Model Selection with Information Complexity in Multiple Linear Regression Modeling

Hongwei Yang
University of Kentucky

Hamparsum Bozdogan
University of Tennessee

This paper aims to introduce to applied researchers a new family of information model selection criteria in multiple linear regression models. These criteria are known as information complexity (*ICOMP*) criteria. The paper provides supportive evidence under the R language to show the effectiveness of *ICOMP* and its tendency to outperform some other traditional criteria: *AIC*, *SBC*, etc. This paper also creates a framework on which to base future work in applying *ICOMP* to more general regression modeling problems in R.

The selection of an appropriate model from a potentially large class of candidate models is an issue that is central to regression, time series modeling, and generalized linear models (McQuarrie & Tsai, 1998). In multiple linear regression, statistical model evaluation and selection involves evaluating a pool of subsets of predictors and selecting the best subset that predicts the response with sufficient accuracy from predictor variables that can be measured cheaply (Miller, 2002). Given a large number of predictor variables, the hope is to identify a small subset of them that gives adequate prediction accuracy for a reasonable cost of measurement. On the other hand, it is well known that, for multiple linear regression models fitted using least squares, the variance of the predicted response values increases monotonically with the number of predictor variables used in the prediction equation, and this increased prediction variability is traded off against reduced prediction bias. The question of how this trade-off should be handled is a critical problem in this field of subset selection in multiple linear regression modeling.

The problem of selecting the best regression subset is not trivial particularly when there are a large number of potential predictors. This is so because, usually without a precise knowledge of the relationship between the response and the predictors, researchers have to find a way of developing, validating, evaluating and selecting regression models and the increase in the number of predictors complicates the process. In addition to theoretical considerations, researchers also rely on data-adaptive approaches to regression model selection. Hypothesis-test-based stepwise regression is one of many data-adaptive model selection techniques that are commonly used today, which adds and/or removes predictors based on partial F or t statistics with arbitrarily set probabilities of entry and removal after controlling the contributions of other predictors, if any, already in the model. However, hypothesis-test-based stepwise regression has known problems. First, there is no guarantee that the final model from stepwise regression is optimal in any specified sense (Tamhane & Dunlop, 1999). Stepwise procedures can sometimes err by identifying a suboptimal regression model as “best” (Kutner, Nachtsheim, & Neter, 2004). Second, the probabilities for entry and removal of predictors are arbitrarily set, so plenty of subjectivity exists in the model search process.

As an alternative to model selection via hypothesis testing, information model selection criteria are recommended for comparing and evaluating competing regression and other statistical models (Burnham & Anderson, 2002). As is compared with the usual methods of hypothesis testing, the use of information criteria in model selection has had a much shorter exposure in statistics. Information criteria belong to the group of relative fit criteria which select the best model from a pool of models that we have specified. Relying on information criteria, we can identify the model that appears to be the best among its competitors (Skrondal & Rabe-Hesketh, 2004), and the model is the best in the sense of optimizing information criteria. So, a critical task for users of information criteria is to set up more appropriate competing models by making use of knowledge regarding the object (Konishi & Kitagawa, 2010). Information criteria can be used with many data-adaptive automatic model selection algorithms including stepwise regression, all-possible-subset regression, and genetic algorithms (Bozdogan, 2004).

There are two approaches to information model selection criteria: 1) Information- theoretic approach, and 2) Bayesian approach (Ando, 2010; Konishi & Kitagawa, 2010). The former approach includes Akaike's Information Criterion or *AIC* (Akaike, 1973; 1987), Consistent Akaike's Information Criteria or *CAIC* (Bozdogan, 1987), etc. The latter approach includes Schwartz Bayesian Criterion or *SBC* (Schwartz, 1978), etc. The *AIC*-type criteria and their variants are constructed as estimators of the Kullback-Leibler

(K-L) information (Kullback & Leibler, 1951) between a statistic's model and the true distribution generating the data. In contrast, the Bayes approach for selecting a model is to choose the model with the largest posterior probability among a set of competing models. Information criteria usually assess how badly a model fits the data while adjusting for the level of complexity of a model (i.e., the number of free parameters, interdependency of parameter estimates, etc.) (Bozdogan, 2004), so the best approximating model is selected as the one that minimizes the criterion. Due to the availability of multiple criteria, matching appropriate selection criteria to a given problem or data set has received much attention in the literature (McQuarrie & Tsai, 1998).

Many information criteria appear similar in form to *AIC* because they all take the form of 1) a penalized log likelihood: a badness/lack of fit term, or a negative log likelihood term, plus 2) a penalty term (Sclove, 1987). For example, the formula for *AIC* is (-2) times the maximized log likelihood function plus 2 times the number of free parameters, with the former term describing lack of fit and the latter penalizing the number of free parameters in the model. In *AIC*, a measure of model complexity is comprised of the number of free parameters (Bozdogan, 2004). Like *AIC*, many other information criteria also contain two terms that serve similar purposes. They usually use the same lack of fit term as *AIC*, but differ in how to penalize model complexity.

Bozdogan's Information Complexity Criterion or *ICOMP* is a relatively new family of model selection criterion (Bozdogan, 2004). Like *AIC* and other criteria, *ICOMP* uses (-2) times the maximized log likelihood to measure the lack of fit of the model. On the other hand, the complexity of the model is measured based on a generalization of the covariance complexity index introduced by Van Emden (1971). Unlike *AIC*, which defines model complexity as number of free parameters, *ICOMP* measures this concept with both the number of free model parameters and the interdependency of parameter estimates. According to Bozdogan (2004), Konishi and Kitagawa (2010), and Mulaik (2009), a generic formula of *ICOMP* is:

$$ICOMP = -2logL(\widehat{\boldsymbol{\theta}}) + 2C(\hat{\Sigma}_M),$$

where $\widehat{\boldsymbol{\theta}}$ is the maximum likelihood estimate of the parameter vector under the model whose covariance matrix is denoted by $\hat{\Sigma}_M = Est.Cov(\widehat{\boldsymbol{\theta}})$, and where C represents a real-valued complexity measure of $\hat{\Sigma}_M$. Usually two types of C measures exist denoted by $C_1(*)$ and $C_{1F}(*)$, respectively. Both of them are designed to *transform* a covariance matrix into a scalar value, which is then used to measure model complexity. The covariance matrix inside the parenthesis of the two complexity measures is called the inverse Fisher Information Matrix (*IFIM*). Bozdogan (2004) developed several *IFIM*s to handle different modeling conditions (e.g., mis-specification resistant vs. otherwise). Loosely speaking, when applying a complexity measure (either by $C_1(*)$ or $C_{1F}(*)$) to *IFIM*, the model complexity part of *ICOMP* is created, which is combined with the lack of fit part to construct an *ICOMP* criterion.

