Open Journal of Statistics, 2016, 6, 70-75
Published Online February 2016 in SciRes. http://www.scirp.org/journal/ojs
http://dx.doi.org/10.4236/ojs.2016.61009

An Alternative Approach to AIC and Mallow's C_p Statistic-Based Relative Influence Measures (RIMS) in Regression Variable Selection

Umeh Edith Uzoma[1], Obulezi Okechukwu Jeremiah[2]

[1]Department of Statistics, Nnamdi Azikiwe University, Awka, Nigeria
[2]Department of Statistics, Abia State Polytechnic, Aba, Nigeria
Email: eu.umeh@unizik.ng, profjeregana@yahoo.com

Received 2 January 2016; accepted 20 February 2016; published 23 February 2016

Abstract

Outlier detection is an important data screening type. RIM is a mechanism of outlier detection that identifies the contribution of data points in a regression model. A BIC-based RIM is essentially a technique developed in this work to simultaneously detect influential data points and select optimal predictor variables. It is an addition to the body of existing literature in this area of study to both having an alternative to the AIC and Mallow's C_p Statistic-based RIM as well as conditions of no influence, some sort of influence and perfectly single outlier data point in an entire data set which are proposed in this work. The method is implemented in R by an algorithm that iterates over all data points; deleting data points one at a time while computing BICs and selecting optimal predictors alongside RIMs. From the analyses done using evaporation data to compare the proposed method and the existing methods, the results show that the same data cases selected as having high influences by the two existing methods are also selected by the proposed method. The three methods show same performance; hence the relevance of the BIC-based RIM cannot be undermined.

Keywords

Relative Influence Measure (RIM), BIC, AIC, Mallow's C_p Statistic, Cook's Distance

1. Introduction

Model selection (variable selection) in regression has received great attention in literature in the recent times. A

How to cite this paper: Uzoma, U.E. and Jeremiah, O.O. (2016) An Alternative Approach to AIC and Mallow's C_p Statistic-Based Relative Influence Measures (RIMS) in Regression Variable Selection. *Open Journal of Statistics*, **6**, 70-75.
http://dx.doi.org/10.4236/ojs.2016.61009

large number of predictors usually are introduced at the initial stage of modeling to attenuate possible modeling biases [1]. As noted by [2], inference under models with too few parameters (variables) can be biased while with models having too many parameters (variables), there may be poor precision or identification of effects. Hence, the need for a balance between under- and over-fitted models is known as variable selection.

Influential observation is a special case of outliers. In the simplest sense, outlying or extreme values are observations which are well separated from the remainder of the data. Outliers result from either (1) the errors of measurement or (2) intrinsic variability (mean shift-inflation of variances or others) and appear either in the form of (i) change in the direction of response (Y) variable, (ii) deviation in the space of explanatory variables, deviated points in X-direction called leverage points or (iii) change in both the directions (direction of the explanatory variable(s) and the response variable). These outlying observations may involve large residuals and often have dramatic effects on the fitted least squares regression function. The influence of an individual case (data point) in a regression model can be adverse causing a significant shift (upward or downward) in the value of the parameters of a model in turn reducing the predictive power of the model. Only few papers dealing with the influence of individual data cases in regression explicitly take an initial variable selection step into account. This problem is handled by [3]-[6].

One objective of regression variable selection is to reduce the predictors to some optimal subset of the available regressors [3]. In literature, several approaches of variable selection exist, which include the stepwise deletion and subset selection. Stepwise deletion includes regression models in which the choice of predictive variables is carried out by an automatic procedure. Usually, this takes the form of a sequence of F-tests or t-tests, but other techniques are possible, such as adjusted R-square, AIC, BIC, Mallow's statistic, PRESS or false discovery rate [7].

[8] proposed the coefficient of determination ratio (CDR) which was based on the value coefficient of determination (R^2) of the linear regression model. [9] developed an outlier detection and robust selection method that combined robust least angle regression with least trimmed squares regression on jack-knife subset. When the detected outliers are removed, the standard least angle regression is applied on the cleaned data to robustly sequence the predictor variables in order of importance. [3] proposed a method called the Relative Influence Measure using the Mallow's C_p and AIC Statistics. These methods are dimensionally consistent, computationally efficient and able to identify influential case, though, failed in asymptotic consistency. [10] in comparing the BIC and AIC, stated that the AIC was not consistent. That is, as the number of observations *n* grows very large, the probability that AIC recovers a true low-dimensional model does not approach unity [11]. [12] supported same argument that the BIC has the advantage of being asymptotically consistent: as $n \to \infty$, BIC will select the correct model.

