Applied Mathematical Sciences, Vol. 9, 2015, no. 12, 557 - 568
HIKARI Ltd, www.m-hikari.com
http://dx.doi.org/10.12988/ams.2015.411989

Performance of Variable Selection Methods in Predicting Language Proficiency Using Language Learning Proficiency

Johannah Jamalul Kiram and Jumat Sulaiman

Faculty of Science and Natural Resources, Universiti Malaysia Sabah
88400 Kota Kinabalu, Sabah, Malaysia

Suyansah Swanto

Faculty of Psychology and Education, Universiti Malaysia Sabah
88400 Kota Kinabalu, Sabah, Malaysia

Wardatul Akmam Din

Centre for the Promotion of Knowledge and Language Learning
Universiti Malaysia Sabah
88400 Kota Kinabalu,Sabah, Malaysia

Abstract

The goal of this study was to compare two multiple regression models generated using two different variable selection methods in order to determine which variable method was more reliable in constructing a better model. Two hundred thirty pre-university students of UMS participated by answering a self-report questionnaire called the Strategy Inventory for Language Learning (SILL), a background questionnaire, and then sat for the Malaysian English University Test (MUET). Selected statistical tests were used to compare models.

Keywords: Variable selection, multiple regression, Language learning strategies, Akaike information criterion

1 Introduction

Building a mathematical model requires a series of process. One of them is the process of variable selection. Variable selection helps select the best subset among the predictors. This helps to produce better quality models that predict more accurate outcomes. However, there are numerous types of variable selection methods that result in different subset of predictors. Generally when a research is done, the stepwise regression and the Akaike's information criterion (AIC) method are the ones that are popularly used where most statistical analysis software have these methods built into the system and ready to be used. With multiple methods being introduced from time to time, the requirement to utilize a method that produces the best model is essential. This study has been restricted to compare these two methods.

As the country's second language, English plays a huge role in defining a pupil's future. It is a compulsory subject taught in primary and secondary school in its formal education system and a widely spoken language in most private universities and colleges. At a pre-university level, it is also a compulsory course as students are required to sit for the Malaysian University English Test (MUET) prior to enrolment for their first degrees. In fact, its results are a pre-requisite for entering certain undergraduate programs at Public Higher Educational Institutions (IPTA). MUET was introduced in 1999 and is a frequently used examination to test English proficiency. There are four components tested in MUET; Listening (45 marks), Speaking (45 marks), Reading (120 Marks) and Writing (90 marks). These four components are used to measure language performance in listening comprehension, grammar skill, communicative ability and writing skill. Proficiency is graded using 6 bands with band 6 being the highest, and band 1 the lowest. There are three exam sessions in a year, which is in March, July or November.

A number of initiatives have been introduced towards using more English in Malaysian education settings over the years. However, an average Malaysian student still finds it hard to master the language adequately especially in terms of verbal fluency, writing compositions in English and also applying proper grammar, which has ultimately caused students to get unsatisfactory examination results. This brings problems for academically able students to get into universities as the mastery of English is considered an advantage in that it help students to gain access to information and all sorts of different knowledge which are mostly written in English.

Language learning strategies (LLS) used by students have often been considered as factors, among other things, that influence their language proficiency and there have been numerous studies on the LLS. Weinstein and Mayer [2] define learning strategies as specific behaviors and thoughts that influence learner's encoding process. It is believed that a learning strategy facilitates the learner's acquisi-

tion, especially in terms of language input storage and retrieval of information. Green and Oxford [12] defines it as "specific action or techniques that students use, often intentionally, to improve their progress in developing second language skills."

This study involved two hundred and thirty pre-university students at Universiti Malaysia Sabah in October 2013. A background questionnaire was handed out to the student. The questionnaire includes gender, age, nationality, state of origin, language used at home, Malaysian Certificate of Education (SPM) results for English language, previous secondary school type, household income and parents' education, together with a self-report questionnaire called the Strategy Inventory for Language Learning (SILL) [15] to identify their learning strategies. Many strategy questionnaires have been constructed [4, 16, 17] but the SILL has been reported to have a higher degree of reliability and validity [14]. It is the most often used strategy questionnaire in many different countries to identify learner's language learning strategy. The participating students also sat for MUET in November of 2013. Their identities were kept confidential and their results were only recorded based on their matriculation number.

