Iraqi Journal of Statistical Science (16) 2010
p.p. [1-12]

Multiple Regression Model Selection by Information Criteria

Zakaria Y. AL-Jammal*

Abstract

In this paper, I considered the problem of selection a model from a collection of candidate models specified by multiple linear regression model. Two criteria are used and compared, they are Akaike information criterion (AIC) and Bayesian information criterion (BIC). A real data set is considered as an application case. I preferred the model that the AIC chosen rather than BIC depending on the values of R^2_{adj} and the $root\ MSE$.

.

اختيار نموذج الانحدار المتعدد باستخدام معايير المعلومات

المستخلص

في هذا البحث تمت دراسة مشكلة اختيار أفضل نموذج من نماذج الانحدار المتعددة باستخدام معيار أكاكي ومعيار بيز للمعلومات . تمت دراسة مجموعتين من البيانات الحقيقية وبالاعتماد على قيمتي معامل التحديد المصحح وجذر متوسط مربعات الخطأ تم اختيار نموذج معيار معلومات أكاكي بوصفه أفضل نموذج مقارنة مع نموذج معيار معلومات بيز.

Keywords: Akaike information criterion, Bayesian information criterion, multiple linear regression, Akaike weight

* Assist. Lecturer / Statistics and Informatics Dept./ Computer Science and Mathematics college / Mosul University / Mosul / IRAQ.

Received:8/3 /2009 _____Accepted: 25 /10 / 2009

1-Introduction

Before three decades and in the recent years, the statistical literature has placed more and more emphasis on model selection criteria. Many model selection criteria were used such as the Akaike information criterion (AIC) (Akaike, 1973), corrected Akaike information criterion ($AICC$)(Hurvich & Tsai, 1989), Bayesian information criterion (BIC) (Schwarz, 1978), and residual information criterion (RIC) (Shi&Tsai, 2002).

All these criteria have been proposed and studied for linear regression models. These criteria involve the sum of two terms. The first is the models log likelihood, which provides a natural assessment of the quality of the fit of the model. The second is a penalty term, which is a function of the numbers of the parameters in the model (Taylor, 2008). We focus on two of the most representative and widely applied model selection criteria in this work, they are AIC and BIC.

Model selection in multiple linear regression model is probably one of the most important problems in statistics (Claeskens & Hjort, 2008).

Many authors applied and made a comparison between using AIC and BIC especially in regression model such as (Yardimci & Erar, 2002), (Cetin & Erar, 2002), (Shi & Tsai, 2002), (Kuha, 2004), and (Ward, 2008). I examined and compared the two criteria and real data set was considered as an application case. Depending on the R^2_{adj} value and the $square\ root\ of\ MSE$ value, we prefer the AIC selecting as the best model.

In sections 2 and 3 Akaike information criterion and Bayesian information criterion. In section were describe 4, the use of the two criteria is examined using a real data set. Finally, sections 5 and 6 show the results and conclusions.

2-Preliminary

Assume that we have a response variable y_i and p predictor variables $x_1, x_2, \ldots, x_p$. The linear regression model is

$$y = X\beta + \varepsilon \qquad \ldots\ldots\ldots\ldots(1)$$

Where $y_{n\times 1}$ is a vector of response variable, $X_{n\times p}$ is a matrix of predictor variables, $\beta_{p\times 1}$ is an unknown vector of coefficients, and $\varepsilon_{n\times 1}$ is an error vector with $E(\varepsilon) = 0$ and $E(\varepsilon\varepsilon') = \sigma^2 I_n$. The estimated regression model can be defined as

$$\hat{y} = X\hat{\beta} \qquad \ldots\ldots\ldots\ldots(2)$$

Where $\hat{\beta}$ is the estimator of the vector of the model parameters obtained by using the least square (LS) or maximum likelihood methods (ML).

