Simulation method of model selection based on Mallows’ Cp Criteria in linear regression

O Darnius* and G Tarigan

Department of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Sumatera Utara

*E-mail: open@usu.ac.id

Abstract. The selection of models involving many independent variables is one of the studies on linear regression modeling. If there are k independent variables, then there are as many as \(2^k\) the number of models to be observed and selected as good as the final model. The increasing number of all models to be studied with the increase of independent variables is the fundamental issue that needs to be determined in selection criteria. This paper reviews the Mallow’s Cp Criteria that is often overestimate the number of independent variables selected in the model. The number of selected variables in the model using Mallow’s Cp criteria is performed by simulation. The simulation is run in two cases, one for the case where there is a correlation between some independent variables and one dependent variable, whereas the second case is the absence of correlation between some independent variables and one dependent variable. The simulation results are used to investigate the overestimate number of independent variables selected in the model.

1. Introduction

One of the main purposes of statistical analysis is to find an appropriate model for a set of data. Selecting variables that "best" explains the data is called variable selection. Variable selection, though well studied in the statistical literature, remains rich in unsolved problems [1], [2] especially as related to missing response data [3]. Researchers are usually interested in predicting the value of a variable Y based on one or more observed variable X which are considered to be related to the variable Y. The variable Y is referred to as a response or dependent variable, and the variables X are referred to as predictors or independent variables. In a more common situation the set of predictor variables measured at one time can be used to predict the response variable at a future time [4].

In regression linear modeling the terms of variable selections, and model selection are considered similar, since both selection have the same purposes. Another method of selection called order selection also has the same persposes with the two. Variable selection, model selection and order selection are methods that lead researchers to good models. The main distiction between the last two methods of selection is in addressing two different questions that usually arise in real problems. The first question is how many variables should be included in the selected model. The second question is which variables should be included in the selected model. The latter question is addressed by model selection, and the first question is addressed by order selection. Empirical comparisons for various data sets and different types of estimators (linear, subset selection, and k-nearest neighbor regression) studied, consistently outperforms AIC for all data sets. [5]
Examining each possible model formed by each possible subset of an available potential set of independent variables may be a reasonably good procedure to select a good model, but this requires numerous calculations. If there are k potential independent variables, \(2^k\) regression equations need to be examined. For example, if \(k=10\), there are \(2^{10}=1024\) regression equation need to be examined. This method of selection called the “all possible subsets method” is reasonable if there are few independent variables involved. For a larger number of independent variables, this method becomes cumbersome and inefficient for selecting model. Some criteria have been proposed to reduce the numerous calculations, such as forward selection criterion, backward selection criterion, cross validation criterion suggested by Alen in 1974, generalized cross validation, Akaike information criterion, and Mallows’ Cp Criteria.

Mallows’ Cp criterion proposed by Mallows in 1972 used in this study to select a good model will be reviewed in this paper. The minimum statistics of Cp is considered as the best model, but the statistics will be large if the ratio of the number of independent variables (p) to the number of observation (n) is high. Consequently the statistics Cp often overestimate the value of p in the model [6]. In this paper the overestimate of p in good model as the result of selecting model based on Mallow’s Cp criterion will be studied throughout a simulation in two cases, the first case is considered for the absence of correlation between some independent variables and one dependent variable, whereas the second case is the existence of correlation between some independent variables and a dependent variable. In the next section, the material and methods for simulation will be clearly stated, and the results will be discussed to the next section.

2. Materials and Methods

A simulation study of model selection in regression linear based on Mallows’ Cp criterion is the materials for this research. Reviews on the criterion as the selection method will be presented before simulation run in R software.

