Performances of Model Selection Criteria When Variables are Ill Conditioned

Peter S. Karlsson¹ · Lars Behrenz¹ · Ghazi Shukur^{1,2}

Accepted: 3 April 2017 / Published online: 25 April 2017
© The Author(s) 2017

Abstract Model selection criteria are often used to find a “proper” model for the data under investigation when building models in cases in which the dependent or explained variables are assumed to be functions of several independent or explanatory variables. For this purpose, researchers have suggested using a large number of such criteria. These criteria have been shown to act differently, under the same or different conditions, when trying to select the “correct” number of explanatory variables to be included in a given model; this, unfortunately, leads to severe problems and confusion for researchers. In this paper, using Monte Carlo methods, we investigate the properties of four of the most common criteria under a number of realistic situations. These criteria are the adjusted coefficient of determination (R^2 -adj), Akaike’s information criterion (AIC), the Hannan–Quinn information criterion (HQC) and the Bayesian information criterion (BIC). The results from this investigation indicate that the HQC outperforms the BIC, the AIC and the R^2 -adj under specific circumstances. None of them perform satisfactorily, however, when the degree of multicollinearity is high, the sample sizes are small or when the fit of the model is poor (i.e., there is a low R^2). In the presence of all these factors, the criteria perform very badly and are not very useful. In these cases, the criteria are often not able to select the true model.

Keywords Information criteria · Model selection · Multicollinearity · Monte Carlo methods

✉ Peter S. Karlsson
peter.s.karlsson@lnu.se

¹ Departments of Economics and Statistics, Linnaeus University, SE 351 95 Växjö, Sweden

² Departments of Economics and Statistics, Jönköping University, 551 11 Jönköping, Sweden

1 Introduction

In many empirical contexts in which one is using multiple regression models to explain a phenomenon, the researcher very often faces a number of competing specifications of the regression model regarding how many and which regressors should be included in the final model. In such situations, it is very common for researchers to base their decisions on model selection criteria. Model selection criteria are used to evaluate a number of competing model specifications; based on the values of these criteria, the researcher makes a final choice by selecting the model for which the criteria yield the highest or smallest value. A number of model selection criteria have been developed over the years; in this paper, we are mainly interested in four information criteria that are commonly used in multiple regression contexts: the adjusted coefficient of determination (R^2 -adj), Akaike's information criterion (AIC), the Hannan–Quinn information criterion (HQC) and the Bayesian information criterion (BIC).

Applying any of these criteria would be straightforward in an ideal context (i.e., when all underlying assumptions in the regression model are met, the R^2 is close to one and the sample size is very large in relation to model complexity). Applying information criteria is however not straightforward when any of these assumptions are not met, for example if R^2 is low, or the sample size is very small. Then the properties of these information criteria can change, and some might not work as well as others in some situations. Our focus in this paper is to determine how these model selection criteria are affected by ill conditioned regressors (Frisch 1934) when one is trying to decide what model to choose and this problem arises in situations in which the explanatory variables are highly inter-correlated. It then becomes difficult to disentangle the separate effects of each explanatory variable on the explained variable. This is of great importance because regressors are non-orthogonal in most empirical applications in social science, especially in economics in which regressors is more likely to be correlated due to links between almost all types of economic activity. Should we therefore base our model selection decisions on model selection criteria when multicollinearity is present? To the authors' knowledge, the performance of model selection criteria has not yet been investigated when the regressors are ill conditioned, i.e. when multicollinearity is present.

There is therefore a need to evaluate and compare the properties of the four selected model selection criteria when the situation is not ideal, especially when various degrees of multicollinearity are present, which will be investigated in this paper.

The paper is organized as follows. In the next section, the four different model selection criteria are presented. In Sect. 3, a number of factors that can affect these properties are introduced. In Sect. 4, the design of our Monte Carlo experiment is presented. In Sect. 5, we describe the results for the rate of successfully identifying the true model for the four different criteria. The conclusions of the paper are presented in the final section.