Models are only approximation of the actual reality. There will never be a right or wrong model. However, discovering a model that closely approximates the outcomes is possible. This study was trying to establish the relationship between LLS and language proficiency by using a mathematical model. We began by comparing two models that used different basic variable selection methods, stepwise regression and Akaike information criterion. We then proceeded on by testing each of the model's adequacy using the Global F-test, root mean square error (RMSE), the R^2 and adjusted R^2. Comparing these two models would aid academicians, teachers and students of the English language in knowing which LLS they should be focusing on in order to attain better language proficiency. This study also helps future studies in modeling the relationship between LLS and language proficiency.

2 Research Design

Two hundred and thirty pre-university science students of University Malaysia Sabah participated in this study. These students were all at the age of 18 years old and were under the Preparatory Centre for Science and Technology 1-year pre-university program. These students were divided into two different lecture groups, randomly selected by the pre-university centre. Students had been informed verbally that they were part of a study to identify their language learning strategies and that there were no right or wrong answers to the questionnaires given. These students were given 20 minutes to answer both the SILL and the background questionnaire simultaneously during a class in October 2013.

2.1 Strategy Inventory for Language Learning

The SILL version 7.0 [15] was used to measure learning strategies preferences. It was a self-report questionnaire divided into six sections, each of which represented a particular strategy, both direct and indirect. The direct strategies were memory, cognitive and compensation. As for the indirect strategies, there were metacognitive, affective and social. There were 50 items and students responded to each item using a 5-point Likert scale with 1 being "Never or almost never true of me", 2 "Usually not true of me", 3 "Somewhat true of me", 4 "Usually true of me", and 5 "Always or almost always true of me". The questionnaire was prepared in both English and Malay.

2.2 Background questionnaire

The background component of the questionnaire requires the research subjects to provide the following information: matriculation number, gender, age, nationality, state, language used at home, SPM result for English Language, previous secondary school type, household income, parents' highest education and the student's use of English whether as a first language, second language or foreign language. This part of the questionnaire was constructed to elicit the background of the participants and to specify the criteria for selection of samples.

2.3 Variable Selection Methods

This section discusses two variable selection methods focused in this paper which are Stepwise regression and the Akaike information criterion. Further explanations are as follows.

2.3.1 Stepwise Regression

This study is a combination of forward selection method and backward elimination method where selection of the predictors depends on the predictor's significance whereby its p-value and the model's R-squared with the existence of that particular predictor is examined. The backward elimination begins by including all the predictors and eliminates the insignificant ones one at a time, whereas the forward selection method just does the total opposite of it [18]. Thus, the stepwise allows you to add or remove predictors one at a time. The number of predictors that remains is determined by the level of significance pre-set for inclusion and exclusion of predictors.

2.3.2 Akaike Information Criterion

The Akaike information criterion (AIC) is by and large looked at as a first model selection criterion that was widely accepted by practitioners. Introduced by Hirotugu Akaike [6], it was based on Kullback-Leibler information [7, 8] where

Akaike discovered a valid connection between Kullback-Leibler information and likelihood theory. A brief explanation can be found in Burnham and Anderson [11].

Most statistical software packages will already provide AIC values for general linear models. The calculation of AIC is divided into two ways; using maximum likelihood estimator where

$$AIC = -2\ln(L) + 2k \tag{1}$$

and using residual sum of square (RSS) where

$$AIC = n\left[\ln\left(\frac{RSS}{n}\right)\right] + 2k \tag{2}$$

where *n* is the sample size, *L* is the maximum likelihood estimate for the model, k is the intercept in the model, and $RSS = \sum(\hat{\epsilon}_i)^2$.

The $\hat{\epsilon}_i$ are estimated residuals in the fitted model. These fitted models are then ranked by AIC and the best approximating model is the one with the lowest AIC value. Next, AIC takes into account how well the model fits the data with model adequacy tests. Ultimately, models with greater numbers of fitted parameters (k) will have higher AIC values. Thus, models with less number of parameters are usually preferred [13].