Under the assumption that ε is multivariate normally distributed. The log likelihood for (1) is

$$log_e L(\beta, \sigma^2) = -\frac{n}{2}[log_e(2\pi) + log_e(\sigma^2)] - \frac{1}{2\sigma^2}\|y - X\beta\|^2 \quad \ldots.(3)$$

The $m.l.e$ for β in this case is the same as in LS method which is

$$\hat{\beta} = (X'X)^{-1}X'y \qquad \ldots\ldots\ldots(4)$$

and the $m.l.e$ for σ^2 is

$$\sigma^2_{m.l.e} = \frac{\|y-X\beta\|^2}{n} = \frac{sse}{n} \qquad \text{.....(5)}$$

2- Akaike Information Criterion (AIC)

The AIC procedure (Akaike, 1973) is used to evaluate how well the candidate model approximates the true model by assessing the difference between the expectations of the vector y under the true model and the candidate model using Kullback – Leibler $(K-L)$ distance (Burnham & Anderson, 2002). The criterion is ;

$$AIC = -2(log\ likelihood) + 2k \qquad \text{.....(6)}$$

where k is the number of the estimated parameters included in the model. The candidate model for which AIC is smallest represents the best approximation to the true model. AIC is an efficient criterion, and it is not a test on the model in the sense of the hypothesis testing (McQuarrie & Tsai, 1998). Hurvich and Tsai (1989) pointed out that the AIC might lead to overfitting in small samples. Thus, they added correction factor to the AIC. It called the corrected Akaike information criterion (AIC_c),which is defined as (Burnham & Anderson, 2002);

$$AIC_c = -2(log\ likelihood) + 2k + \frac{2k(k+1)}{n-k-1} \qquad \text{.....(7)}$$

where n represents the sample size.

In itself, the value of the AIC has no meaning. Two measures associated with the AIC can be used to compare models, which they have easy

interpretation, they are the $delta\ AIC\ (\Delta_i)$ and $Akaike\ weight\ (w_i)$ (Burnham & Anderson, 2002). The Δ_i is a measure of each model relative to the best model, and it is calculated as;

$$\Delta_i = AIC_i - \min AIC \qquad \ldots\ldots\ldots\ldots(8)$$

where AIC_i is the AIC value for model i, and $\min AIC$ is the AIC value of the best model. As a rule of thumb, $\Delta_i < 2$ suggests substantial evidence for the model, values between 4 and 7 indicate that the model has considerably less support, whereas $\Delta_i > 10$ indicates that the model is very unlikely (Burnham & Anderson, 2002).

$Akaike\ weights\ (w_i)$ provide another measure of the strength of evidence for each model, and represent the ratio of Δ_i values for each model relative to the whole set of R candidate models(Burnham & Anderson, 2002);

$$w_i = \frac{e^{-\Delta_i/2}}{\sum_{i=1}^{R} e^{-\Delta_i/2}} \qquad \ldots\ldots\ldots\ldots(9)$$

The interpretation of w_i is the probability that the model is the best among the whole set of candidate models (Burnham & Anderson, 2002).

3-Bayesian Information Criterion (BIC)

The second important model selection criterion that is frequently compared to AIC is the Bayesian (or Schwarz) information criterion. It was proposed by Schwarz (1978) in an evaluation criterion for models defined in terms of their posterior probability (Konishi & Kitagaw, 2008). Let $f\left(x_n \middle| \hat{\theta}\right)$

be a statistical model estimated by the maximum likelihood method. Then the BIC is given by (Konishi & Kitagaw, 2008)

$$BIC = -2\log_e f\left(x_n|\hat{\theta}\right) + k\log_e(n) \quad \text{……..(10)}$$

The model that minimizes the value of BIC can be selected as the optimal model for the data. For the linear regression model, the BIC is(Konishi & Kitagaw, 2008)

$$BIC = n\log_e(\hat{\sigma}^2_{m.l.e}) + k\ \log_e(n) \quad \text{……..(11)}$$

BIC is a consistent estimator, it assigns more weight to complex models than does AIC (McQuarrie & Tsai, 1998). Generally, BIC is interpreted as a rough approximation to the logarithm of the Bayes factor. BIC assumes equal priors on each model, so that if R models are considered, the prior on each is $1/R$ (Ward, 2008).

4-Application Case

In this section, one real data set is considered as an application case to compare the criteria.