2 Information Criteria and Model Selection in a Multiple regression Context

Model selection criteria have become popular tools to use for the specification of regression models. One reason for this is that the researcher can use model selection criteria mechanically for choosing between model alternatives, which does not require them to be nested. They are easy to apply, easy to interpret and available in almost all statistical program packages. There are a number of different model selection criteria in the literature, but in this paper, we focus on four model selection criteria that are commonly used in a multiple regression context: the (R^2 -adj), the (AIC), the (HQC) and the (BIC). Useful reviews of these criteria are to be found in [Atkinson \(1981\)](#) and [Amemiya \(1980\)](#). Our standpoint is that data have been generated by an underlying data generating process (true model) and that its regressors can be observed in the set of potential predictors available for building the model. This is in accordance to one of the two philosophies of model selection, the consistent criteria, which asymptotically identifies the correct model with probability one. Among the four criterion investigated in this paper, then BIC and HQC are consistent criterion. The other philosophy of model selection is the efficient criteria of which the goal is to find the model that best approximate the true model that is either of infinite dimensions and hence not possible to exactly model the data generating process (DGP), or the DGP is not included in the set of candidate models. For the latter philosophy, then an efficient criterion chooses the model with minimum mean square error distribution ([Shibata 1980](#)), and among the included criterion in this paper, then AIC belongs to this philosophy. But it should be noted that [Akaike \(1973\)](#) made the assumption that the true model belongs to the set of candidate models when he derived the AIC ([McQuarrie and Tsai 1998](#)). Furthermore, it should be noted that both consistency and efficiency are two asymptotic properties under which the criterion were derived, and therefore it is of interest to investigate their performance in a small sample setting which is often encountered in empirical applications in social science.

In more detail, the AIC is based on information theory using the Kullback–Leibler divergence for measuring the distance between true model and candidate model (note this measure is not a true distance function since it does not fulfill all properties that a distance measure should have), whereas the BIC is based on Bayes' factors. Although they are based on different approaches and philosophes, the criterion in a regression context are very similar to one another, and the only difference between them is in the penalty term, which can be seen below in Eqs. (2.1)–(2.3). Note that R^2 -adj (2.4) is different from the AIC, BIC and HQC. The general forms of the information criteria in a linear multiple regression model contexts are:

$$Criteria = n \ln \left(\frac{SSE}{n} \right) + (penalty\ term),$$

where n is the number of observations and $SSE = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$, $\hat{Y}_i$ is the predicted value from the estimated multiple regression and $\ln(\cdot)$ is the natural logarithm.

The evaluated model selection criteria for a multiple regression in this study are:

$$\text{AIC} = n \ln \left(\frac{\text{SSE}}{n} \right) + 2p, \quad (2.1)$$

$$\text{BIC} = n \ln \left(\frac{\text{SSE}}{n} \right) + p \ln(n), \quad (2.2)$$

$$\text{HQC} = n \ln \left(\frac{\text{SSE}}{n} \right) + p \ln(\ln(n)), \quad (2.3)$$

$$R_{adj}^2 = 1 - \frac{\text{SSE}/(n-p)}{\text{SST}/(n-1)}, \quad (2.4)$$

where p is equal to the number of parameters that are estimated in the multiple regression model.

3 The Factors that Might Affect the Performance of the Model Selection Criteria

For the model selection criteria, a number of factors can affect the rate of selecting the true model. In this paper, we focus solely on cases in which the data generating process is in the form of a linear multiple regression model. In this situation, the question arises about whether one should have fixed or stochastic regressors in the regression model. In statistics, the classical regression model has fixed regressors, and inference is conducted conditional upon the observed data. The ideal setting for a model with fixed regressors is an experimental design, but this is not a realistic model for empirical applications in many social sciences in which regressors are generated by stochastic processes over which the researcher has no control. As a result, a more realistic model would be a multiple regression model with stochastic regressors (Greene 2003; Edgerton 1996), as seen in Eq. (3.1) below. Based on this, we have specified the DGP with both stochastic error terms and stochastic regressors in order to make the evaluation similar to the situation in which these criteria are used in empirical applications in the social sciences.