A typical multiple linear regression model with k independent variable X and one dependent variable Y is given by

\[ Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \epsilon_i, \text{ for } i = 1, 2, \ldots, n. \]  

(1)

Where \(Y_i\) denotes the \(i^{th}\) observable random variable Y, \(X_{ij}\) denotes the \(i^{th}\) observable random variable \(X_j\) (\(j = 1, 2, \ldots, k\)), \(\beta_0, \beta_1, \ldots, \beta_k\) denote the unknown parameters of the model, and \(\epsilon_i\) assumed independent and normally distributed with mean zero and variance \(\sigma^2\) denotes the error term in observation i. If \(Y, X, \beta, \text{ and } \epsilon\) are defined as matrix notations respectively in the order of \(nx1, nx(k+1), (k+1)x1\) and \(nx1\), then equation 1 can be simply expressed as

\[ Y = X \beta + \epsilon \]  

(2)

Here \(\epsilon\) is independent normally distributed with expectation, \(E(\epsilon) = 0\), and dispersion (covariance), \(\text{Cov}(\epsilon) = \sigma^2 I\), for I is an \(nxn\) identity matrix, and \(X\) is usually designated as the design matrix of the model.

Before the simulation method is stated, the method of selecting a good model that can be used to predict the observable independent random variable Y is reviewed by least square estimator method. In addition the Mallows’ Cp criterion is discussed.

Asume we have a fixed model as in equation 2. The principle of least square is to find the \(\hat{\beta}\) which minimizes the sum of error square, \(\epsilon^t \epsilon\), where the superscript t symbolizes the transpose matrix. The sum of square can be written as a function of \(\beta\), as follows,

\[ S(\beta) = (Y - X\beta)^t (Y - X\beta). \]  

(3)

Since \(S(\beta)\) is a non negative-real value of the quadratic function, the existence of a finite minimum \(S(\beta)\) is
assured. The solution for \( \beta \), denoted as \( \hat{\beta} \) minimized \( S(\beta) \) as the result of normal equation solution, called as least square estimator of \( \beta \):
\[
\hat{\beta} = (X^t X)^{-1}(X^t Y)
\]  
(4)

Where the estimate \( \hat{\beta} \) is unbiased, \( E(\hat{\beta}) = \beta \), with covariance \( \sigma^2(X^t X)^{-1} \), and the least square prediction model is \( \hat{Y} = X \hat{\beta} \). The residual sum of squares (RSS) of this model is defined as the summation of squared difference between the observed variable \( Y \) and least square estimate \( \hat{Y} \), and it is denoted as:
\[
\text{RSS} = (Y - X \hat{\beta})^t(Y - X \hat{\beta})
\] 
(5)

Since the \( E(Y) = X \beta \), and Sum of Square Error (SSE) is defined as the summation of the squared difference between expected value of \( Y \) and its least square estimate \( \hat{Y} \), then
\[
\text{SSE} = (X \beta - X \hat{\beta})^t(X \beta - X \hat{\beta})
\] 
(6)

A canonical form which is a transformation form of linear model using a new basis can separate the errors and estimate of model (2). If a new basis \( \{ \alpha_1, \alpha_2, \ldots, \alpha_k, \ldots, \alpha_n \} \) is a new orthonormal basis of space spanned by \( \{ X_1, X_2, \ldots, X_B, \ldots, X_n \} \) in \( n \)-Euclidean space \( V_n \), we can write \( Y = \sum_{i=1}^{n} \alpha_i Z_i \), because \( \{ Z_i \} \) are the coordinates of \( Y \) relative to the new basis. The canonical form can derived the mean of sum square errors (MSSE) to the formula as follows:
\[
\text{MSSE} = E(\text{SSE}) = k \sigma^2 + \sum_{i=k+1}^{n} \epsilon_i
\] 
(7)

Where \( k \sigma^2 \) is the variance and \( \sum_{i=k+1}^{n} \epsilon_i \) is the bias of the fitted model, [7].

Order selection means that the order of independent variables entering the model has been determined in advance. For example, if we have \( k \) independent variables \( \{ X_1, X_2, \ldots, X_k \} \) and their order is indicated by their indexes \( j=1,2,\ldots,k \), then models are examined in the following order:
\[
Y = \beta_0 + \beta_1 X_1 + \epsilon
\]
\[
Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon
\]
\[
\vdots
\]
\[
Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \epsilon
\]

Our goal is to select a good model that has minimum MSSE. In this case the problem is simply to select the order, say \( p \), for \( p = 1, 2, \ldots, k \). That is the number of independent variables included in the model.