This data include 4 predictor variables and have $n = 350$. The predictor variables are x_1 (smoking of mother), x_2(dairy products intake of mother), x_3(blood pressure of mother), and x_4(hemoglobin of mother) (AL-Mola,2007).Since the response variable did not have the normal distribution, the log transformation is taken to assure that the response will have the normal distribution. The response variable is $\log_e y_{FBLL} = natural\ logarithm\ of\ the\ fetal\ blood\ lead\ level$,

Because only 4 variables are available, the R possible candidate models is $2^4 - 1 = 15$. Table (1) shows the summary of the 15 models.

Table (1): Summary of the 15 models for log_e fetal blood lead level ranked according to the minimum values of AIC and BIC

Model	k	Rank	AIC	Δ_i	w_i	Rank	BIC	Δ_i	w_i
$x_1x_2x_3x_4$	***6***	***1***	***-259.503***	***0***	***0.725458***	2	-240.21313	1.6958	0.2583774
$x_2x_3x_4$	5	2	-257.341	2.161	0.246129	***1***	***-241.909***	***0***	***0.6032648***
$x_1x_2x_4$	5	3	-252.655	6.847	0.023643	3	-237.223	4.686	0.057937
x_2x_4	4	4	-249.449	10.05	0.004757	4	-237.874	4.035	0.0802267
$x_1x_2x_3$	5	5	-235.565	23.93	4.6E-06	8	-220.133	21.776	1.127E-05
x_2x_3	4	6	-235.687	23.81	4.89E-06	5	-224.1097	17.799	8.231E-05
$x_1x_3x_4$	5	7	-233.693	25.80	1.8E-06	9	-218.261	23.648	4.42E-06
x_1x_2	4	8	-232.191	27.31	8.51E-07	7	-220.616	21.293	1.435E-05
x_2	3	9	-231.605	27.89	6.35E-07	6	-223.888	18.021	7.367E-05
x_3x_4	4	10	-229.679	29.82	2.42E-07	10	-218.1055	23.803	4.089E-06
x_1x_4	4	11	-228.771	30.73	1.54E-07	11	-217.197	24.712	2.596E-06
x_4	3	12	-223.717	35.78	1.23E-08	12	-216.001	25.908	1.428E-06
x_1x_3	4	13	-209.288	50.21	9.05E-12	15	-197.714	44.195	1.526E-10
x_3	3	14	-208.1	51.40	5E-12	13	-200.383	41.526	5.798E-10
x_1	3	15	-207.428	52.07	3.57E-12	14	-199.7125	42.196	4.146E-10

5-Results and Comparisons

From table (1), AIC suggests that model $x_1x_2x_3x_4$ as the best approximate model. The estimated regression model is

$$\hat{y}_i = e^{3.77-0.241x_1-0.22x_2+0.4x_3-0.202x_4}$$

where the $t-values\ and\ p-values$ for the four estimated parameters are $[6.4(0.00)-2.03(0.043), -5.34(0.0001), 2.97(0.003), -5.15(0.0001)]$. The second best model is $x_2x_3x_4$, where $2 < \Delta_i = 2.161 < 3$ and the $w_i = 0.246$ which indicates that it has a 24.6% chance of being the best one among the other candidate models. The rest candidate models except model $x_1x_2x_4$ represent poor approximation since Δ_i values > 10.

The BIC suggests that the model $x_2x_3x_4$ is the best, since it has the smallest value, $\hat{y}_i = e^{0.295-0.228x_2+0.42x_3-0.193x_4}$, where the $t-values\ and\ p-values$ are $[6.9(0.0001), -5.53(0.0001), 3.15(0.0018), -4.92(0.0001)]$.

The w_i for this model is about 60.3%. The second best model is the full model $(x_1x_2x_3x_4)$, where $\Delta_i = 1.69 < 2$, and $w_i = 25.8\%$. We observe that there is a different between AIC and BIC in choosing the best model. Both of them suggest that the $full\ model$ and $x_2x_3x_4$ model be the best, the difference just in ranking.

Let us use the $videnceratio\ (ER) = \frac{w_j}{w_i}$, where model j is compared against model i (Burnham & Anderson, 2002). For AIC, the ER between $full\ model$ and $x_2x_3x_4$ is $ER = \frac{0.7254}{0.2461} = 2.974$, would indicate

that the $full\ model$ is only 2.974 more likely than $x_2x_3x_4$ model to be the best, given the rest candidate models. Whereas in BIC, the $ER = 2.334$, which indicates that the $x_2x_3x_4$ model is only 2.334 more likely than the $full\ model$.