Although the use of *AIC*, *CAIC*, and *SBC* in regression analysis is well documented in the literature (Burnham & Anderson, 2002; Claeskens & Hjort, 2008; McQuarrie & Tsai, 1998; Miller 2002) partially because they have been made readily available by major statistics programs, the research on applying *ICOMP* to regression modeling is very limited. Bozdogan and Haughton (1998) examined the performance of six *ICOMP* criteria using only the $C_1(*)$ measure of complexity in its early stage of development. Since then, more *ICOMP* criteria have been created that have extended the way model complexity is measured. So, this paper revisits the topic of *ICOMP*-based regression model selection using more recent *ICOMP* criteria that approach model complexity from beyond the $C_1(*)$ perspective to include the $C_{1F}(*)$ measure. Also, prior implementations of *ICOMP* have used MATLAB®, a program preferred mainly by engineers/mathematicians. Coding *ICOMP* in R is desired because R is more readily available and is better accepted in non-engineering/non-math fields

In sum, this study aims to achieve the following: 1) familiarizing applied researchers using regression with *ICOMP*, 2) comparing the performance of *ICOMP* in regression with that of other criteria, and 3) creating *ICOMP* routines in R (available upon request from the authors) to present the criteria in a better accepted environment.

Before continuing, some key general issues in model selection are briefly discussed:

Best approximating model: This is the model in the pool of candidate models that is "closet" to the true model (Bozdogan & Haughton, 1998). The objective of modeling is to obtain a "good" model, rather than the true model (Konishi & Kitagawa, 2010). This true model, which in the background generated the data, might be very complex and almost always unknown. For working with the data, it may be more practical to work instead with a simpler, but almost-as-good model, and, hence, the best approximating model. A true model can be defined explicitly only in some special situations such as in computer simulations. In this paper, the *good* model and the *best* model are both used to refer to the best approximating model.

Consistency: A model selection criterion is considered to be consistent if the probability of selecting the best approximating model converges to one as the sample size goes to infinity. Because an infinitely large sample is impossible to obtain, the paper focuses on the behavior of *ICOMP* criteria as the sample size is finite and keeps increasing. If the performance of *ICOMP* improves as sample size increases, it provides supportive evidence of *ICOMP* being consistent.

Overfitting and underfitting: Statistical modeling has to balance simplicity (i.e., fewer parameters in a model, lower variability in the predicted response, but with more modeling bias) against complexity (i.e., more parameters in a model, higher variability in the predicted response, but with smaller modeling bias). Statistical model selection criteria have to seek a proper balance between overfitting (i.e., a model with too many parameters, more than actually needed) and underfitting (i.e., a model with too few parameters, not capturing the right signal) (Claeskens & Hjort, 2008). A criterion underfits/overfits a model when it selects a model that contains fewer/more parameters than does the best approximating model (Bozdogan & Haughton, 1998).

Theoretical Framework

A multiple linear regression model under normality is defined by:

$$\underset{(n\text{x}1)}{\mathbf{y}} = \underset{(n\text{x}q)}{\mathbf{X}}\underset{(q\text{x}1)}{\boldsymbol{\beta}} + \underset{(n\text{x}1)}{\boldsymbol{\varepsilon}} \tag{1}$$

where $\mathbf{y}$ is an (nx1) vector of observed values of the response variable, $\mathbf{X}$ is an (nxq) full rank matrix representing n observations with each one measured on k variables and $q = k + 1$, $\boldsymbol{\beta}$ is a (qx1) matrix of unknown regression coefficients, and $\boldsymbol{\varepsilon}$ is an (nx1) vector of i.i.d. random errors. Further, suppose $\mathbf{y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$ and $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I})$ with σ^2 being the unknown variance of random errors.

To evaluate how well an estimated regression model under Equation (1) fits the observed data, *ICOMP* criteria are presented below. *ICOMP* criteria share the same badness/lack of fit term as *AIC*, *CAIC*, etc., which equals (-2) times the maximized log likelihood function, but *ICOMP* criteria measure model complexity differently.

Badness/Lack of Fit Term of ICOMP

Given the multiple regression model in Equation (1) the maximum likelihood estimates or MLE's of $\boldsymbol{\beta}$ and σ^2 are given by:

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{X}'\mathbf{y}\,, \tag{2}$$

$$\hat{\sigma}^2 = \frac{1}{n}(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})'\,(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) \tag{3}$$

Hence, the maximized log likelihood function is

$$logL(\widehat{\boldsymbol{\beta}}, \hat{\sigma}^2) = -\frac{1}{2}\,n\,log(2\pi) - \frac{n}{2}\,log(\hat{\sigma}^2) - \frac{1}{2}\,n \tag{4}$$

The badness/lack of fit part of *ICOMP* is thus:

$$-2logL(\widehat{\boldsymbol{\beta}}, \hat{\sigma}^2) = n\,log(2\pi) + n\,log(\hat{\sigma}^2) + n \tag{5}$$

Model Complexity Term of ICOMP

The model complexity term of *ICOMP* takes various forms, so various versions of *ICOMP* can be defined. Basically, this term is defined as the complexity of inverse the Fisher Information Matrix or *IFIM* (Bozdogan, 2004). There are two ways to measure the complexity of a matrix, namely $C_1(*)$ and

$C_{1F}(*)$. There are also two different forms of *IFIM*, namely *IFIM* and mis-specified *IFIM*. Presented next are three approaches to model complexity in *ICOMP* with different combinations of 1) complexity measure ($C_1(*)$ vs. $C_{1F}(*)$) and 2) *IFIM* (*IFIM* vs. mis-specified *IFIM*).

The first approach to *ICOMP* complexity takes the $C_1(*)$ complexity of $\mathbf{F}^{-1}$, denoted by $C_1(\mathbf{F}^{-1})$, where $\mathbf{F}^{-1}$ is the estimated inverse Fisher Information Matrix of the regression model given by

$$\mathbf{F}^{-1} = Est.Cov(\widehat{\boldsymbol{\beta}},\sigma^2) = \begin{bmatrix} \hat{\sigma}^2\,(\mathbf{X}'\mathbf{X})^{-1} & \mathbf{0} \\ \mathbf{0}' & \frac{2\hat{\sigma}^4}{n} \end{bmatrix}$$

Now invoking the complexity measure the C_1 to $\mathbf{F}^{-1}$ we have the scalar value of its complexity given by:

$$C_1(\mathbf{F}^{-1}) = \frac{s}{2}log\left[\frac{tr(\mathbf{F}^{-1})}{s}\right] - \frac{1}{2}log\left|\mathbf{F}^{-1}\right|, \tag{6}$$

where $s = dim(\mathbf{F}^{-1}) = rank(\mathbf{F}^{-1})$ (7)