Multiple regression model with stochastic regressors

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N(0, \sigma^2 \mathbf{I}) \quad (3.1)$$

where $\mathbf{y}$ is a stochastic $n \times 1$ vector, $\boldsymbol{\beta}$ is a $(k+1) \times 1$ parameter vector, $\mathbf{X} = [\mathbf{1X}_{(1)} \dots \mathbf{X}_{(k)}]$ is a matrix with dimension $n \times (k+1)$ in which we have a k different stochastic regressors in its k last columns and each regressor $\mathbf{X}_{(j)}$ is drawn from a uniform distribution defined in the interval from 0 to 100, $\boldsymbol{\varepsilon}$ is an $n \times 1$ disturbance vector with error variance equal to σ^2 . Furthermore, no correlation between any regressor and the error term is assumed, i.e. $\text{Cov}(\mathbf{X}_{(j)}; \boldsymbol{\varepsilon}) = 0 \forall j = 1, \dots, k$.

The main focus of this paper is to evaluate the performance of model selection criteria when various degrees of multicollinearity are present. This is because, for many empirical applications in, for example, economics, there is more likely to be correlation among regressors than orthogonal regressors due to links between almost all types of economic activity within an economy. As a result, the features of most

economic variables suggest that one will most likely find multicollinearity in empirical applications of multiple regression. Based on this and the fact that many economists choose their final models with model selection criteria, it is therefore of great interest to investigate the performance of these model selection criteria when various degrees of multicollinearity are present. In conjunction with multicollinearity, it is also of great interest to investigate performance under different signal-to-noise ratios, measured here by the coefficient of determination (R^2). We expect that the performances of these criteria will become poorer as R^2 decreases. In addition, because some model selection criteria are based on asymptotic results, it is also of interest to investigate the performance of these criteria for various sample sizes. Their performance should worsen as sample size decreases. The sample size, however, needs to be set in relation to model complexity. As a result, one should expect that as the number of parameters in the multiple regression model increases, the sample size should have to increase as well in order to keep the result at comparable levels. In this paper, we calculate the n/k ratio, where n is the sample size and k is the number of regressors in the true model. We then use this ratio to make the settings comparable for different values of k . Hence, by using the same n/k ratio for different values of k , we can compare results for different dimensions with the expectation that the criteria's performance will be weakened as model complexity k increases. An interesting fact is that the number of possible models increases as k increases. Let K_{max} denote the number of available regressors in a specific application; for example, $K_{max}=8$ yields 255 different models, whereas $K_{max}=10$ yields 1023 different models and $K_{max}=15$ yields 32,767 different models. As a result, the number of competing models increases at a rate much higher than linear growth, and it is therefore expected that the performance of model selection criteria will deteriorate as K_{max} increases.

The final aspect that is expected to affect the performance of the model selection criteria is the distribution of the error term in Eq. (3.1), which is assumed to be normally distributed under the classical assumptions of multiple regression models (Greene 2003). As a result, any deviation from the normal distribution is expected to affect the performance of any of the model selection criteria evaluated. This especially concerns distributions that do not retain properties possessed by the normal distribution, such as distributions with higher tail probabilities and/or non-skewness. This is reflected in the Monte Carlo experiment by generating the error terms in Eq. (3.1) from the student t-distribution and the exponential distribution.

4 The Monte Carlo Experiment

In order to evaluate the four different model selection criteria we need to perform a Monte Carlo simulation study under some different contexts. To make the evaluation similar to that in empirical applications, the data generating process is specified with both stochastic error terms and stochastic regressors, as in Eq. (4.1) below. Hence, in each replication of the Monte Carlo experiments, the error terms are drawn from either normal distributions, student t distributions or exponential distributions, and the regressors are drawn from a multivariate uniform distribution. We chose a uniform distribution for the regressors because we were initially interested in investigating

whether the observed range of these regressors would have any impact on the result of the study. The range was controlled by simply changing the upper limit of the uniform distributions from which the regressors were drawn. However, it was discovered that this did not have any impact on the result; as a result, we only present the case when the uniform distribution has the largest definition interval.

4.1 The Design of the Monte Carlo Experiment

An issue that arose in the Monte Carlo experiments was how to control the R^2 level in the simulations when the stochastic regressors were drawn from a multivariate uniform distribution. In this paper, we have adopted the following procedures: first, the level of variance for the errors generated in the model was kept fixed for each R^2 level, while the correlation structure among the stochastic regressors was altered in the Monte Carlo experiments.