Hence, the specific objectives of this paper are to propose a relative influence measurewith an indication of whether the fit of the selected model improves or deteriorates owing to the presence of an observation (case) and that retains asymptotic consistency and hence not violating the sampling properties of the model parameters.

2. Existing Methods

2.1. Cook's Distance and the Influence Measure

Let *V* be the set of indices corresponding to the predictor variables selected from the full data set and let $\hat{y}(V)$ be the prediction vector based on the selected variables and calculated from the full data set. Also let $\hat{y}_{(-i)}(V)$ be the prediction vector based on the variables corresponding to *V*, but calculated from the full data set without case *i*. [3] noted that $\hat{y}_{(-i)}(V)$ contains prediction for case *i*, although this case is not used in calculating $\hat{y}_{(-i)}(V)$. The conditional Cook's distance for the i^{th} case is

$$\left\|\hat{y}(V)-\hat{y}_{(-i)}(V)\right\|^2 \tag{1}$$

approximately scaled. Here, $\|\bullet\|$ denotes the Euclidean norm. Repeating the variable selection using the data without case *i* as pointed out by [12], this selection yields a subset $V_{(-i)}$ of indices with $V_{(-i)}$ possibly different from *V*. Hence, the unconditional Cook's distance is

$$\left\|\hat{y}(V)-\hat{y}_{(-i)}\left(V_{(-i)}\right)\right\|^2 \tag{2}$$

approximately scaled. Since the unconditional version explicitly takes the selection effect into account [13] argued that it is preferable. As explained in the literature, a measure say M calculated from the complete data set can as well be calculated from the reduced data set as $M_{(-i)}$ and then quantify the influence of case i in terms of a function $f(\bullet)$ of M and $M_{(-i)}$. The $f(\bullet)$ has to be based on the difference in the value of the selection criterion before and after omitting case i. This difference $M - M_{(-i)}$ may then be divided by M in order to calculate the relative change in the selection criterion. As proposed by [3], the influence measure for the i^{th} case when the Cook's distance is used becomes

$$f\left(M, M_{(-i)}\right) = \frac{M - M_{(-i)}}{M} \tag{3}$$

2.2. Mallow's C_p Estimate and the Influence Measure

Let Y be an $n \times 1$ vector of response in a linear regression with corresponding $n \times p$ design matrix X of explanatory variables. A traditional model is

$$Y = X\beta + \varepsilon \tag{4}$$

where $\varepsilon \sim N\left(0, \sigma_\varepsilon^2\right)$, β and σ_ε^2 are unknown parameters. Usually, β is a $k \times 1$ vector of parameters and $k = p + 1$. $X_{n \times p}$ often contains redundant or unimportant variables. Let RSS be the usual sum of squares from the OLS fit of (4), then $\hat{\sigma}^2 = \dfrac{RSS}{n - p - 1}$ is the commonly used unbiased estimator of σ^2. Consider the subset V of $\{1, \cdots, p\}$ and let $RSS(V)$ be the residual sum of squares from the least squares fit using only the regressors corresponding to the indices in V together with an intercept. The C_p statistic corresponding model is

$$C_p(V) = \frac{RSS(V)}{\hat{\sigma}^2} + 2(v+1) - n \tag{5}$$

where v is the number of indices in V. Variable selection based on (5) entails calculating $C_p(V)$ for each subset of $\{1, \cdots, p\}$ and selecting the variables corresponding to $\hat{V}$; the subset minimizing (5). This approach is based on the fact that for a given V, $\hat{\sigma}^2 C_p(V)$ is an estimate of the expected squared error if a (multiple) linear regression function based on the variables corresponding to V is used to predict Y^*, a new (future) observation of the response random vector Y. Therefore, choosing $\hat{V}$ to minimize (5) is equivalent to selecting the variables which minimize the estimated expected prediction error. As proposed by [3], the influence measure for the i^{th} case when the C_p criterion is used becomes