For the regression model in Equation (1), $s = dim(\mathbf{F}^{-1}) = rank(\mathbf{F}^{-1}) = q$. Further suppose the eigenvalues of $Est.Cov(\widehat{\boldsymbol{\beta}},\sigma^2)$ are $\lambda_1, \lambda_2, \ldots, \lambda_q$. Therefore,

$$\begin{aligned} C_1(\mathbf{F}^{-1}) &= \frac{q}{2}log\left[\frac{tr(\mathbf{F}^{-1})}{q}\right] - \frac{1}{2}log\left|\mathbf{F}^{-1}\right| \\ &= \frac{q}{2}log\left[\frac{\sum_{j=1}^{q}\lambda_j}{q}\right] - \frac{1}{2}log\left|\prod_{j=1}^{q}\lambda_j\right| \\ &= \frac{q}{2}log\left[\frac{\bar{\lambda}_a}{\bar{\lambda}_g}\right] \end{aligned}$$

where $\bar{\lambda}_a = \frac{1}{q}\sum_{j=1}^{q}\lambda_j$ is the arithmetic mean of the eigenvalues of $\mathbf{F}^{-1}$ and $\bar{\lambda}_g = \left[\prod_{j=1}^{q}\lambda_j\right]^{\frac{1}{q}}$ is the corresponding geometric mean.

The second approach to *ICOMP* complexity takes the $C_{1F}(*)$ complexity of $\mathbf{F}^{-1}$ denoted by $C_{1F}(\mathbf{F}^{-1})$. This second complexity measure is used to avoid the problematic situation where $C_1(\mathbf{F}^{-1})$ becomes zero; it measures the relative variation in the eigenvalues and is given by:

$$C_{1F}(\mathbf{F}^{-1}) = \frac{1}{4\bar{\lambda}_a^2}\sum_{j=1}^{q}(\lambda_j - \bar{\lambda}_a)^2 \;. \tag{8}$$

The third approach to *ICOMP* complexity uses both $\mathbf{F}^{-1}$ and its outer product form $\mathbf{R}$. For the regression model in Equation (1), the estimated outer product form of the Fisher Information Matrix is given by:

$$\mathbf{R} = \begin{bmatrix} \frac{1}{n\hat{\sigma}^4}\mathbf{X}'\mathbf{D}^2\mathbf{X} & \mathbf{X}'\mathbf{1}\frac{Sk}{2\hat{\sigma}^3} \\ (\mathbf{X}'\mathbf{1}\frac{Sk}{2\hat{\sigma}^3})' & \frac{(Kt-1)}{4\hat{\sigma}^4} \end{bmatrix}, \tag{9}$$

where $\mathbf{D}^2 = \text{diag}[\hat{\epsilon}_1^2, \hat{\epsilon}_2^2, , \ldots, \hat{\epsilon}_n^2]$ with , $i = 1,2, \ldots, n$, being squared residuals from the fitted regression model, *Sk* is the estimated residual skewness, *Kt* the estimated residual kurtosis, and $\mathbf{1}$ is an (nx1) vector of ones. Formulas for *Sk* and *Kt* are respectively given by:

$$Sk = \frac{\frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i^3}{\hat{\sigma}^3} \quad , \qquad \text{and} \qquad Kt = \frac{\frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i^4}{\hat{\sigma}^4}$$

With $\mathbf{F}^{-1}$ and $\mathbf{R}$, the mis-specified version of the estimated *IFIM* can be defined:

$$Est.Cov(\widehat{\boldsymbol{\beta}},\sigma^2)_{Mis} = \mathbf{F}^{-1}\mathbf{R}\mathbf{F}^{-1}$$

Therefore, the third approach to *ICOMP* complexity takes the $C_1(*)$ complexity of $\mathbf{F}^{-1}\mathbf{R}\mathbf{F}^{-1}$ denoted by $C_1(\mathbf{F}^{-1}\mathbf{R}\mathbf{F}^{-1})$. This version of *ICOMP* provides a protection against model mis-specification (Bozdogan, 2004).

ICOMP and Non-ICOMP Criteria

Based on the information presented previously, formulas for several *ICOMP* criteria are given below, along with formulas for several non-*ICOMP* criteria.

$$AIC = nlog(2\pi) + nlog(\hat{\sigma}^2) + n + 2(k+1) \tag{10}$$

$$AIC_C = nlog(2\pi) + nlog(\hat{\sigma}^2) + n + 2\left[\frac{n(k+1)}{n-k-2}\right] \tag{11}$$

$$CAIC = nlog(2\pi) + nlog(\hat{\sigma}^2) + n + [log(n)+1]k \tag{12}$$

$$SBC = nlog(2\pi) + nlog(\hat{\sigma}^2) + n + [log(n)]k \tag{13}$$

$$\begin{aligned} ICOMP_{C1} &= nlog(2\pi) + nlog(\hat{\sigma}^2) + n + 2C_1(\mathbf{F}^{-1}) \\ &= nlog(2\pi) + nlog(\hat{\sigma}^2) + n + 2\left[\frac{q}{2}log\left(\frac{\bar{\lambda}_a}{\bar{\lambda}_g}\right)\right] \end{aligned} \tag{14}$$

$$\begin{aligned} ICOMP_{C1F} &= nlog(2\pi) + nlog(\hat{\sigma}^2) + n + 2C_{1F}(\mathbf{F}^{-1}) \\ &= nlog(2\pi) + nlog(\hat{\sigma}^2) + n + 2\left[\frac{1}{4\bar{\lambda}_a^2}\sum_{j=1}^{q}(\lambda_j - \bar{\lambda}_a)^2\right] \end{aligned} \tag{15}$$

Finally, according to the mis-specified *IFIM* or $Est.Cov(\hat{\boldsymbol{\beta}},\sigma^2)_{Mis}$, the mis-specified *ICOMP* can be defined by:

$$\begin{aligned} ICOMP_{Mis} &= nlog(2\pi) + nlog(\hat{\sigma}^2) + n + 2C_1\left[Est.Cov(\hat{\boldsymbol{\beta}},\hat{\sigma}^2)_{Mis}\right] \\ &= nlog(2\pi) + nlog(\hat{\sigma}^2) + n + 2C_1[\mathbf{F}^{-1}\mathbf{R}\mathbf{F}^{-1}] \end{aligned} \tag{16}$$

Further analyses are based on the seven criteria presented above. Data sources and the simulation protocol are detailed in the next section.

Monte Carlo Simulation Examples

Simulation Protocol

Determining the effectiveness of an information criterion involves evaluating cumulative model selection results from repeated random sampling: running the simulation repeatedly and finding the number of times that the best approximating model is identified by each criterion. Data sets used in the study are generated using Monte Carlo methods (Bozdogan & Haughton, 1998). The study simulates data sets where the true regression model has five predictors, namely $\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{x}_3$, $\mathbf{x}_4$, and $\mathbf{x}_5$. And the analysis is performed respectively for three sample sizes, namely n = 50, 100, and 1000.

