03006543         |
| MCP          | 0.03006543         |
Based on Table 1, it can be seen that the optimum lambda value for the three penalty functions is almost the same. After obtaining the optimum lambda value, then it is done simultaneously penalties for free variables will result in a sparse solution.

Based on the optimum lambda value in Fig. 1, the SCAD penalty function produces a sparse solution with 6 independent variables as important variables for the model as in Fig. 1a. LASSO produces more independent variables to be included in the regression model, while the MCP mostly presses the parameter to zero and leaves only 5 independent variables as important variables for the model as in Fig. 1c.

Furthermore, the estimated value of the parameter coefficients of the regression model is generated simultaneously with the penalized process of the parameter coefficients whose values are close to zero. This free variable which corresponds to the parameter coefficient is referred to as an important variable for the regression model, as presented in Table 2 below.

**Figure 1.** Variable penalty process (a) SCAD, (b) LASSO, (c) MCP, (d) Ridge.

**Table 2.** Parameters coefficient of variable selection results.

| Parameters Coefficient (SCAD) | Parameters Coefficient (LASSO) | Parameters Coefficient (MCP) |
|-------------------------------|-------------------------------|-------------------------------|
| 1.7107618 (V100)             | 0.93441717 (V100)             | 1.5850401 (V100)             |
| -1.8273808 (V115)            | -1.14079728 (V115)            | -1.7548294 (V115)            |
| 0.3310645 (V185)             | 0.06528688 (V185)             | 0.3965482 (V185)             |
| 0.6306875 (V192)             | 0.49756907 (V192)             | 0.8252179 (V192)             |
| 1.1780653 (V196)             | 0.57587991 (V196)             | 1.4130881 (V196)             |
| 0.9468941 (V166)             | -0.03062178 (V3)              | 0.2612070 (V21)              |
Based on Table 2, SCAD produces 6 independent variables that are important for the model, while MCP produces 5 important variables for the model. Only the 166th variable is not in the MCP but is in SCAD. LASSO produces 11 independent variables that are important for the model by adding the 3rd, 21st, 98th, 111th and 124th free variables which neither SCAD nor MCP choose these variables as important variables to be included in the model. The following are the independent variables that correspond to the parameter coefficients resulting from the selection of variables which are factors that influence the average value of school mathematics in Indonesia as in Table 3.

Table 3. The factors that influence the average value of school mathematics in Indonesia result from the selection of variables.

| Factors (SCAD)                                      | Factors (LASSO)                                      | Factors (MCP)                                      |
|-----------------------------------------------------|------------------------------------------------------|----------------------------------------------------|
| Environmental insight into the greenhouse effect (V100) | Environmental insight into the greenhouse effect (V100) | Environmental insight into the greenhouse effect (V100) |
| Environmental insight on the issue of air pollution (-V115) | Environmental insight on the issue of air pollution (-V115) | Environmental insight on the issue of air pollution (-V115) |
| Attended chemistry class in the previous year (V185) | Attended chemistry class in the previous year (V185) | Attended chemistry class in the previous year (V185) |
| Attending study guidance of science (V192)          | Attending study guidance of science (V192)           | Attending study guidance of science (V192)          |
| Listening the teacher's explanation in science subjects (V196) | Listening the teacher's explanation in science subjects (V196) | Listening the teacher's explanation in science subjects (V196) |
| The teacher does not provide feedback for reinforcement (V166) | The teacher does not provide feedback for reinforcement (V166) | Mothers who do not have S2 qualifications (-V3) |
| Students conduct experiments (V21)                  | Environmental insight about increasing greenhouse gases (-V98) | Environmental insight into the problem of clearing forests for land use (-V111) |
|                                                   | Environmental insight into the problem of lack of clean water (V124) |                                                   |

Description: the sign (-) on the variable symbol shows a negative sign on the parameter coefficient.

In Table 3, it can be explained that SCAD has 6 independent variables that enter the model. It is seen that the increasing proportion or percentage of students who know about the issue of greenhouse effects and can explain well about the issue also increases the average value of mathematics of the student's school. On the other hand, with the increasing proportion of students who are aware of the issue of air pollution but unable to explain it properly it actually decreases the average value of school mathematics.
For the percentage of students who take chemistry tutoring in the previous year and science learning guidance in an integrated and intensive manner that year will increase the average value of school mathematics. The increasing percentage of students who did not listen to the teacher's explanations in science learning for several meetings turned out to be directly proportional to the average value of school mathematics. This might happen in the sense that there is a possibility that students do not listen to the teacher's explanations in several meetings of science subjects but listen to mathematics. For MCP, it is almost the same as SCAD except the 166th variable, which is the proportion of teachers who almost never provide feedback for reinforcement which is not an important variable for the MCP model.