The R^2 level is estimated by using one replication in each setting (except for changes in sample size in the Monte Carlo experiments) where the number of observations is very large ($n = 10,000$ observations with orthogonal regressors). The authors argue that the estimated R^2 value should then be very close to the corresponding true population R^2 because the number of observations would be very large. As a result, the level of variance used on the generated errors in the model is kept fixed for each R^2 level while the correlation structure among the stochastic regressors is altered in the experiments.

The DGP used in the Monte Carlo experiments

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} = \mathbf{c}\mathbf{d}, \quad (4.1)$$

where the only difference between Eqs. (3.1) and (4.1) is the error term which now consists of two parts where $\mathbf{c}$ is a fixed scalar that controls the variability of the stochastic part $\mathbf{d}$; more details on the specification of the error term are available in Table 3.

The correlation structure among the stochastic regressors

In each Monte Carlo experiment, the number of potential regressors is equal to $K_{\max}$ of which k regressors belong to the true DGP. Then, we generate a matrix ($\mathbf{U}$) consisting of $K_{\max}$ random variables (u_j) where each random variable is drawn from a uniform distribution such that $u_j \sim \text{Uniform}[0, 10] \quad \forall j = 1, \dots, k_{\max}$ and $\text{Cov}(u_j, u_h) = 0 \quad \forall j \neq h$. Then, the first k first columns in $\mathbf{U}$ will be used as regressors, i.e. they will be inserted in the k last columns in $\mathbf{X} = [\mathbf{1}\mathbf{X}_{(1)} \dots \mathbf{X}_{(k)}]$ in Eq. (4.1). Thus, the regressors are initially uncorrelated in the set of potential regressors, which will be altered in the Monte Carlo experiments. Our choice of the regressors' correlation structure is a Toeplitz structure that is determined by one parameter ϕ , as shown below in Eq. (4.2). By using this type of correlation structure, we can generalize the same pattern to any dimension. Note that if ϕ is set to zero, then this yields uncorrelated regressors, and for cases in which ϕ is larger than zero, it generates a correlation structure such that an individual regressor will not be correlated at the

same level as the rest of the regressors, which is a realistic setting to find in empirical applications.

$$\begin{aligned} \mathbf{P}_x = \text{Toeplitz}(\phi) &= \begin{bmatrix} \phi^0 & \phi^1 & \dots & \phi^{K_{\max}-1} \\ \phi^1 & \ddots & \ddots & \vdots \\ \dots & \ddots & \ddots & \phi^1 \\ \phi^{K_{\max}-1} & \dots & \phi^1 & \phi^0 \end{bmatrix} \\ &= \begin{bmatrix} 1 & \phi^1 & \dots & \phi^{K_{\max}-1} \\ \phi^1 & \ddots & \ddots & \vdots \\ \dots & \ddots & \ddots & \phi^1 \\ \phi^{K_{\max}-1} & \dots & \phi^1 & 1 \end{bmatrix}, \end{aligned} \quad (4.2)$$

where $\mathbf{P}$ denotes the population correlation matrix.

By post-multiplying $\mathbf{U}$ by the transpose of the matrix root from the Cholesky decomposition of $\mathbf{P}$, we transform the regressors so that they are correlated with the correlation structure in question. Then, the first k regressors of these correlated $K_{\max}$ potential regressors are used in the DGP, i.e. inserted as the k last columns in $\mathbf{X}$ in Eq. (4.1).