Suppose $z_i \sim \mathcal{N}(0,1)$, i = 1, 2, . . ., 6. The following simulation protocol is used:

$$\mathbf{x}_i = \sqrt{1-\alpha_1^2}z_i + \alpha_1 z_6 \quad \text{when } i = 1, 2, 3$$

$$\mathbf{x}_i = \sqrt{1-\alpha_2^2}z_i + \alpha_2 z_6 \quad \text{when } i = 4, 5.$$

α_1 and α_2 are parameters controlling the degree of multicollinearity, and α_1^2 = 0.3 and α_2^2 = 0.5 to yield a reasonable covariance structure for $\mathbf{X} = \{\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3,\mathbf{x}_4,\mathbf{x}_5\}$. Given **X** already generated using the above protocol, the focus is now on obtaining $\boldsymbol{\beta}$. Here, $\boldsymbol{\beta}$ is generated from the eigenvectors of $(\mathbf{X}'\mathbf{X})$. Three $\boldsymbol{\beta}$ vectors are obtained from $(\mathbf{X}'\mathbf{X})$ and used to produce three sets of $(\mathbf{X}\boldsymbol{\beta})$ values having different degrees of variability, namely $\boldsymbol{\beta}_{max}$, $\boldsymbol{\beta}_{min}$, and $\boldsymbol{\beta}_{int}$. The eigenvector corresponding to the largest eigenvalue of $(\mathbf{X}'\mathbf{X})$ is denoted as $\boldsymbol{\beta}_{max}$, that corresponding to the smallest eigenvalue as $\boldsymbol{\beta}_{min}$, and that equal to ½($\boldsymbol{\beta}_{max}+\boldsymbol{\beta}_{min}$) as $\boldsymbol{\beta}_{int}$. So, according to Johnson and Wichern (1992), $(\mathbf{X}\boldsymbol{\beta}_{max})$ possesses the largest variability, $(\mathbf{X}\boldsymbol{\beta}_{min})$ the smallest variability, and $(\mathbf{X}\boldsymbol{\beta}_{int})$ the intermediate variability. Given **X** and $\boldsymbol{\beta}$, $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. Here, $\boldsymbol{\varepsilon}$ is simulated from a normal distribution with a mean of 0 and a user-specified variance, σ^2.

Two Modeling Conditions

Given **X** and **y**, the performance of information criteria is examined under two conditions. One condition has the true model included in the pool of candidate models, whereas the other one does not. The good model is to be identified in both conditions. When the true model is in the pool, the good model is just the true model. Otherwise, the good model is the one that is "closest" to the true model.

When the True Model is Included

This part of the analysis assesses the number of times that *ICOMP* criteria successfully identify the true model, which *ICOMP* criteria overfit a model, and that *ICOMP* criteria underfit a model. To add more competing models to the pool, two additional variables $\mathbf{x}_6$ and $\mathbf{x}_7$ are added to $\mathbf{X}$ with both of them generated from an exponential distribution Exp (0.1). A total of seven models are evaluated and compared using information criteria, namely $\{\mathbf{x}_1\}$, $\{\mathbf{x}_1,\mathbf{x}_2\}$, …, and $\{\mathbf{x}_1,\mathbf{x}_2, ..., \mathbf{x}_K\}$, $K = 3, 4, …, 7$. The true model is the one with five predictors: $\{\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3,\mathbf{x}_4,\mathbf{x}_5\}$.

When the True Model is Not Included

This part of the analysis assesses the number of times that *ICOMP* criteria select the good model minimizing the K-L distance between the true model and each estimated model. Here, $\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{x}_3$, and $\mathbf{x}_4$ are used to create the pool of candidate models. A total of four models are created, evaluated, and compared using information criteria, namely $\{\mathbf{x}_1\}$, $\{\mathbf{x}_1,\mathbf{x}_2\}$, $\{\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3\}$, and $\{\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3,\mathbf{x}_4\}$. The true model is still the one with five predictors: $\{\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3,\mathbf{x}_4,\mathbf{x}_5\}$, although it is not in the pool of competing models. The model in the pool that minimizes the K-L distance from the true model is the one with four predictors: $\{\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3,\mathbf{x}_4\}$. Hereafter, Models 1 through 7 refer to the regression models with 1 through 7 predictor variables, respectively. For example, Model 3 is the regression model that contains just three predictors $\mathbf{x}_1$ through $\mathbf{x}_3$, or $\{\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3\}$.

Simulation Results

With the True Model Included

Tables 1, 2, and 3 present the model selection results from the case when the true model is included, with Table 1 corresponding to $\boldsymbol{\beta}_{max}$, Table 2 to $\boldsymbol{\beta}_{int}$, and Table 3 to $\boldsymbol{\beta}_{min}$. In each table, seven model selection criteria are scored to evaluate seven regression models: Models 1 to 7 described above under three sample sizes (i.e., small, medium, and large): $n_{min} = 50$, $n_{int} = 100$, and $n_{max} = 1000$. Since it is Model 5 that simulates the data, the goal of using model selection criteria is to identify this model as the best model.

Under each $\boldsymbol{\beta}$ by n combination, two sets of simulations are run. In the first set of simulations, a total of 100 runs are performed, whereas in the second set, as many as 10,000 runs are performed. So, cells in each table contain two integers separated by a forward slash sign which are frequencies of each competing model being selected under the two sets of simulations (100 runs/10,000 runs), respectively. Model selection results from the two sets of simulations are compared with each other in a few aspects: frequency and/or percentage of identifying the best approximating model, etc. Conclusions are drawn from the patterns found from both sets of simulations. Given any inconsistency in results between the two sets of simulations, those from the second set with a larger number of simulations prevail, because they explore a larger model space.

In addition to model selection frequencies in each of the tables, Figures 1 and 2 present the average percentage of the true model (Model 5) selection as a function of sample size and variability in ($\mathbf{X}\boldsymbol{\beta}$), respectively. Finally, Figure 3 compares all seven criteria in terms of the range of percentages of each of Models 1 through 7 being selected.