$$f\left(C_p(V), C_p\left(V_{(-i)}\right)\right) = \frac{C_p(V) - C_p\left(V_{(-i)}\right)}{C_p(V)} \tag{6}$$

where $C_p\left(V_{(-i)}\right)$ is calculated as in (5) but with the i^{th} case omitted. In calculating $C_p\left(V_{(-i)}\right)$, the estimator for the error variance $\hat{\sigma}^2$ is obtained from the full data set.

2.3. The AIC Estimate and the Influence Measure

The AIC is based on the maximized log-likelihood function of the model under consideration. Suppose $\varepsilon \sim N\left(0, \sigma_\varepsilon^2\right)$, and ignoring constant terms, the maximized log-likelihood for the model corresponding to a subset V is given by $-n \log \dfrac{\{RSS(V)/n\}}{2}$. This is a non-decreasing function of the number of selected regressors. [13] therefore included a penalty termviz; $v + 2$, which equals the number of parameters which have to be estimated. Multiplying the resulting expression by −2 yields

$$\widetilde{AIC}(V) = n \log\left\{\frac{RSS(V)}{n}\right\} + 2(v+2) \tag{7}$$

See [14] for details. It is known that $AIC(V)$ does not perform when the number of parameters to be estimated is large compared to the sample size (typically cases where $40(v+2) > n$. In such a case, a modified

version of (7) should be

$$AIC(V) = n\log\left\{\frac{RSS(V)}{n}\right\} + \frac{2n(v+2)}{n-v-3} \tag{8}$$

Variable selection based on (5) and (8) calculating the criterion for each subset V of $\{1,\cdots,p\}$ and selecting the variables corresponding to the minimizing subset. This is equivalent to selecting the variables which maximize a penalized version of the maximum log-likelihood. As proposed by [3], the influence measure for the i^{th} case when the AIC criterion is used becomes

$$f\left(AIC(V), AIC\left(V_{(-i)}\right)\right) = \frac{AIC(V) - AIC\left(V_{(-i)}\right)}{AIC(V)} \tag{9}$$

The value of $AIC\left(V_{(-i)}\right)$ in (9) is obtained by using either (7) or (8) but with the i^{th} case omitted.

2.4. The Proposed BIC-Based Relative Influence Measure

A popular alternative to AIC as proposed by [15] is the Bayesian Information Criterion (BIC) [16]-[18].

The BIC is formally defined as

$$BIC = -2\ln\hat{L} + v\ln n \tag{10}$$

$\hat{L} = p\left(y|\hat{\theta}, M\right)$, where $\hat{\theta}$ are theparameter values that maximize the likelihood function. The BIC is an asymptotic result derived under the assumptions that the data distribution is in the exponential family. That is, the integral of the likelihood function $p\left(y|\hat{\theta}, M\right)$ times the prior probability distribution $p\left(\theta|M\right)$ over the parameters θ of the model M for fixed observed data y is approximated as

$$-2\ln p\left(y|M\right) \approx BIC = -2\ln\hat{L} + v\left(\ln n - \ln 2\pi\right) \tag{11}$$

Under the assumption that the model errors or disturbances are independent and identically distributed according to a normal distribution and that the boundary condition that the derivative of the log likelihood with respect to the true variance is zero, this becomes (up to an additive constant, which depends only on n and not on the model).

$$BIC = n\ln\hat{\sigma}_e^2 + v\ln n \tag{12}$$

where $\hat{\sigma}_e^2$ is the error variance. The error variance in this case is defined as

$$\hat{\sigma}_e^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2 \tag{13}$$

which is a biased estimator for the true variance. In terms of residual sum of squares, the BIC is defined thus

$$BIC(V) = n\log\left\{\frac{RSS(V)}{n}\right\} + v\log n \tag{14}$$

The BIC is an increasing function of the error variance σ_e^2 and an increasing function of v. That is, unexplained variations in the dependent variable and the number of explanatory variables increase the value of BIC. Hence, lower BIC implies either fewer explanatory variables, better fit or both.