It should be noted that there isn't any basis to say that AIC selects a better approximating model than BIC. The R^2_{adj} for the two selected models in AIC is (0.1526) and in BIC is (0.145), also the $square\ root\ of\ MSE$ is (0.685) and (0.688) respectively. Figure (1) and (2) show that they have constant variance property , figures (3) and (4) show that the error has normal distribution. So, in this case the selected model by AIC will be chosen as the best model than the selected model by BIC since the $R^2_{adj}(AIC) > R^2_{adj}(BIC)$ and the $quare\ root\ of\ MSE(AIC) < square\ root\ of\ MSE(BIC)$.

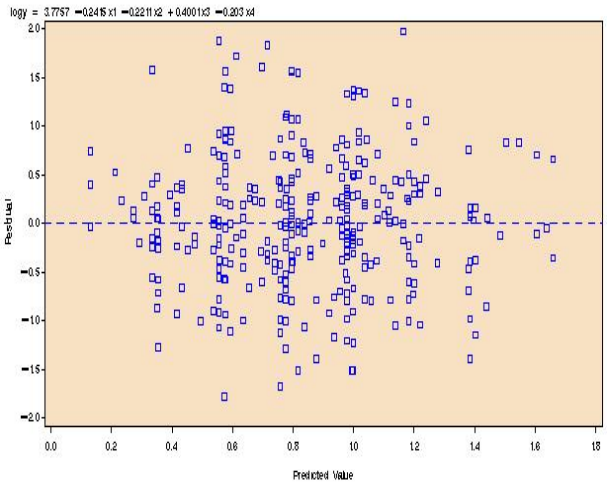
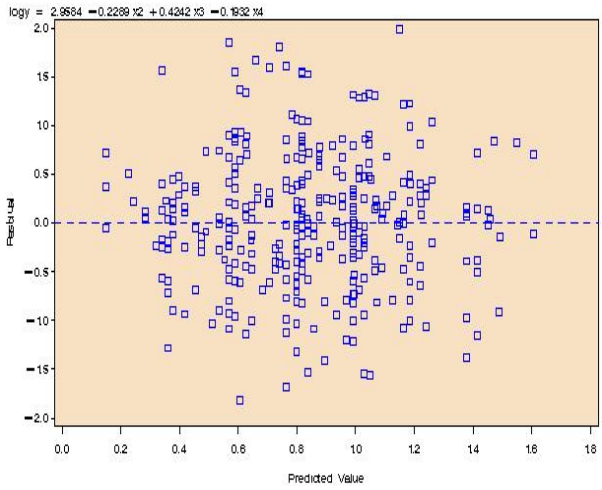


Figure (1): constant variance property of the selected model by $AIC\ (FBLL)$



Figure(2): constant variance property of the selected model by $BIC\ (FBLL)$

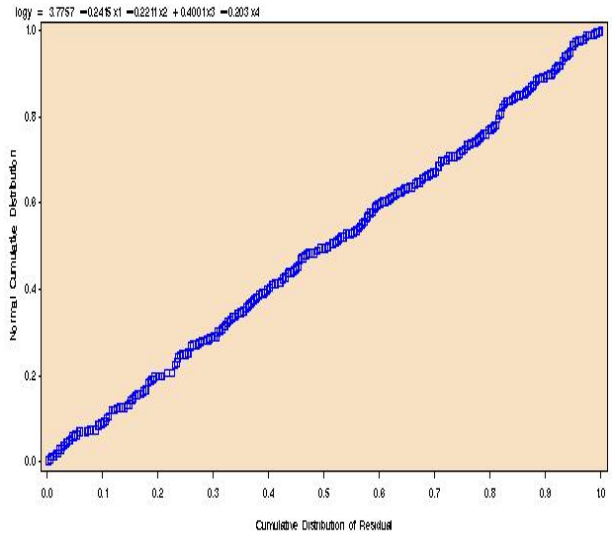
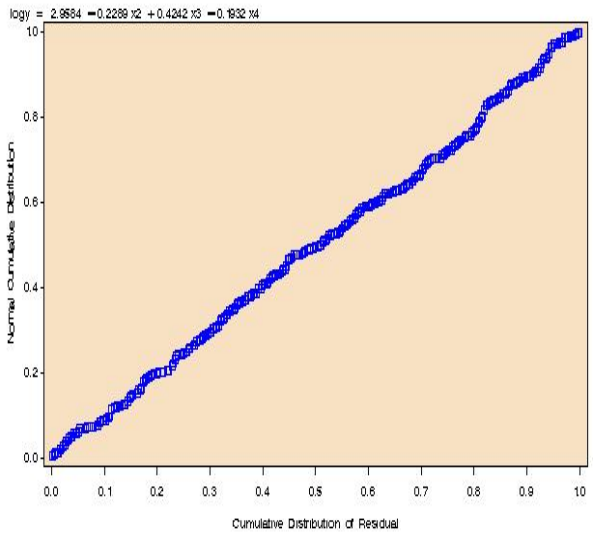