2.4 Constructing Linear Model

The main objective of this study is to construct the best mathematical model based on multiple regression analysis, using the best possible subset of predictors. Basically, regression analysis is concerned with the study of the relationship between one variable called the explained or dependent variable and one or more other variables called independent or explanatory variable [3]. Multiple regression involves more than one independent variable. It is the more commonly used statistical analysis tool when it comes to explaining the relationships between a dependent variable (Y) and a number of independent variables (X_i). It creates a regression rule that explains the mean of the distribution of a single response variable for particular values of explanatory variables. For regression models, there are two common forms of mathematical models that can be constructed which are linear models and non linear models. Examples of non linear models cases are as follows; [1, 5, 9, 10]. However, this paper deals with linear models which is constructed as follows:

$$Y_i = \beta_0 + X_{i1}\beta_1 + \cdots + X_{ik}\beta_k + \mu_i \tag{3}$$

where the $Y_i's$ are random variables, X_{ij} are known constants, μ_i are independent random variables, β_i and σ^2 are unknown parameters, and *k* stands for the number

of explanatory variables. The parameters that need to be estimated are β_i for all values of *i* using the method of ordinary least square [3]. These parameters are to be estimated to produce the best population regression model possible. The point here is to choose $\hat{\beta}_i$ that gives the least amount of disturbance possible.

To construct simple linear regression, this model comes with a number of assumptions. Hence, the assumptions of multiple linear regression models are [3]:

i) The expected value of the error term μ is zero. That is, $E\ (\mu_i|X_i) = 0$
ii) The variance of each μ_i is constant. Also known as homoscedastic. That is, $Var(\mu_i|X_i) = \sigma^2$
iii) No autocorrelation. That is, $cov(\mu_i\mu_j) = 0;\ i \neq j.$
iv) No exact collinearity exists between X_i and X_j where $i \neq j.$
v) The error term μ follows normal distribution. That is, $\mu_i \sim N(0, \sigma^2)$.

In this study, we assumed that the data obtained followed a linear model and that all these assumptions were met.

After removing every possible variable to improve the model, it is then followed by checking the model's adequacy. We proceeded on to see the prediction error of these explanatory variables using the Root Mean Squared Error (RMSE). The RMSE is the standard deviation of the random error, ε. It is a measure of difference between the estimated values and the actual values. The lower the value, the better the model is [3]. The formula of RMSE is given as

$$RMSE = \sqrt[2]{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)}{n}} \tag{4}$$

The model's prediction accuracy was then tested by using R-squared, which is the proportion of variability in data set. It is an intuitive scale that ranges from zero to one. The value of the model's R-squared gets closer to one as the model's prediction gets better [3]. Hence, when the value of the model's R-squared is approaching zero, the model is said to be a weak predictor. The formula for calculating R-squared is

$$R^2 = \frac{SS_{reg}}{SS_y} \tag{5}$$

where $SS_{reg} = \sum_{i-1}^{n}(\hat{y}_i - \bar{y})^2$ and $SS_y = \sum_{i-1}^{n}(y_i - \bar{y})^2$.

When there are more than one predictor variables (independent variables), the Adjusted R-squared has to be looked into because the value of R-squared is influenced by the number of independent variables in the model. The adjusted R-squared is the proportion of total variance that is explained by the model. It

encompasses the model's degree of freedom. The formula for the Adjusted R-squared is

$$Adjusted\ R^2 = 1 - (1 - R^2)\frac{n-1}{n-m-1} \quad (6)$$

where *m* is the number of independent variable included in the model.

Two models are nested if both contain the same terms but one has at least one additional term. Neither of the models in this paper are full models. However, they differ in terms of variable selected. Another way to decide which model is better is to hypothesize

H_0= reduced model is adequate
H_1= otherwise.

To test this hypothesis, we use F-test, where

$$F = \frac{\left(\frac{drop\ in\ SSE}{Number\ of\ extra\ terms}\right)}{s^2\ for\ full\ model}$$

where SSE is the sum of squared residuals.

3 Results and Discussion

The analysis began by fitting the model into multiple linear regression model. Then we analyzed its quantile-quantile plots. Next, model adequacy test were done using root mean square error, R^2, and the adjusted R^2. Finally, we tested the hypothesis of which is a better model using ANOVA, where the F-test was calculated.

3.1 The fitted model

Fitting the model begins by selecting the variables according to the variable selection method. Beginning with the full model where,

response variable (Y) = Proficiency (MUET results);

independent variables (X_i) = memory, cognitive, comprehensive, metacognitive, affective and social.

Thus, the multiple regression model in the form of equation (3) before estimating the parameters and variable selection, is

$$Y_i = \beta_0 + X_{i1}\beta_1 + X_{i2}\beta_2 + X_{i3}\beta_3 + X_{i4}\beta_4 + X_{i5}\beta_5 + X_{i6}\beta_6 + \varepsilon_i \quad (7)$$

where $i = 1, 2, 3, \ldots, n,\ n = 56$, and $X_{i1}, X_{i2}, \ldots, X_{i6}$ represents each independent variable as stated before, respectively. We begin by comparing the fitted model as selected using both variable selection methods.