Suppose that we have a linear model with \( p^{th} \) order, say \( Y = X_p \beta_p + \epsilon \) as define in equation (2), with the subscript \( p \) used at \( X \) and \( \beta \) describes that the model only involves \( p \) independent random variables. The canonical form of this model resulted \( \text{RSS}_p = \sum_{i=p+1}^{n} Z_i^2 \), the expectation of \( Z_i \), \( E(Z_i) = 0 \), for \( i > p \), and \( E(Z_i^2) = \text{Var}(Z_i) = \sigma^2 \). Therefore, the expectation of residual sum of squares \( \text{RSS}_p \), \( E(\text{RSS}_p) = (n-p) \sigma^2 \), and \( E(\hat{\sigma}^2) = E(\frac{\text{RSS}_p}{n-p}) = \sigma^2 \). If we assume that \( Y_i \) are from normal distribution, then \( Z_i \) are also normal, and \( \frac{\text{RSS}_p}{\sigma^2} \sim \chi^2_{n-p} \),[8]. Similarly, if we decrease the order of the model from \( p \) to \( p-1 \), \( \frac{\text{RSS}_p}{\sigma^2} \sim \chi^2_{n-p+1} \). Now let \( \Delta_p \) denote the difference of these residuals sum of squares, \( \Delta_p = \text{RSS}_{p-1} - \text{RSS}_p \). It follows that \( \frac{\Delta_p}{\sigma^2} \sim \chi^2_1 \).

A criterion of model selection proposed by Mallows in 1972 based on consideration of mean of sum squared errors . The MSSE for a model that involves \( p \) independent variables can be expressed as follows:
\[
\text{MSSE}_p = E(\text{RSS}_p) + 2p \sigma^2 - n \sigma^2
\]
The statistics $C_p$ is equation 8 below is an obvious choice of estimator for $MSSE_p$. [9]. In this criterion, we assume that the full model involving all $k$ independent variables is correct, and the reduced model with any subset of size $p$ of independent variables, for $p < k$, is evaluated by

$$C_p = RSS_p + 2p\sigma^2 - n\sigma^2$$  \hspace{1cm} (8)

In the practice, the $\sigma^2$ is usually estimated by its unbiased estimator, $\hat{\sigma}^2 = \frac{RSS}{n-p}$, for $RSS_k$ is the residual of sum squares of the full model. In this criterion the model that has minimum $C_p$ statistics is considered as the best model.

The method of simulation in this paper is set in two cases. For the first case, that is the absence of correlation between some independent variables and one dependent variable is set as follows; the first step, we simulate 500 observations of ten independent random variables $X_1, X_2, ..., X_{10}$ and one dependent variable $Y$ from a normal distribution with mean zero and variance 1. The second step, we use the forward method to order the independent random variables enter the model sequentially. Therefore, the problem is reduced into the order selection problem. And the choice of a good model is then just the problem of choosing the number of independent variables $p$ included in the model. Finally, the statistics of Mallows’ $C_p$ is used as the criterion for selecting a good model. This simulation is run 500 times by keeping the independent variables fixed but not for the dependent variable, $Y$.

For the second case, that is for the existence of correlation between some independent variables and a dependent variable is set as follows; the first step, we simulate 500 observations of ten independent random variables $X_1, X_2, ..., X_{10}$, and error term $\epsilon$ from a normal distribution with mean zero and variance 1. The second step, we generate the dependent variable $Y$ using a model $Y = 2X_1 + 5X_2 + \epsilon$. The third step, we use the forward method to order the independent random variables enter the model sequentially. Therefore, the problem is reduced into the order selection problem. And the choice of a good model is then just the problem of choosing the number of independent variables $p$ included in the model. Finally, the statistics of Mallows’ $C_p$ is used as the criterion for selecting a good model. This simulation is run 500 times by keeping the independent variables fixed but for the dependent variable, $Y$ is generated differently through resampling of $\epsilon$.

3. Results and Discussions
The result of simulation for the first case, the absence of correlation between some independent variables and one dependent variable the absence, is presented in table 1 for 5 simulations, and the distribution of the number of $p$ selected in the model is shown by figure 1, and the result of simulation for the second case, the existence of correlation between some independent variables and a dependent variable, is presented in table 2, for 5 simulations, and the distribution of the number of $p$ selected in the model is shown by figure 2.