The model selection results are examined in the following three aspects:

(1) The increase in sample size tends to improve the performance of all seven criteria in identifying the true model, or Model 5, and this supports the consistency property of all seven criteria. This trend is indicated relatively clearly in all seven line graphs in Figure 1, particularly when the number of runs is larger. In that figure, when the number of runs is 10,000, with an increase in sample size (from 50 to 100, again to 1,000), each line graph keeps showing an upward trend, which indicates that the average percentage of successfully identifying the true model is increasing. When the number of runs is only 100, five of the seven information criteria present an upward trend with an increase in sample size. Two of them, AIC_{C} and $ICOMP_{\mathrm{C1F}}$,

Table 1. Frequency of Model Selection Given Maximum Variability with True Model (100/10,000 runs)

Criterion	*n*	1	2	3	4	5*	6	7
AIC	50	0/0	0/0	0/6	2/143	72/7179	15/1506	11/1166
	100	0/0	0/0	0/0	0/0	78/7582	12/1467	10/951
	1000	0/0	0/0	0/0	0/0	73/7822	15/1337	12/841
*AIC*c	50	0/0	0/0	0/11	2/210	79/8112	12/1085	7/582
	100	0/0	0/0	0/0	0/1	84/8052	9/1251	7/696
	1000	0/0	0/0	0/0	0/0	73/7874	15/1313	12/813
CAIC	50	0/0	0/1	0/52	6/526	89/8940	4/381	1/100
	100	0/0	0/0	0/0	0/15	99/9698	1/252	0/35
	1000	0/0	0/0	0/0	0/0	99/9947	1/49	0/4
SBC	50	0/0	0/1	0/30	3/371	87/8751	7/613	3/234
	100	0/0	0/0	0/0	0/9	97/9490	3/414	0/87
	1000	0/0	0/0	0/0	0/0	99/9890	1/100	0/10
$ICOMP_{C1}$	50	0/0	0/0	0/0	0/41	95/9437	4/407	1/115
	100	0/0	0/0	0/0	0/0	97/9532	3/387	0/81
	1000	0/0	0/0	0/0	0/0	96/9615	4/331	0/54
$ICOMP_{C1F}$	50	0/0	0/0	0/0	0/22	50/5152	31/2849	19/1977
	100	0/0	0/0	0/0	0/0	53/5041	26/2969	21/1990
	1000	0/0	0/0	0/0	0/0	37/5007	39/3028	24/1965
$ICOMP_{Mis}$	50	0/0	0/0	0/0	0/100	93/9236	6/532	1/132
	100	0/0	0/0	0/0	0/2	97/9423	3/482	0/93
	1000	0/0	0/0	0/0	0/0	94/9578	6/362	0/60

Table 2. Frequency of Model Selection Given Intermediate Variability with True Model (100/10,000 runs)

Criterion	*n*	1	2	3	4	5*	6	7
AIC	50	0/6	0/142	10/644	14/1696	54/5190	13/1282	9/1040
	100	0/1	1/53	2/305	14/1204	61/6221	12/1334	10/882
	1000	0/0	0/0	0/0	0/6	73/7818	15/1337	12/839
*AIC*c	50	0/9	0/198	12/854	17/2033	57/5541	8/877	6/488
	100	0/2	1/62	2/343	15/1333	66/6510	9/1123	7/627
	1000	0/0	0/0	0/0	0/6	73/7870	15/1313	12/811
CAIC	50	1/111	2/602	21/1658	20/2459	51/4864	4/240	1/66
	100	0/16	4/292	9/917	26/2144	60/6428	1/181	0/22
	1000	0/0	0/0	0/0	0/25	99/9922	1/49	0/4
SBC	50	1/51	1/412	18/1316	20/2318	54/5292	4/434	2/177
	100	0/11	4/212	9/751	21/1954	63/6681	3/329	0/62
	1000	0/0	0/0	0/0	0/23	99/9867	1/100	0/10
$ICOMP_{C1}$	50	0/0	0/5	0/99	10/831	85/8548	4/403	1/114
	100	0/0	0/5	0/31	7/447	90/9051	3/386	0/80
	1000	0/0	0/0	0/0	0/1	96/9614	4/331	0/54
$ICOMP_{C1F}$	50	0/0	0/1	0/42	5/411	45/4761	31/2822	19/1963
	100	0/0	0/1	0/11	5/172	49/4863	25/2966	21/1987
	1000	0/0	0/0	0/0	0/0	37/5007	39/3028	24/1965
$ICOMP_{Mis}$	50	0/0	0/31	2/195	10/1232	81/7924	6/492	1/126
	100	0/0	0/9	1/78	6/639	90/8719	3/467	0/88
	1000	0/0	0/0	0/0	0/2	94/9576	6/362	0/60

* The true model

Table 3. Frequency of Model Selection Given Minimum Variability with True Model (100/10,000 runs)

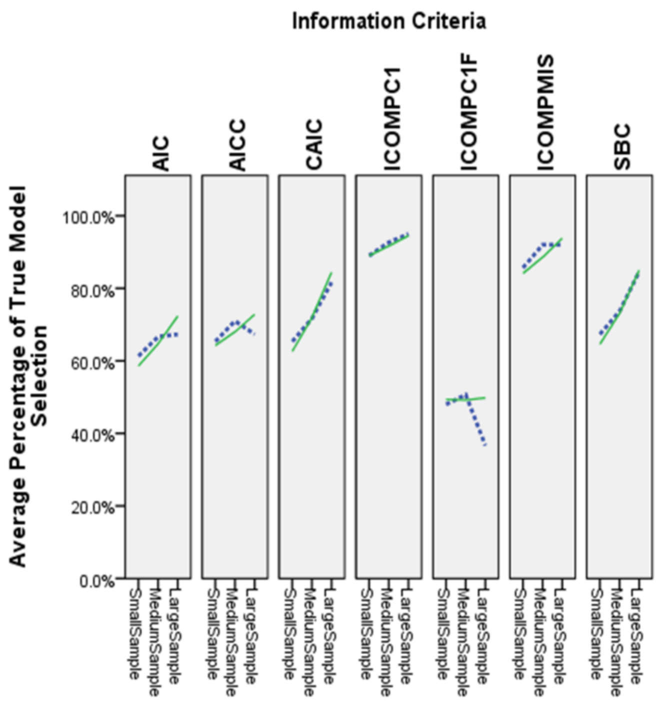
Criterion	*n*	1	2	3	4	5*	6	7
AIC	50	0/0	3/133	4/755	13/1599	58/5205	14/1271	8/1037
	100	0/0	1/67	6/647	15/1567	61/5608	8/1262	9/849
	1000	0/0	0/19	10/504	13/1500	56/6059	13/1162	8/756
*AIC*c	50	0/0	3/184	5/969	15/1886	60/5616	12/867	5/478
	100	0/0	1/77	6/735	17/1693	63/5832	6/1050	7/613
	1000	0/0	0/19	11/509	12/1508	56/6098	13/1139	8/727
CAIC	50	1/9	5/576	13/1854	21/2244	56/4977	3/263	1/77
	100	0/0	4/331	19/1799	20/2252	56/5430	1/161	0/27
	1000	0/0	1/106	20/2060	31/2344	47/5462	1/27	0/1
SBC	50	0/2	4/405	7/1494	20/2149	61/5337	6/450	2/163
	100	0/0	2/228	14/1507	21/2173	61/5741	2/283	0/68
	1000	0/0	1/87	19/1832	23/2287	56/5734	1/56	0/4
$ICOMP_{\mathrm{C1}}$	50	0/0	0/2	2/80	6/680	87/8719	4/404	1/115
	100	0/0	0/2	0/50	6/564	91/8917	3/386	0/81
	1000	0/0	0/0	1/30	2/458	93/9130	4/328	0/54
$ICOMP_{\mathrm{C1}F}$	50	0/0	0/0	0/24	1/265	49/4894	31/2843	19/1974
	100	0/0	0/0	0/6	3/183	50/4856	26/2967	21/1988
	1000	0/0	0/0	0/0	1/83	36/4924	39/3028	24/1965
$ICOMP_{\mathrm{Mis}}$	50	0/0	0/16	2/194	10/1091	83/8080	4/493	1/126
	100	0/0	0/6	0/123	8/889	89/8428	3/463	0/91
	1000	0/0	0/0	2/52	4/543	88/8989	6/356	0/60