With stepwise regression, the final model suggested to estimate language proficiency is

$$\hat{y} = 137.744 + 8.549\, x_2 + 8.056\, x_4 - 13.238\, x_5 \quad (8)$$

where cognitive (x_2), metacognitive (x_4) and affective (x_5) are the only independent variables that were significant.

With AIC, the final model suggested to estimate language proficiency is

$$\hat{y} = 140.481 - 3.985\, x_1 + 10.437\, x_2 + 8.196\, x_4 - 12.448\, x_5 \quad (9)$$

where memory (x_1) was retained in the model as opposed to the stepwise model. These models appear similar with high intercepts.

Quantile-quantile (QQ) plots for both models were plotted out and the results are as in Figure 1(a) and Figure 1(b). Both plots appeared to be slightly skewed which may suggest that the data follows chi-square instead of normal distribution.

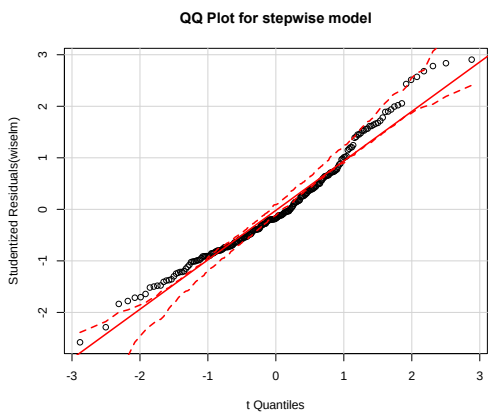

Figure 1(a): Stepwise QQ

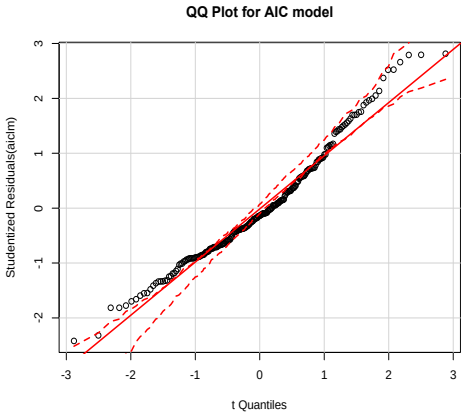

Figure 1(b): AIC QQ

3.2 Root Mean Square Error

The RMSE for both models also proved to be similar, as in Table 1. However, the RMSE for the model using AIC variable selection was slightly smaller, proving that keeping memory strategy in the model, does improve the model. Despite the difference, the high value of RMSE for both models needs to be investigated as it shows that the sum of the difference between the predicted $\hat{y}$ and the real data y is large.

Table 1: RMSE values of each model

Model	RMSE
Stepwise model	22.21
AIC model	22.16

3.3 R^2 and Adjusted R^2

The R^2 value of the model using stepwise variable selection is 0.156 whereas the model using AIC variable selection is 0.148. As shown in Table 2, both values of the R^2 and adjusted R^2 for both models tends to approach zero with the model using stepwise regression variable selection being slightly higher at 0.141 as compared to the model using AIC variable selection at 0.137. Contradicting the calculated RMSE in the previous section, these values suggest that the model using Stepwise regression variable selection give slightly better prediction.

Table 2: The R^2 and Adjusted R^2 for each model

Model	R^2	Adjusted R^2
Stepwise model	0.156	0.141
AIC model	0.148	0.137

3.4 Comparing nested models

In this study, the reduced model is said to be the model using Stepwise variable selection (Stepwise model). Using F-test to compare the R^2, the R statistical software uses the command "anova(model1, model2)". The command gives the result as in Table 3. The p-value suggested insignificant as it was not less than $\alpha = 0.05$. Hence, failing to reject the hypothesis where the reduced model is adequate.

Table 3: ANOVA

Model	RSS	Sum of Square	F-test	p-value
AIC model	111511			
Stepwise model	110510	1001.8	2.0396	0.1546

4 Conclusions

Both models have been tested for adequacy and it was found the model using AIC variable selection has slightly better RMSE values as compared to the model using stepwise regression variable selection. This suggests that the model using AIC variable selection has better consistency in predicting. However, the adjusted R^2 values for both model suggests that the model using Stepwise variable selection has better prediction. While using F-test to compare the residual sum of squares, the model using AIC variable selection is proven to be more adequate.