In the simulation of the first case, even there is no correlation between the ten independent random variables and one dependent variable $Y$ from a normal distribution with mean zero and variance 1, the distribution of estimate $p$, the number of independent random variables selected in the models, in figure 1 shows some independent random variables included in models, but it deeply decreases as the number of estimate $p$ increases. For 5 simulations that are tabulated in table 1, three selected models do not include any random variable, but there are two selected models that still included two and three independent random variables each.

The simulation results of the second case, where the correlations are set between the two independent random variables and one dependent random variable describe in table 2, and figure 2. The result shown in table 2 indicated that the two independent random variables always included in the selected model. Whereas the distribution of estimate $p$, the number of independent variables selected in the models, whose correlations to the dependent random variables are set, indicated that the independent random variables included in models always chosen, and it decreases as the number of estimate $p$ increases.
Figure 1. Distribution of $p$ estimate for the first case simulation

Figure 2. Distribution of $p$ estimate for the second case simulation

Table 1. The simulation result for the first case

| Simulation Number | $C_p$ Values | $p$ estimate | Selected Variables |
|-------------------|--------------|--------------|--------------------|
| 1                 | -4.68        | 0            | None               |
| 2                 | -3.81        | 0            | None               |
| 3                 | 0.57         | 0            | None               |
| 4                 | -1.44        | 2            | $X_5, X_9$         |
| 5                 | -1.38        | 3            | $X_1, X_8, X_{10}$ |

Table 2. The simulation result for the second case

| Simulation Number | $C_p$ Values | $p$ estimate | Selected Variables |
|-------------------|--------------|--------------|--------------------|
| 1                 | -1.71        | 2            | $X_1, X_2$         |
| 2                 | -1.11        | 3            | $X_1, X_2, X_8$   |
| 3                 | 3.50         | 5            | $X_1, X_2, X_3, X_6, X_{10}$ |
| 4                 | 1.95         | 5            | $X_1, X_2, X_3, X_8, X_{10}$ |
| 5                 | 2.30         | 6            | $X_1, X_2, X_3, X_6, X_8, X_{10}$ |
4. Conclusions
The proposed method of simulation for selecting a good model in regression linear used order selection method. The criterion to find the order of independent random variables entering the model used the forward criterion. The selection criterion to find a good model in this study used Mallows’ $C_p$ criterion. Two cases of simulations are run, one case is conducted by considering the absence of correlation between the ten independent variables and one dependent variable, whereas the another case is conducted by considering the existence of correlation between two independent variables and one dependent variable.

The results for both cases shows that the Mallows $C_p$ criterion frequently overestimate the selected model, even for the absence of correlation the criteria selected some of variables included in the model.

5. Acknowledgement
In this opportunity, I would like to express my sincere gratitude to the Research Department of the Universitas Sumatera Utara that has provided financing aid for this research under the contract number of No: 89/UN5.2.3.1./PPM/KP-TALENTA USU/2018. I would also like to thank the Dean of FMIPA, Universitas Sumatera Utara, for his given permit to me to carry out this research and to present it in this International seminar.

References
[1] Rao C and Wu Y 2001 Lecture Notes-Monograph Series 38 1
[2] Miller A J 2002 Subset Selection in Regression (London: Chapman and Hall)
[3] Chaurasia A and Harel O 2012 Health Serv. Outcomes Res. Method. 12 219
[4] Miller A J 1990 Subset Selection in Regression (London: Chapman and Hall)
[5] Cherkassky V & Yunqian M 2003 Neural Computation 15 1691
[6] Hocking R R 1972 Technometrics 14 967
[7] Darnius O 1996 A Simulation Study on Order and Model Selection Thesis
[8] Christensen R 1987 Plane Answers to Complex Quaeastions: The Theory of Linear Models (USA: Spring-Verlag)
[9] Thompson M L 1978 International Statistics Review 46 1
[10] Nemes S, Steineck G and Dickman P W 2010 BMC Medical Research Methodology 10 108