* The true model

have a turning point when the sample size is medium, indicating that they perform the best when the sample is neither largest nor smallest. This observation under only 100 simulations is not consistent with that when the number of runs is 10,000, thus we consider it to be untrustworthy due to the small number of simulations. Finally, the performance of $ICOMP_{\mathrm{C1F}}$ does not seem to be very consistent with that of the rest. Its performance under 10,000 runs of simulations increases only slightly when the sample size jumps from 50 to as large as 1,000, whereas all other criteria show a marked increase in the average percentage of identifying the true model when increasing the sample size.

(2) The increase in the variability of (**Xβ**) tends to improve the performance of all seven criteria. This trend is clearly indicated in Figure 2 for both sets of simulations for six of the seven criteria (excluding $ICOMP_{\mathrm{CIF}}$); and, the two trend lines representing 100 and 10,000 simulations in each of the six graphs almost completely overlap, so that they are almost indistinguishable from each other. When sample size increases from 50 to 1,000, a marked increase in the average percentage of identifying the true model is observed for *AIC* (approximately from 60% to 78%), AIC_{C} (approximately from 60% to 80%), *SBC* (approximately from 60% to 96%), and *CAIC* (approximately from 58% to 98%). A relative moderate increase is observed for $ICOMP_{\mathrm{C1}}$ (approximately from 90% to 99%) and $ICOMP_{\mathrm{Mis}}$ (approximately from 90% to 98%). These two *ICOMP* criteria are already successful at as high as 90% of the time when (**Xβ**) assumes the minimum variability, so there is not much room for improvement for the two of them given more variability in (**Xβ**). Finally, $ICOMP_{\mathrm{C1F}}$ fails to meet our expectations again this time. When the other criteria are becoming more and more capable of identifying the true model with increasing variability in (**Xβ**), the increase in the performance of $ICOMP_{\mathrm{C1F}}$ is negligible under the larger set of simulations.

(3) An overall comparison of all seven criteria is found in Figures 1, 2, and 3. In Figures 1 and 2, it can be seen that on average both $ICOMP_{\mathrm{C1}}$ and $ICOMP_{\mathrm{Mis}}$ tend to outperform non-*ICOMP*



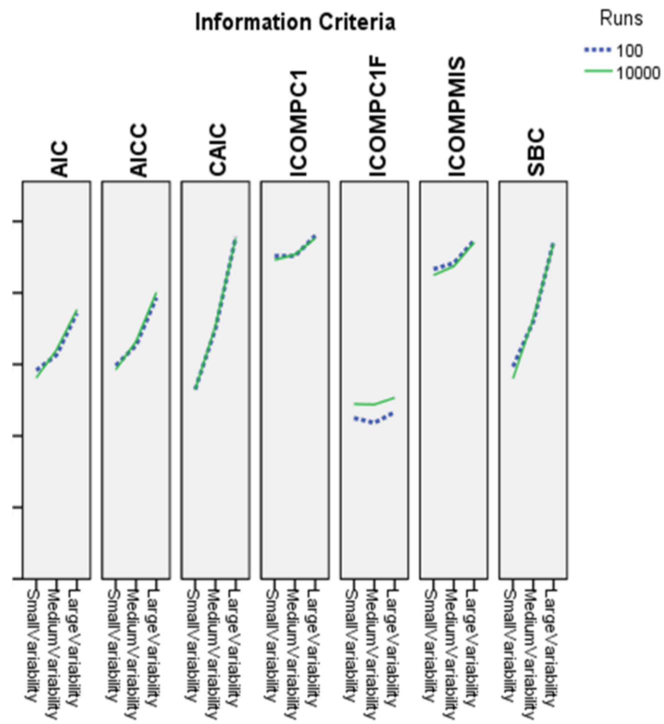


Figure 1. Comparison of average percentage of true model selection (Model 5) as a function of sample s under 100 and 10000 runs of simulations.

Figure 2. Comparison of average percentage of true model selection (Model 5) as a function of variability under 100 and 10000 runs of simulations.

criteria: *AIC*, *AIC*$_C$, *SBC*, and *CAIC*, and, in Figure 3, the range of percentages of successfully identifying the true model from each simulation condition tends to be higher for the two *ICOMP* criteria than for all other criteria. However, *ICOMP*$_{C1F}$ does not seem to perform as well as the other two *ICOMP* criteria, and is probably the worst of all seven criteria in terms of the likelihood of identifying the true model. The bad performance of this criterion is due to its tendency to select more complex models, either Model 6 or Model 7. In Figure 3, such an overfitting tendency of *ICOMP*$_{C1F}$ is clearly observed. This criterion is much more likely to select either Model 6 or Model 7 than all other criteria, thus causing it to be less successful in identifying the true model, or Model 5.

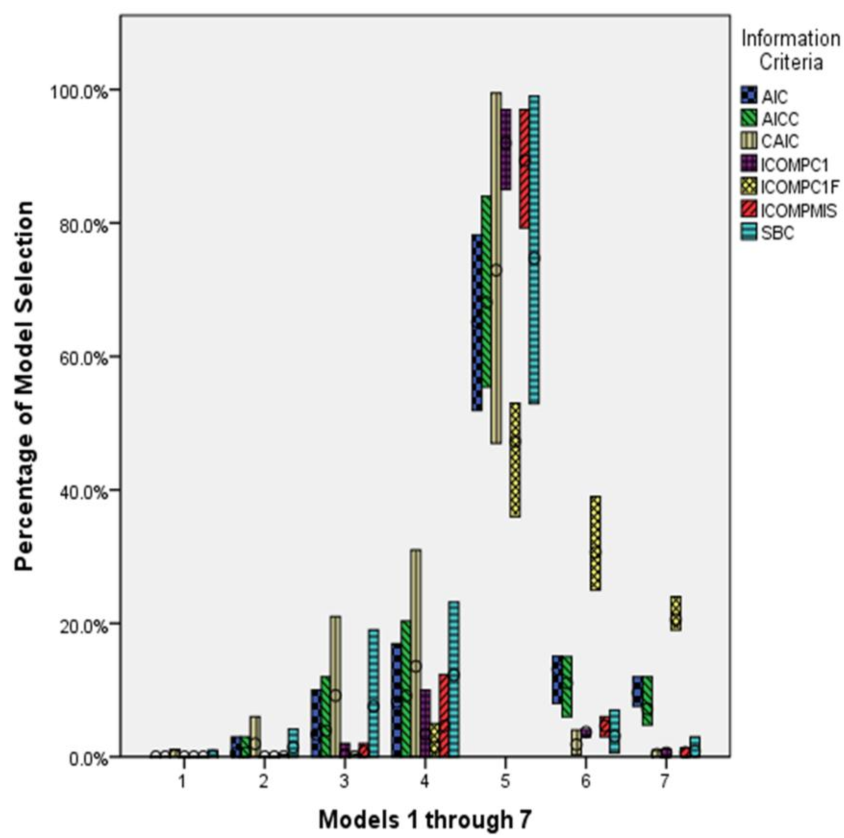


Figure 3. Comparison of range of percentages of model selection by all seven information criteria with true model.

