Model selection in regression linear: a simulation based on akaike’s information criterion

O Darnius*, Normalina and A Manurung
FMIPA, Universitas Sumatera Utara, Indonesia

*Corresponding author: open@usu.ac.id

Abstract. Akaike’s Information Criterion (AIC) was firstly announced by Akaike in 1971. In linear regression modelling, AIC is proposed as a model selection criterion since it estimates the quality of each model relative to other models. In this paper we demonstrate the use of AIC criterion to estimate p, the number of selected variables in regression linear model through a simulation study. We simulate two particular cases, namely orthogonal and non-orthogonal cases. The orthogonal case is run where there is totally no correlation between any independent variable and one dependent variable, whereas for the orthogonal case is run where there is a correlation between some independent variables and one dependent variable. The simulation results are used to investigate of the overestimate number of independent variables selected in the model for two cases. Although the two cases produce the overestimate number of independent variables, most of the time the orthogonal case still provide less overestimate of independent variables than the non orthogonal case.

1. Introduction
Model selection is one of the most widely used of all statistical procedures and a common task in regression analysis. In regression linear modelling the terms of model selection and variable selection are considered similar, since both selections have the same purposes. Variable selection, the method of searching for subset(s) of variables that "best" explain the response through a model, makes the term is similar to model selection, where "best" is defined with respect to a specific purpose such as model interpretation or prediction. The knowledge and practice of variable selection methods were developed when small data grew into early-size big data circa late 1960s / early 1970s [1]. Sometimes it is desirable to select the most important predictors without losing too much prediction accuracy [2].

The main distinction between the 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, and 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 variable selection. Variable selection, though well studied in the statistical literature, remains rich in unsolved problems[3], [4] especially as related to missing response data [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, 2k = 1024 regression equation need to be examined [6].

Empirical comparisons for various data sets and different types of estimators (linear, subset selection, and k-nearest neighbour regression) studied, consistently outperforms AIC for all data sets.
Some variable selection methods have been proposed in literature such as backward elimination, forward selection and stepwise regression. Some other criterion-based procedures like: Mallows Cp, the Akaike’s Information Criterion (AIC), the Bayes Information Criterion (BIC), and Information complexity (ICOMP) have been widely used in many studies. These procedures choose efficient variables after obtaining parameter estimators. In this paper the AIC will be used as selection criterion to estimate \( p \), the number of selected variables in regression linear model through simulation. Two particular cases, namely orthogonal and non-orthogonal cases will be run in the simulation. The orthogonal case is run where there is totally no correlation between any independent variable and one dependent variable, whereas for the orthogonal case is run where there is a correlation between some independent variables and one dependent variable. The number of models that must be considered in the selection of linear regression models will increase in line with the increase in the independent variables that must be taken into account. If there are \( n \) independent variables, then there are \( 2^n \) models that must be considered. In this study, the accuracy of the model selection method with the Akaike’s information criterion is examined. The study is conducted with a simulation.

2. Methods

2.1. Model and Notation

A typical multiple linear regression model with \( k \) independent variable \( X \), and one dependent variable \( Y \) is given by:

\[
Y_t = \beta_0 + \beta_1 X_{t1} + \beta_2 X_{t2} + \cdots + \beta_k X_{tk} + \epsilon_t, \quad \text{for} \ t = 1, 2, ..., n. \tag{1}
\]

where \( Y_t \) denotes the \( i \)-th observable random variable \( Y \), \( X_{ij} \) denotes the \( i \)-th observable random variable \( X_j \) (\( j = 1, 2, ..., k \)), \( \beta_0, \beta_1, ..., \beta_k \) denote the unknown parameters of the model, and \( \epsilon_t \) assumed independent and normally distributed with mean zero and variance \( \sigma^2 \) denotes the error term in observation \( i \). Equation 1 can be denoted in matrix form as follows:

\[
Y = X\beta + \epsilon \tag{2}
\]

where \( Y \) is a \( n \)-dimensional vector, \( X \) is a \( n \times (k + 1) \) matrix, \( \beta \) is a \( (k + 1) \)-dimensional vector, and \( \epsilon \) is the \( n \)-dimensional (uncorrelated) error term which is normally distributed with zero mean and constant variance \( (\sigma^2) \). The principle of least square is applied to find the estimate of parametr \( \beta \), which minimizes the sum of error square, \( \epsilon^t \epsilon \), the superscript \( t \) symbolizes the transpose matrix. A canonical form which is a transformation form of the linear model using a new basis can separate the the errors and estimate of model (2). If a new basis \( \{a_1, a_2, ..., a_k, ..., a_n\} \) is a new orthonormal basis of space spanned by \( \{X_1, X_2, ..., X_k, ..., X_n\} \) in \( n \)-Euclidean space \( V_n \), we can write \( Y = \sum_{i=1}^{n} a_i Z_i \), for\{\( Z_i \}\} are the coordinates of \( Y \) relatives to the new basis. The canonical form can derived the mean of sum square errors (MSSE) to the formula as follows:

\[
\text{MSSE} = E(\epsilon^2) = k\sigma^2 + \sum_{i=k+1}^{n} \epsilon_i \tag{3}
\]

Where \( k\sigma^2 \) is the variance and \( \sum_{i=k+1}^{n} \epsilon_i \) is the bias of the fitted model, [8].