A motivation for the parameter setting used in the Monte Carlo experiments

As displayed in Table 1, we have chosen three main settings in which the number of regressors in the DGP is equal to 3, 5 and 10, with five extra unrelated regressors in each setting. These numbers were decided mainly by reviewing a number of empirical papers in the social sciences and noticing that it was very unusual to find a paper with more than ten continuous regressors. In this study, we have used three levels on the n/k ratio in which the largest of them was 21, which corresponds to 210 observations when we have a DGP where k is equal to 10. It is not likely that the available data sets will have large sample sizes when looking at empirical applications in the social sciences, where variables are observed over time. This is especially true for empirical applications in macroeconomics in which the normal frequency for which variables are observed is on a yearly basis. Based on this fact, we have therefore restricted the largest value on the n/k ratio to 21 in the Monte Carlo experiments. In regards to the ϕ parameter, which controls the correlation structure among the regressors, we have chosen a relatively large resolution. This is because the main focus of this paper is how correlated regressors affect the model selection criteria, which, as previously mentioned, is then investigated under different settings. The motivation for the largest values of ϕ is that they actually can occur in real applications (for example, in allocation models and demand systems in which the regressors are logarithm of prices of different close commodities that increase over time, such as prices for low-fat milk, and standard milk; or prices for different kinds of meat commodities) or in other settings in which the regressors are lagged values of themselves and/or of the dependent variable.

In regards to the beta coefficients used that are specified in Table 2, these values were chosen in an ad hoc manner. It is our belief that the choice of these values will

Table 1 Specifications of the Monte Carlo experiments

Factor	Symbol	Design
Number of explanatory variables in the DGP, i.e., in the true model	k	3; 5; 10
Maximum number of explanatory variables in the pool of potential regressors and within parentheses is the dimension of the DGP	kmax	8 (3); 10 (5); 15 (10)
n/k ratio (sample size in relation to dimension of the DGP)	n/k	3; 12; 21
Multicollinearity parameter, i.e., the phi parameter in the Toeplitz matrix (Eq. 4.2)	Φ	0; 0.1; 0.2; 0.3; 0.4; 0.5; 0.6; 0.7; 0.8; 0.9; 0.95; 0.99

Table 2 Specification of the beta coefficient in the DGP used in the Monte Carlo experiments

Dimension of the DGP	Beta coefficients in DGP for a given k (note that the first coefficient is the intercept)
k = 3	$\beta' = [10 \ 2 \ 8 \ 1]$
k = 5	$\beta' = [10 \ 2 \ 8 \ 1 \ -2 \ 3]$
k = 10	$\beta' = [10 \ 2 \ 8 \ 1 \ -2 \ 3 \ -15 \ 4 \ 0.5 \ 3 \ -7]$

Table 3 Specification of the distribution of the error term ($\mathbf{e} = c\mathbf{d}$) in the DGP

Stochastic part (d)

$$\mathbf{d} \sim N_n(\mathbf{0}, \mathbf{I})$$

$$\mathbf{d} \sim t_\nu(\mathbf{0}, \frac{\nu}{\nu-2}\mathbf{I}), \text{ where } \nu = 5$$

$$\mathbf{d}' = [d_1 \dots d_n], \text{ where each } d_i = \exp(\lambda) - \lambda^{-1} \text{ and } \lambda = 1, \text{ so that } E[d_i] = 0 \text{ and } Var(d_i) = 1 \ \forall i = 1, \dots, n$$

Fixed part (c)

$$k = 3: c = 10, 50, 100, 150$$

$$k = 5: c = 10, 50, 100, 150$$

$$k = 10: c = 10, 150, 300, 400$$

not have any noticeable impact on the result. The final parameter setting in the Monte Carlo experiments is the specification of the error term. As described in Sect. 3, it is of interest to compare the performance of the model selection criteria for situations in which the classical assumption of normal distributed errors in the DGP is met (some of the investigated criteria are based on the likelihood that in turn is based on normal distributed error terms). This is then changed to the student t distribution and the exponential distribution, as specified in Table 3 above.

We have had a much richer variety of different values on the simulation parameters, but in order to keep the results presentable in this paper, we have chosen to present

the above setting. For example, we also tried another correlation structure among the regressors based on the rho-model (ones in the diagonal and the parameter rho in the off-diagonals) in the Monte Carlo simulations without finding any major change in results; as a result, we chose to leave it out. Furthermore, the number of sample sizes reported for each value of k is reduced to three different sizes for each k in the paper. This was because the pattern could be discovered from a smaller number of values. Moreover, in the Monte Carlo simulation study, we also tried three different ranges on the regressors, but this did not yield any noticeable differences, so we chose to only present the largest range used in the simulation study.