With the True Model Excluded

Tables 4, 5, and 6 present the model selection results from the case when the true model is excluded, with Table 4 corresponding to $\boldsymbol{\beta}_{max}$, Table 5 to $\boldsymbol{\beta}_{int}$, and Table 6 to $\boldsymbol{\beta}_{min}$. In each figure, seven model selection criteria are scored to evaluate four regression models: Models 1 to 4 described above, with Model 4 being the best approximating model of the true model: Model 5. Three different sample sizes (i.e., small, medium, and large) are used, namely n_{min} = 50, n_{int} = 100, and n_{max} = 1000.

Similar to the previous case with the true model included, under each $\boldsymbol{\beta}$ by n combination, two sets of simulations are performed for the purpose of cross-validating model selection results. The first set contains 100 runs of simulations whereas the second set 10,000 runs. So, cells in each of Tables 4, 5, and 6 also contain two integers separated by a forward slash sign which represent frequencies of each competing model being selected under the two sets of simulations (100 runs/10,000 runs), respectively.

Besides, Figures 4 and 5 present the average percentage of the best approximating model (Model 4) selection as a function of sample size and variability in ($\mathbf{X}\boldsymbol{\beta}$), respectively. Finally, Figure 6 compares all seven criteria using the range of percentages of each of Models 1 through 4 being selected under each simulation condition.

Under the second case, where Model 4 is the best, similar patterns of criterion performance are found. In Figure 4, the two lines of 100 and 10,000 simulations both show a continuing upward trend with an increase in sample size for all seven criteria (i.e., the $ICOMP_{C1}$ line for the smaller number of simulations

Table 4. Frequency of Model Selection Given Maximum Variability Without True Model (100/10,000 runs)

Criterion	n	1	2	3	4*
AIC	50	0/0	1/1	2/278	97/9721
	100	0/0	0/0	0/9	100/9991
	1000	0/0	0/0	0/0	100/10000
*AIC*c	50	0/0	1/2	5/346	94/9652
	100	0/0	0/0	0/14	100/9986
	1000	0/0	0/0	0/0	100/10000
CAIC	50	0/0	1/25	9/828	90/9147
	100	0/0	0/0	1/76	99/9924
	1000	0/0	0/0	0/0	100/10000
SBC	50	0/0	1/12	7/611	92/9377
	100	0/0	0/0	1/51	99/9949
	1000	0/0	0/0	0/0	100/10000
$ICOMP_{C1}$	50	0/0	0/0	0/40	100/9960
	100	0/0	0/0	0/0	100/10000
	1000	0/0	0/0	0/0	100/10000
$ICOMP_{C1F}$	50	0/0	0/0	0/31	100/9969
	100	0/0	0/0	0/0	100/10000
	1000	0/0	0/0	0/0	100/10000
$ICOMP_{Mis}$	50	0/0	0/0	1/153	99/9847
	100	0/0	0/0	0/3	100/9997
	1000	0/0	0/0	0/0	100/10000

Table 5. Frequency of Model Selection Given Intermediate Variability Without True Model (100/10,000 runs)

Criterion	n	1	2	3	4*
AIC	50	0/38	4/392	25/2023	71/7547
	100	0/1	2/218	12/1226	86/8555
	1000	0/0	0/0	0/3	100/9997
*AIC*c	50	0/45	5/481	27/2244	68/7230
	100	0/2	3/234	12/1285	85/8479
	1000	0/0	0/0	0/3	100/9997
CAIC	50	1/238	7/1068	39/3041	53/5653
	100	0/35	7/634	21/2133	72/7198
	1000	0/0	0/2	0/10	100/9988
SBC	50	1/145	6/828	33/2762	60/6265
	100	0/23	5/513	20/1919	75/7545
	1000	0/0	0/1	0/8	100/9991
$ICOMP_{C1}$	50	0/5	2/83	8/747	90/9165
	100	0/0	0/39	4/380	96/9581
	1000	0/0	0/0	0/0	100/10000
$ICOMP_{C1F}$	50	0/5	2/58	4/575	94/9362
	100	0/0	0/25	3/266	97/9709
	1000	0/0	0/0	0/0	100/10000
$ICOMP_{Mis}$	50	0/6	1/159	13/1092	86/8743
	100	0/0	0/57	7/579	93/9364
	1000	0/0	0/0	0/0	100/10000

* The best approximating model

may deviate a little bit, though), thus supporting their property of consistency. In Figure 5, such a continuing upward trend is also observed for all seven criteria when the variability in (**Xβ**) increases. Finally, the performance of *ICOMP* criteria is generally better than that of non-*ICOMP* criteria. This is true of all three *ICOMP* criteria. In Figures 4 and 5, the average performance of each *ICOMP* criterion under smallest sample size or smallest (**Xβ**) variability is generally the same as or even better than that of each non-*ICOMP* criterion under largest sample size or largest (**Xβ**) variability. In Figure 6, the range of percentages of successfully identifying the best approximating model under each simulation condition tends to be higher for the three *ICOMP* criteria than for the four non-*ICOMP* criteria. Although $ICOMP_{C1F}$ performs less satisfactorily in the previous case that includes the true model, it performs as well as the other two *ICOMP* criteria in this second case. Such an increase in performance is probably

Table 6. Frequency of Model Selection Given Minimum Variability Without True Model (100/10,000 runs)