A simulation method proposed in this paper uses order selection, namely the selection of the independent variables included in the model has been determined in advance. For example, if there are \( k \) independent variables, \( X_1, X_2, ..., X_k \), with the order indicates by indexes \( j=1, 2, ..., k \), then models are examined in the following order:

\[
Y = \beta_0 + \beta_1 X_1 + \epsilon \\
Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon \\
\vdots \\
Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_{ik} + \epsilon
\]

In this case, the selection problem is simply to select the order, say \( p \), for \( p = 1, 2, ..., k \). That is the number of independent variables included in the model.

Suppose that we have a linear model with \( p \)-order, say \( Y = X_p \beta_p + \epsilon \) as defined in equation (2), and the subspet \( p \) used at \( X \) and \( \beta \) indicates that the model only involves \( p \) independent random...
variables, then the canonical form of this model resulted residual sum of square, $RSS_p = \sum_{i=p+1}^{n} Z_i^2$, and the expectation of $Z_i$, $E(Z_i) = 0$, for $i > p$, and $E(Z_i^2) = Var(Z_i) = \sigma^2$. Therefore, the expectation of residual sum of squares $RSS_p$, $E(RSS_p) = (n-p) \sigma^2$, and $E(\sigma^2) = E \left( \frac{RSS_p}{n-p} \right) = \sigma^2$.

If $Y$ are from normally distributed, then $Z_i$ are also normal, and $\frac{RSS_p}{\sigma^2} \sim \chi^2_{n-p}$. Analogically, if the order of the model is decreed by 1, from $p$ to $p-1$, then $\frac{RSS_{p-1}}{\sigma^2} \sim \chi^2_{n-p+1}$. If we let $\Delta_p$ denote the difference of these residuals sum of squares, $\Delta_p = RSS_{p-1} - RSS_p$, then $\frac{\Delta_p}{\sigma^2} \sim \chi^2_1$.

A criterion of model selection proposed by Akaike in 1971 based on an expected log-likelihood of any model with $p$ independent variables, for $p < k$. The $AIC_p$ statistic is then given by equation (4);

$$AIC_p = RSS_p \text{Exp} \left( \frac{\Delta_p}{n} \right)$$

In this criterion the model with minimum $AIC_p$ value is considered a good model.

2.2. Simulations Method

In order to investigate the distribution of $p$ order estimate, the best model with one independent variable, we performed two cases of simulations, following the method used by [6] when using Mallow Cp criterion. Firstly, we simulate the orthogonal case, where there is no correlation between some independent variables and one dependent variable. The steps we used for the simulations as follows; in 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 next step, we use the forward method to order the independent random variables entering model sequentially to reduce the problems into order selection problem. Consequently 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 $AIC_p$ is used as the criterion for selecting a good model. The simulation is run 500 times by keeping the independent variables fixed but not for the dependent variable, $Y$.

Secondly, we simulate the non-orthogonal case, it means there is correlation between some independent variables and one dependent variable. The steps we used for the case, as follows; for the first step, 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 next step, we use the forward method to order the independent random variables entering model sequentially which reduced the problem into order selection problem. Finally, the statistics of $AIC_p$ is used as the criterion for selecting a good model.

3. Results and Discussion

The result of the simulations for the orthogonal case in terms of order selection using Akaike’s information criterion is the distribution of order estimate $p$ which is the number of independent variables selected in the model considered as a good model. The distribution is depicted in Figure 1, and the simulation result for the first five runs of the order estimate $p$ is tabled in Table 1.

For the non-orthogonal case, the results of the simulations in terms of order selection using Akaike’s information criterion is also the distribution of order estimate $p$ which is the number of independent variables selected in the model considered as a good model. The distribution is depicted in Figure 2, and the simulation result for the first five runs of the order estimate $p$ is tabled in Table 2.

The distributions show us that the Akaike’s information criterion frequently over estimate the selected model. For the non-orthogonal case, it can be seen from the table that although the criterion successfully selected the variables which are correlated to the dependent variables, but most of the time it still selected the variables those are not correlated to the dependent variable. Likewise in the orthogonal case, the criterion is also overestimate the order estimate $p$. Although, the criterion did not select any of the independent variable in this case, it occurred only one out of five times as shown in
The table 2 shows that 5 out of 500 observations in the simulation of non-orthogonal case, the Akaike information criterion consistently selected the independent variables $X_1$, and $X_2$ which are previously set to be correlated to the dependent variable.

4. Conclusion

The model selection method to find a good regression linear model proposed in this study is Akaike’s information criterion. The distribution of $p$, the number of independent variables selected in the model is estimated through simulation. The simulation is run in two cases, i.e., orthogonal case, and non-orthogonal case. The simulation produced overestimate number of independent variables selected in the model. Although the result showed that the Akaike’s information criterion frequently
overestimated the model, but the criterion consistently selected the independent variables those have correlations to the dependent variable.

References
[1] Bruce R 2010 J. Target. Meas. Anal. Mark. 18 65
[2] Chen Y, Du P and Wang Y 2014 WIREs Comput. Stat. 6 1
[3] Rao C and Wu Y 2001 Lect. Notes-Monogr. Ser. 38 1
[4] Miller A J 2002 Lond.: Chapman Hall
[5] Chaurasia A and Harel O H 2012 Serv Outcomes Res Method
[6] O Darnius and G Tarigan 2018 J. Phys.: Conf. Ser. 1116 022008
[7] Cherkassky V and Yunqian M 2003 Neural Comput. 15 1691
[8] Akaike H 1974 IEEE Trans. Autom. Control. 19 716
[9] Burnham K and Anderson D 2004 Sociol. Methods Res. 33 261
[10] Harel O 2009 J. Appl. Stat. 36 1109