4.2 Performance Criteria for Evaluating Model Selection performance

A number of ways to evaluate model selection criteria exist in the literature, and one common way is the rate of successfully identifying the true model. In this paper, our standpoint is that there exists a true model and the researcher is interested in interpreting the estimated model; as a result, it is of great importance to explore how the rate of successfully identifying the true model is affected under the previously mentioned settings. In some studies (e.g., [Gagné and Dayton 2002](#)), the authors also looked at the proportion of identified models that had at least 70% of the true model's regressors, treating over- and under-identified models separately. However, we argue that such an evaluation is not proper in our context because we will use non-orthogonal regressors in the Monte Carlo experiment, which is a major contrast to other studies. As a result, if the sole purpose of the model is to make predictions, not to interpret the estimated model, then one could simply include all the available regressors, bearing in mind examples from introductory statistical textbooks covering the multicollinearity problem, in which the signs of the estimated parameters changes as one introduces new regressors to the model or leaves an important regressor out. These measures are not of interest when interpretation is the sole focus of the empirical investigation. Thus, our standpoint is that one needs to identify the true model to allow for fruitful interpretation, and therefore, the only measure of interest here is the rate of successfully identifying the true model.

5 Analysis of Model Selection Criteria in a Multiple Regression setting

In this section we analyze the output from the Monte Carlo experiment along with the main dominating factors affecting the properties of the different selection criteria.

The effect of R^2

The performances of all the model selection criteria in this study deteriorate as the R^2 decreases, and this does not come as a surprise as this simply indicates that performance deteriorates as the signal-to-noise ratio decreases and the data contain less information about the true model. The most surprising result is the very weak ability to find the true data generating model in situations where R^2 is low. This is not an unrealistic situation to find in empirical applications in social sciences. Note that in

many applications in economics, especially when using data on the micro level, it possible to exhibit R^2 scores below 0.2. Thus, it is very doubtful whether researchers should base their choice of final model on model selection criteria. For the setting investigated here a thumb rule should.

The effect of k

If R^2 is close to one (i.e., in our settings, approximately 0.99), and given that we hold the n/k ratio constant, then performance is improved as we let k increase. However, this effect is reversed for lower levels of R^2 , and at these levels are more realistic settings to find in empirical studies, our conclusion is that the performance of a model selection criterion deteriorates as the complexity of the model increases. For example, if R^2 is approximately 0.95, then an increase of the dimension of the true model from $k=5$ to $k=10$ causes the rate of successfully identifying the true model to be one fourth of the former (A3.3.2 vs A2.3.2). This can mainly be explained by the increase in the number of competing models for selecting the true model, which increases at a much higher rate than linear growth. For instance, $K_{max}=10$ yields 1023 different models to be evaluated compared to the case in which $K_{max}=15$, which yields 32,767 different models to be evaluated.

The effect of multicollinearity

Multicollinearity has a clear impact on the model selection criteria's ability to find the true model. The most interesting aspect here is how the impact of multicollinearity on the model selection criteria changes as R^2 decreases. As a result, the tangible effect of multicollinearity is pushed downward such that the rate of successfully identifying the true model is weakened as Φ increases. Thus, in empirical applications in social science where we would expect to find low R^2 , the impact of multicollinearity becomes more severe.

The effect of n/k

From the simulation result, given the case in which we have no multicollinearity, it is obvious that BIC suffers the most from small samples. For example, as we can see in "Appendix A2", Figure A2.1.1, when $n/k=3$ (which is here considered a small sample), the other criteria outperform BIC. However, as the n/k ratio increases to 12, which is displayed in Figure A2.2.1, the BIC outperforms the other model selection criteria. The latter result is enhanced when the n/k ratio increases to 21 in Figure A2.3.1.

The effect of distribution of the noise part

In regards to the noise part of the DGP, we are mainly interested in seeing how a violation of normal distributed errors can affect the results. This is done in two ways. The first is performed by using errors that are generated from the t-distribution, which has higher probability in the tails than the normal distribution. The second is performed by using errors that are generated from an exponential distribution, which we then subtract from the expected value so that the errors in the Monte Carlo experiments have an expected value equal to zero but with a skewed distribution.