Despite these calculations, the results suggested shows ambiguity in the data. While Stepwise regression variable selection selects variable by looking at its significance, AIC variable selection selects variables based on the calculated AIC, neither can it be said to produce a "better" model. The QQ-plot also suggested that the data may be following chi-square distribution instead of normal distribution, however too vague to be concluded as either or. Further investigations needs to be done to create a well-improved model to help produce better predictions that will aid academicians, teachers and students in mastering better English.

References

[1] A. Barrea and M. Hernández. Pareto front for chemotherapy schedules. *Applied Mathematical Sciences,* **6** (2012), no. 116, 5789-5800.

[2] C. E. Weinstein and R. E. Mayer. The teaching of learning strategies. *Innovation Abstracts*, **5** (1986), no. 32, 3-4.

[3] D. Gujarati, Essentials of Econometrics. Boston [MA]: McGraw-Hill/Irwin (1993) p. 201, 203, 417.

[4] E. Bialystok, The role of conscious strategies in second language proficiency. *Modern Language Journal,* **65** (1981)**,** 24-35. http://dx.doi.org/10.1111/j.1540-4781.1981.tb00949.x_

[5] E. S. Lakshminarayanan and M. Sumathi. On representation of age-dependent stretched exponent in the extended Weibull model. *International Journal of Contemporary Mathematical Sciences,* **6** (2011), no. 4, 177-190.

[6] H. Akaike. Information theory as an extension of the maximum likelihood

principle. In: Petrov BN, Csáki F (eds) Second international symposium on information theory. Akadémiai Kiadó, Budapest, (1973) pp 267–281.

[7] H. Akaike. Information measures and model selection. *International Statistical Institute*, **44** (1983), 277-291.

[8] H. Akaike. Information theory and an extension of the maximum likelihood principle. Pages 610-624. in S. Kotz, and N. L. Johnson (Eds.) Breakthroughs in statistics, **1** (1992). Springer-Verlag, London.
http://dx.doi.org/10.1007/978-1-4612-0919-5_38

[9] H. Wijayanto, I. M. Sumertajaya, A. Fitrianto, and S. Wahyuni. Statistical models for chili productivity, *Applied Mathematical Sciences,* **8** (2014), no. 2, 69-79. http://dx.doi.org/10.12988/ams.2014.311616

[10] I. Nizovtseva. Nonlinear model of the mushy layer in the time-dependent crystallization of sea water in ice cracks, *Advanced Studies in Theoretical Physics,* **7** (2013), no. 21, 1011-1016. http://dx.doi.org/10.12988/astp.2013.39112

[11] K. P. Burnham, and D. R. Anderson. Multimodel inference understanding AIC and BIC in model selection. *Sociological methods & research,* **33.2** (2004): 261-304. http://dx.doi.org/10.1177/0049124104268644

[12] M. Green and R. Oxford, A closer look at learning strategies, L2 proficiency, and gender. *TESOL Quarterly,* **29** (1995), 261-297.
http://dx.doi.org/10.2307/3587625

[13] M. R. E. Symonds and A. Moussalli. A brief guide to model selection, multimodel inference and model averaging in behavioural ecology using Akaike's information criterion. *Behavioral Ecology and Sociobiology,* **65.1** (2011), 13-21.
http://dx.doi.org/10.1007/s00265-010-1037-6

[14] R. L. Oxford. Employing a questionnaire to assess the use of language learning strategies. *Applied Language Learning,* **7** (1996), no. 1 & 2, 25-45.

[15] R. Oxford, *Language Learning Strategies: What Every Teacher Should Know*. New York: Newbury House (1990).

[16] R. Politzer, An exploratory study of self-reported language learning behaviors and their relation to achievement. *Studies in Second Language Acquisition,* **6** (1983)**,** 54-65. http://dx.doi.org/10.1017/s0272263100000292

[17] R. Politzer and M. McGroarty. An explanatory study of learning behaviours and their relationship to gains in linguistic and communicative competence. *TESOL Quaterly,* **19** (1985)**,** 103-124. http://dx.doi.org/10.2307/3586774

[18] Wang, George CS, and Chaman L. Jain. *Regression analysis: modeling & forecasting*. Institute of Business Forecastig, 2003.

Received: December 8, 2014; Published: January 9, 2015