Based on (14), the proposed influence measure for the i^{th} case when the BIC criterion is used becomes

$$f\left(BIC(V), BIC\left(V_{(-i)}\right)\right) = \frac{BIC(V) - BIC\left(V_{(-i)}\right)}{BIC(V)} \tag{15}$$

(2.15) can take the form of

$$f\left(BIC(V), BIC\left(V_{(-i)}\right)\right) = 1 - \frac{BIC\left(V_{(-i)}\right)}{BIC(V)} \tag{16}$$

Suppose $r = \dfrac{BIC\left(V_{(-i)}\right)}{BIC\left(V\right)}$, by invoking trichotomy law of real numbers, (16) can be rewritten as

$$f\left(BIC\left(V\right), BIC\left(V_{(-i)}\right)\right) = 1 - r \tag{17}$$

$$\begin{cases} \text{if } r = 1, \text{ then } i^{th} \text{ observation has no influence} \\ r = 0, \quad i^{th} \text{ observation is the only outlier in the data} \\ 0 < r < 1, \text{ then } i^{th} \text{ observation is influential} \end{cases}$$

The values of $BIC\left(V_{(-i)}\right)$ in (15, 16 and 17) are obtained by using (14) but with the i^{th} case omitted. Steel and Uys (2007) claimed that influence measure can be calculated for all selection criteria where the particular criterion is a combination of some sort of goodness-of-fit measure and a penalty function (such a penalty function usually include the number of predictors of the particular selected model as one of its components [19]. Closely evaluating (14), it is clear that $v \log n$ is a huge penalty term compared to the penalty term in (5) and (8) and hence it gives a good model fit of the data set.

3. Results

The results above **Table 1** show that the method was able to detect cases 33 and 41 as having high influence on the model given that their respective RIMs are relatively larger than others just as the AIC and Mallow's C_p Statistic-based RIM detected. The method proposed here for simultaneously detecting influential data points and variable selection, detects outliers one at a time. However, further study can be embarked upon to detect multiple influential data points all at a time while selecting optimal predictor variables.

The problems of masking and swamping were not covered in this study. Masking occurs when one outlier is not detected because of the presence of others; swamping occurs when a non-outlier is wrongly identified owing to the effect of some hidden outliers. Therefore, further studies can be carried out to detect influential outliers and simultaneously select optimal predictor variables while incorporating the solutions to problems of masking and swamping.

Table 1. BIC-based RIM for the Evaporation data contained in [20].

Case Omitted	Variables Selected	Influence Measure (BIC)	
1	1, 3, 6, 9	0.01797914	
2	1, 3, 6, 9	0.03826104	
14	1, 3, 6, 9	0.0179898	
15	1, 3, 6, 9	0.01751532	
31	1, 3, 4, 8, 9	0.03268965	
32	1, 3, 6, 9	0.02016968	
33	**6, 9, 10**	**0.05645203**	***high influence measure comparable to the Steel & Uys (2007) paper results that used the C_p and AIC
34	1, 3, 6, 9	0.01791702	
40	6, 9, 10	0.03512789	
41	**1, 3, 6, 9**	**0.06042516**	***high influence measure comparable to the Steel & Uys (2007) paper results that used the C_p and AIC
42	1, 3, 6, 9	0.02053766	
45	1, 3, 6, 9	0.01754306	
46	1, 3, 6, 9	0.01905877	

Again, because it was not intended initially to carry out a test of convergence through which we can compare the computational cost of the three methods, this work avoided the task of re-sampling which was done by Steel and Uys (2007). Meanwhile Steel and Uys (2007) did not run any test of convergence after bootstrapping rather they calculated the estimated average prediction error to substantiate their results. Hence, their additional task of re-sampling is a repetition of the results they achieved with their methods and as a result it is not necessary in this study. One can further implement these existing methods by adding a test of convergence after re-sampling.

4. Conclusion

Two things are unique about this paper namely a new approach to detecting influential outlier and then the conditions for the interpretation of the result. The later is achieved by invoking the trichotomy law of real numbers. The proposed method penalizes models hugely as the sample size becomes very large and hence has greater likelihood of choosing a better model while detecting influential data cases one at a time.