Figure (3): Normal probability plot of the error of the selected model by $AIC\ (FBLL)$



Figure(4): Normal probability plot of the error of the selected model by $BIC\ (FBLL)$

6-Conclusion

As a result of this study, it can be observed that both AIC and BIC provide a good approaches in model selection. Unfortunately, the two criteria were different from sharing the choosing best model. Depending on R^2_{adj} and the $root\ MSE$, I prefer to chose the best model that found using AIC. From figures (1) and (2) and using Breusch-Pagan test for heteroskedasticity we show that the selected models have constant variance property and figures (3) and (4) show that the errors of the selected models have normal distribution.

7-Refrences

1- Akaike, H., (1973),” Information Theory and an Extension of the Maximum Likelihood Principle”, In Second International Symposium on Information Theory.

2- Al-Mola, Z. W., 2007, "Maternal and Umbilical Cord Blood Lead Levels and Pregnancy Outcomes Hospital Based Enquiry", M.Sc thesis, College of Medicine, Mosul University.

3- Burnham, K., P. and Anderson, D., R., (2002),” Model Selection and Multimodel Inference: A Practical Information-Theoretical Approach”, 2d ed., Springer- Verlag, New York.

4- Cetin, M., C., and Erar, A., (2002),”Variable Selection with Akaike Information Criteria: A Comparative Study”, Hacettepa Journal of Mathematics and Statistics, Vol.31, pp.89-97.

5- Claeskens, G. and Hjort, N. L,(2008),” Model Selection and Model Averaging”, Cambridge University Press, Cambridge.

6- Hurvich, C., M. and Tsai, Chi, (1989), “Regression and Time Series Model Selection in Small Samples", Biometrika, 76, pp.297-307.

7- Konishi, S. and Kitagawa, G., (2008),”Information Criteria and Statistical Modeling’, Springer Science and Business media, New York.

8- Kuha, J., (2004),” AIC and BIC Comparisons of Assumptions and performance”, Sociological Methods and Research, Vol.33, No.2, pp.188-229.

9- McQuarrie, A., D., R., and Tsai,Ch.,(1998)”Regression and Time Series Model Selection” ,World Scientific Publishing Company, Singapore.

10- Schwarz, G.,(1978),”Estimating the dimension of a model”, Annals of Statistics, 6,pp.461-464.

11- Shi,P. and Tsai, Ch.,(2002),”Regression Model Selection – a Residual likelihood approach”, Journal of Royal Statistics,Soc.B,64,pp.237-252.

12- Taylor, J.,W., (2008),"Exponentially Weighted Information Criteria for Selecting among Forecasting Models", International Journal of Forecasting, Vol.24, No.3, pp.513-524.

13- Ward, E., J.,(2008),"A Review and Comparison of Four Commonly Used Bayesian and Maximum Likelihood Model Selection Tools", Ecological Modeling, 211,pp.1-10.

14- Yardimci,A. and Erar,A.,(2002),"Bayesian Variable Selection in Linear Regression and a Comparison", Hacettepa Journal of Mathematics and Statistics, Vol.31,pp.63-76.