LASSO added that the qualifications of mothers who are not masters qualified have a declining impact on the average value of school mathematics, the increasing proportion of students conducting experiments turns out to be directly proportional to the response, and insight into students' knowledge of greenhouse gases, forest clearing and lack of clean water as important variables entered into the model.

3.1. Goodness of fit

From Table 4, it can be explained that based on the RMSE criteria, SCAD has the best performance compared to LASSO, MCP and Ridge. LASSO has a BIC and RMSE value that is greater than SCAD and also has more free variables to be included in the model so that the model becomes more complex. MCP performance is almost the same as SCAD based on RMSE and BIC criteria. When viewed based on the complexity of the model even the MCP produces a simpler model than SCAD. Ridge has the worst performance compared to the other three penalty functions for all criteria. The results obtained are in accordance with the results of the study [6] where SCAD has the best performance that produces a simple and almost unbiased model.

| Penalty Function | BIC      | RMSE    |
|------------------|----------|---------|
| SCAD             | 117.9556 | 30.54303|
| LASSO            | 157.7216 | 31.75476|
| MCP              | 117.3784 | 31.79519|
| Ridge            | 1108.910 | 45.89411|

4. Conclusion

Factors that influence the average value of school mathematics have been assessed based on several penalty functions via convex penalized likelihood. SCAD is able to produce sparse solutions with a simple and almost unbiased model. Ridge regression produces the worst performance based on RMSE and GIC criteria. LASSO produces more complex models with excessive bias, while MCP is almost as good as SCAD and even produces simpler models.

Acknowledgement

Thank you to Prof. Dr. Khairil Anwar Notodiputro and Dr. Bagus Sartono for his guidance, direction and input on this research. Thanks also to the Statistics Study Program Faculty of Mathematics and Natural Sciences State University of Jakarta for the support.

References

[1] L Breiman 1996 Heuristics of instability and stabilization in model selection The Annals of Statistics 24 2350–2383
[2] R Tibshirani 1996 Regression Shrinkage and Selection via The LASSO J R Statist. Soc (B) 58 267-288
[3] I Frank and J Friedman 1993 A statistic review of some chemometrics regression tools with
discussion Technometrics 35 109-148

[4] C H Zhang and J Huang 2008 The sparsity and bias of the lasso selection in high-dimensional linear regression The Annals of Statistics 36 1567–1594

[5] C H Zhang 2010 Nearly unbiased variable selection under minimax concave penalty The Annals of Statistics 38 894-942

[6] J Fan and R Li 2001 Variable selection via nonconcave penalized likelihood and its oracle properties The American Statistical Association 96 1348-1360

[7] H Zou and T Hastie 2005 Regularization and variable selection via the elastic net The Royal Statistical Society: Series B (Statistical Methodology) 67 301–320

[8] H Zou 2006 The adaptive lasso and its oracle properties The American Statistical Association 101 418–1429

[9] Y Lin and H H Zhang 2006 Component selection and smoothing in smoothing spline analysis of variance models The Annals of Statistics 34 2272–2297

[10] J Lv and Y Fan 2009 A unified approach to model selection and sparse recovery using regularized least squares The Annals of Statistics 37 3498–3528

[11] X Shen, W Pan and Y Zhu 2011 Likelihood-based selection and sharp parameter estimation The American Statistical Association To appear

[12] L Dicker, B Huang and X Lin 2011 Variable selection and estimation with the seamless-l0 penalty Statistica Sinica To appear

[13] E L Lehmann 1983 Theory of point estimation Pacific Grove CA: Wadsworth and Brooks/Cole

[14] C J Stone, M Hansen, C Kooperberg and Y K Truong 1997 Polynomial Splines and Their Tensor Products in Extended Linear Modeling (with discussion) The Annals of Statistics 25 1371–1470

[15] G Wahba 1990 Spline Models for Observational Data, Philadelphia: SIAM

[16] P J Green, and B W Silverman 1994 Nonparametric regression and generalized linear models: a roughness penalty approach, London: Chapman and Hall