To our surprise, this does not seem to have any major impact on the performance of the model selection criteria. Although there is some effect, the magnitude is very low.

The importance of each factor in comparison to each other

When it comes to the effect of all five factors, then these are related to each other such as if the n/k ratio is very large (say a n/k ratio larger than 1000) then one achieve the same performance for a much lower R^2 level compared to small sample case (say a n/k ratio smaller than 21) with a larger R^2 . In addition we also have case in which the performance of the model selection criteria is less affected by multicollinearity for very large samples. In other word, to find a noticeable effect on the criteria we need a much higher degree of multicollinearity in large samples. Hence, this complicates disentangle the factors' individual effects. From the simulation result in this paper we have found that among the five factors, it is the n/k ratio that is most important since a large sample size mitigates the problem of multicollinearity and simultaneously improves the performance of the investigated criteria with or without multicollinearity present.

Further, if we have situation where the n/k ratio is smaller than say 22, then the second most important criteria is the R^2 level, since this affects both the performance of model selection criteria with or without multicollinearity present, hence it affects maximum power for each model selection criterion (this normally happens when $\phi = 0$) and also affects at which point (ϕ) for which the power starts to diminish.

The third important factor is the degree of multicollinearity, which showed to always have an negative effect on model selection criteria for very high level of correlation structure among the regressors, which here corresponds to a ϕ around 0.99. The least important factor, but not unimportant, is the model complexity which corresponds to the number of regressors in the model (k), but one should keep in mind that with increasing k , then the power for detecting the true model decreases for each investigated model selection criterion.

A comparison of the relative performances among the four model selection criteria

Based on the results from the Monte Carlo experiments displayed in "Appendix A1–A3", R^2 -adj almost always has the lowest rate of successfully identifying the true model; as a result, it is relatively unaffected by any changes in the settings. This is especially clear for changes in the correlation structure among the regressors because the curves in the figures in "Appendix A1–A3" are very flat for R^2 -adj. A more notable difference is observed for the three other criteria. If R^2 is approximately 0.99, then the BIC is always the best out of the four criteria, but when R^2 decreases, the result shifts so that the HQC, and in some cases, the AIC, will outperform the BIC for large multicollinearity, i.e., large ϕ values, although the HQC is still best. This result is enhanced as R^2 decreases, where the above changes in performance are then shifted downwards with respect to the level of multicollinearity for which the HQC and the AIC outperform the BIC (see for example Figures A2.2.1-4 and A2.3.1-4 in "Appendix A2"), and in some settings, the BIC will not achieve the highest rate of successfully identifying the true model for any levels of multicollinearity (see for example Figure A2.3.4 in "Appendix A2"). Hence, the HQC is a better alternative if the R^2 is less than 0.9 and there are middling or higher levels of multicollinearity, which in this paper

is reflected by a ϕ value larger than 0.4. In the light of different sample sizes, the same prevailing pattern is displayed in the results from the Monte Carlo experiments, in which BIC suffers most from small samples. Thus, in light of the results discussed here, we recommend that one always consider the HQC in empirical applications in the social sciences because it is very likely to have a situation in which R^2 is low and some degree of multicollinearity is present, but if R^2 is above 0.9 and sample size is large then it is advised to use BIC. Further, we advise not to use both AIC and R^2_{adj} in neither situations due to that they are always outperformed by the other two model selection criteria.

Some thumb rules for the practitioner

To give guidance to the practitioner we offer some thumb rules for when the investigated model selection criteria can be used in a multiple regression application without any seriously affected performance with respect to the five factors.

- To use model selection criteria in a setting with potential regressors less or equal to ten, then the following should be satisfied: n/k ratio ≥ 6 ; R^2 level ≥ 0.9 ; degree of multicollinearity measured by $\phi < 0.8$ and $k \leq 10$.
- To use model selection criteria in a setting with potential regressors more than ten, but less or equal to fifteen, then the following should be satisfied: n/k ratio ≥ 6 ; R^2 level ≥ 0.95 ; degree of multicollinearity measured by $\phi < 0.8$ and $k \leq 15$.