References

[1] Fan, J. and Li, R. (2001) Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties. *Journal of the American Statistical Association*, **96**, 1348-1360. http://dx.doi.org/10.1198/016214501753382273

[2] Burnham, K.P. and Anderson, D.R. (2004) Kullback-Leibler Information as a Basis for Strong Inference in Ecological Studies. *Wildlife Research*, **28**, 111-119. http://dx.doi.org/10.1071/WR99107

[3] Steel, S.J. and Uys, D.W. (2007) Variable Selection in Multiple Linear Regression: The Influence of Individual Cases. *ORiON*, **23**, 123-136. http://dx.doi.org/10.5784/23-2-52

[4] Cook, R.D. (1977) Detection of Influential Observations in Linear Regression. *Technometrics*, **19**, 15-18. http://dx.doi.org/10.1080/00401706.1977.10489493

[5] Cook, R.D. (1986) Assessment of Local Influence. *Journal of the Royal Statistical Society, Series B*, **48**, 133-169.

[6] Belsley, D.A., Kul, E. and Welsch, R.E. (1980) Regression Diagnostics. Wiley, New York. http://dx.doi.org/10.1002/0471725153

[7] Tibshirani, R.J. (1997) The LASSO Method for Variable Selection in the Cox Model. *Statistics in Medicine*, **16**, 385-395. http://dx.doi.org/10.1002/(SICI)1097-0258(19970228)16:4<385::AID-SIM380>3.0.CO;2-3

[8] Zakaria, A., Howard, N.K. and Nkansah, B.K. (2014) On the Detection of Influential Outliers in Linear Regression Analysis. *American Journal of Theoretical and Applied Statistics*, **3**, 100-106. http://dx.doi.org/10.11648/j.ajtas.20140304.14

[9] Shahriari, S., Faria, S., Goricalves, A.M. and Van Aelst, S. (2014) Outlier Detection and Robust Variable Selection for Least Angle Regression. Computational Science and Its Application-ICCSA, Vol. 8581, Springer-Verlag, New York, 512-522.

[10] Wagenmakers, E.J. and Farrell, S. (2004) AIC Model Selection Using Akaike Weights. *Psychonomic Bulletin and Review*, **11**, 192-196. http://dx.doi.org/10.3758/BF03206482

[11] Bozdogan, H. (1987) Model Selection and Akaike's Information Criterion (AIC): The General Theory and Its Analytical Extensions. *Psychometrika*, **52**, 345-370. http://dx.doi.org/10.1007/BF02294361

[12] Guetta, D. (2010) High Dimensional Variable Selection. www.columbia.edu/.../part IIIEssay.pdf

[13] Leger, C. and Altman, N. (1993) Assessing Influence in Variable Selection Problems. *Journal of the American Statistical Association*, **88**, 547-556.

[14] Akaike, H. (1973) Information Theory and an Extension of the Maximum Likelihood Principle. *The* 2^{nd} *International Symposium on Information Theory*, Budapest, 267-281.

[15] Schwarz, G. (1978) Estimating the Dimension of a Model. *Annals of Statistics*, **6**, 461-464. http://dx.doi.org/10.1214/aos/1176344136

[16] Burnham, K.P. and Anderson, D.R. (2002) Model Selection and Multi-Model Inference. Springer, New York.

[17] Hastie, T., Tibshirani, R. and Freidman, J. (2001) The Elements of Statistical Learning: Data Mining, Inference and Prediction. Springer-Verlag, New York. http://dx.doi.org/10.1007/978-0-387-21606-5

[18] Kass, R.E. and Raftery, A.E. (1995) Bayes Factors. *Journal of the American Statistical Association*, **90**, 773-795. http://dx.doi.org/10.1080/01621459.1995.10476572

[19] Kundu, D. and Murali, G. (1996) Model Selection in Linear Regression. *Computational Statistics and Data Analysis*, **22**, 461-469. http://dx.doi.org/10.1016/0167-9473(96)00008-4

[20] Freund, R.J. (1979) Multicollinearity etc.: Some "New" Examples. *Proceedings of the Statistical Computing Section*, American Statistical Association, USA, 111-112.