6 Conclusions

A number of studies have evaluated model selection criteria in which the settings involved orthogonal regressors and very large sample sizes. In this paper, we have relaxed those unrealistic settings, which will most likely not reflect an empirical application where these criteria are used for selecting the final model. Our opinion is that this evaluation of model selection criteria should as much as possible reflect settings that are often experienced by practitioners in social sciences. We have therefore focused on designing Monte Carlo experiments where the variables exhibit various degrees of multicollinearity in relatively small samples. The results from this investigation indicate that the HQC outperforms the BIC, the AIC and the R^2 -adj under specific conditions when R^2 is less than 0.9 and the degree of multicollinearity is larger than 0.4. Furthermore, to the authors' surprise, the performances of these criteria were not as good as expected, which was particularly true under the conditions that reflect common empirical applications in social science. Thus, based on these results, our final recommendation is that practitioners should not base their model building decisions only on the model selection criteria. These criteria should have only the role of guidance without having the final decision. After this guidance, practitioners are advised to conduct a battery of diagnostic tests in order to evaluate the adequacy of the selected model(s).

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

Appendix A1

In this setting the true model has 3 regressors and the sample sizes are $n = 10, 36, 63$ starting with 10 in the left-most column.

Standard Normal distribution

Student t-distribution with $df=5$

Exponential distribution with $\lambda=1$

Appendix A2

In this setting the true model has 5 regressors and the sample sizes are $n = 15, 60, 105$ starting with 15 in the left-most column.

Standard Normal distribution

Student t-distribution with df=5

Exponential distribution with $\lambda=1$

Appendix A3

In this setting the true model has 10 regressors and the sample sizes are $n = 30, 120, 210$ starting with 30 in the left-most column.

Standard Normal distribution

Student t-distribution with $df=5$

Exponential distribution with $\lambda=1$

References

- Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csake (Eds.), *Second international symposium on information theory* (pp. 267–281). Budapest: Akademiai Kiado.
- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19, 716–723.
- Akaike, H. (1978a). A new look at the Bayes procedure. *Biometrika*, 65, 53–59.
- Akaike, H. (1978b). A Bayesian analysis of the minimum AIC procedure. *Annals of the Institute of Statistical Mathematics*, 30, 9–14.
- Amemiya, T. (1980). Selection of regressors. *International Economic Review*, 21, 331–354.
- Atkinson, A. C. (1981). Likelihood ratios, posterior odds and information criteria. *Journal of Econometrics*, 16, 15–20.

- Bai, Z. D., Krishnaiah, P. R., Sambamoorthi, N., & Zhao, L. C. (1992). Model selection for log-linear model. *Sankhya B*, 54, 200–219.
- Bozdogan, H. (1987). Model-selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, 52, 345–370.
- Burnham, K. P., & Anderson, D. R. (1998). *Model selection and inference: A practical information—Theoretic approach*. New York: Springer.
- Draper, N. R., & Smith, H. (1998). *Applied regression analysis* (3rd ed.). New York: Wiley.
- Edgerton, D. L. (1996). Should stochastic or non-stochastic exogenous variables be used in Monte Carlo experiments? *Economics Letters*, 53(2), 153–159.
- Frisch, R. (1934). *Statistical confluence analysis by means of complete regression systems*. Oslo: University Institute of Economics, Publication 5.
- Gagné, P., & Dayton, C. M. (2002). Best Regression model using information criteria. *Journal of Modern Applied Statistical Methods*, 1(2), 479–488.
- Greene, W. H. (2003). *Econometric analysis*. Upper Saddle River: Pearson Education.
- Hannan, E. J., & Quinn, B. G. (1979). The determination of the order of an autoregression. *Journal of the Royal Statistical Society: Series B*, 41, 190–195.
- Hurvich, C. M., & Tsai, C.-L. (1991). Bias of the corrected AIC criterion for underfitted regression and time series models. *Biometrika*, 78, 499–509.
- McQuarrie, A. D. R., & Tsai, C.-L. (1998). *Regression and time series model selection*. Singapore: World Scientific.
- Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6, 461–464.
- Shibata, R. (1980). Asymptotically efficient selection of the order of the model for estimating parameters of a linear process. *Annals of Statistics*, 8, 147–164.