Criterion	*n*	1	2	3	4*
AIC	50	3/127	5/572	20/2720	72/6581
	100	1/10	2/296	29/2533	68/7161
	1000	0/0	0/58	22/2094	78/7848
*AIC*c	50	3/177	8/680	20/2921	69/6222
	100	1/10	3/315	29/2665	67/7010
	1000	0/0	0/58	23/2099	77/7843
CAIC	50	8/604	14/1269	29/3513	49/4614
	100	1/48	9/767	43/3765	47/5420
	1000	0/0	1/196	39/4023	60/5781
SBC	50	8/401	10/1042	26/3322	56/5235
	100	1/31	8/618	39/3549	52/5802
	1000	0/0	1/167	37/3821	62/6012
$ICOMP_{C1}$	50	1/3	1/76	6/883	92/9038
	100	0/1	0/34	2/677	98/9288
	1000	0/0	0/4	10/523	90/9473
$ICOMP_{C1F}$	50	1/3	0/32	6/582	93/9383
	100	0/1	0/9	1/388	99/9602
	1000	0/0	0/1	2/184	98/9815
$ICOMP_{Mis}$	50	1/12	3/159	6/1509	90/8320
	100	0/2	0/54	10/1077	90/8867
	1000	0/0	0/8	10/616	90/9376

* The best approximating model

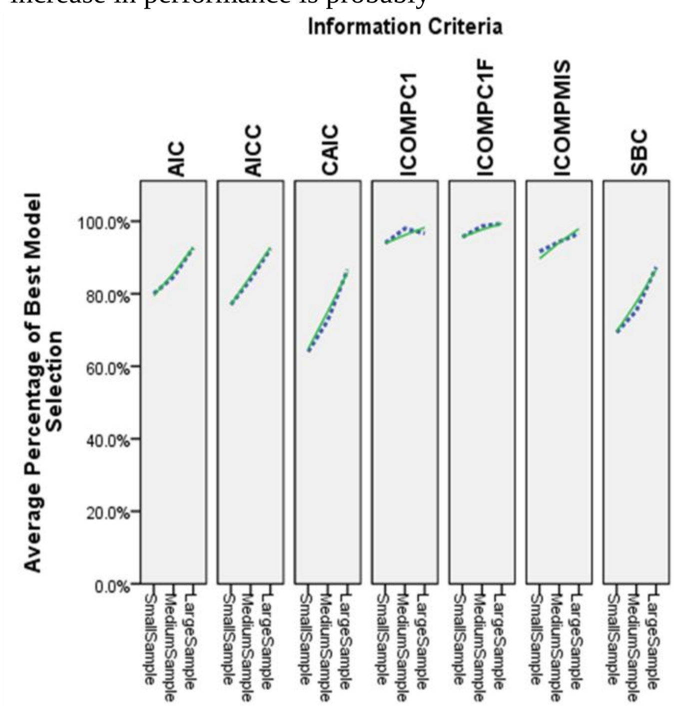


Figure 4. Comparison of average percentage of best approximating model selection (Model 4) as a function of sample size under 100 and 10000 runs of simulations.

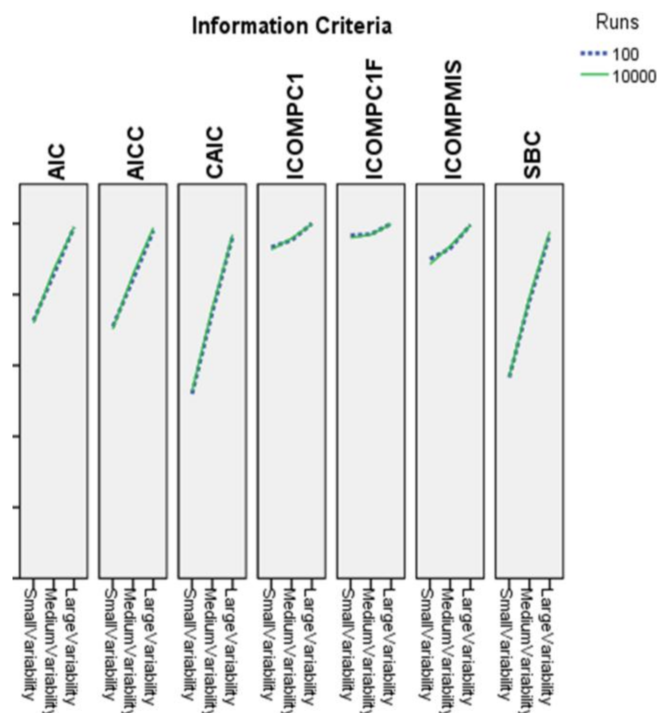


Figure 5. Comparison of average percentage of best approximating model selection (Model 4) as a function of variability under 100 and 10000 runs of simulations.

because this criterion tends to overfit a model and the best approximating model in the second case is already the most complex model. In other words, in both cases, $ICOMP_{C1F}$ tends to select a more complex model and, in the second case only, the most complex model happens to be the best approximating model.

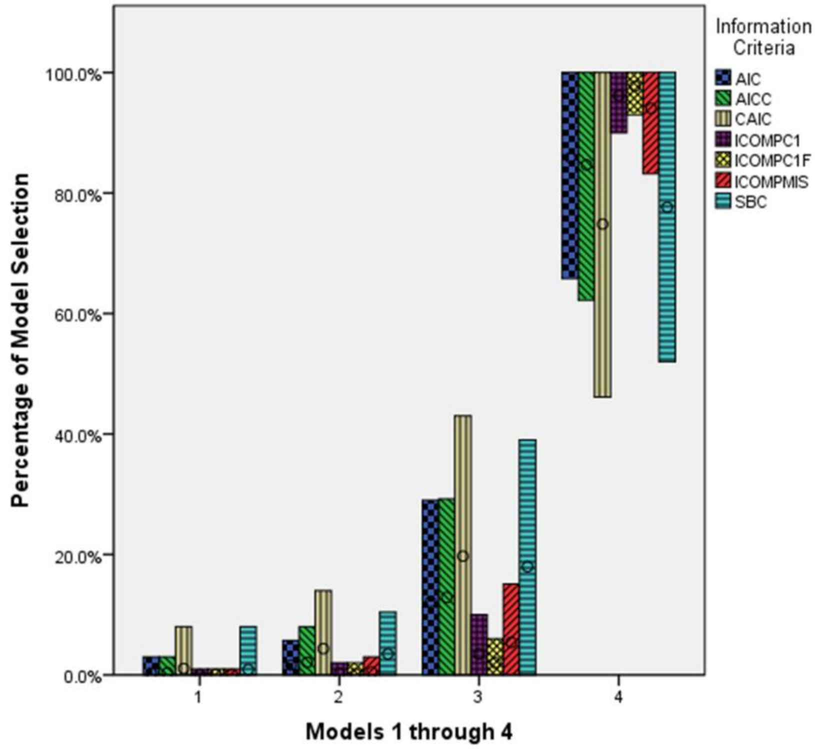


Figure 6. Comparison of range of percentages of model selection by all seven information criteria without true model.

Conclusion

The paper provides support for the use of two *ICOMP* criteria in multiple linear regression to supplement existing information criteria commonly found in major statistics programs: *AIC*, *CAIC*, *SBC*, etc. The two recommended *ICOMP* criteria are $ICOMP_{C1}$ and $ICOMP_{Mis}$. However, this paper has some reservations for the third *ICOMP* criterion, or $ICOMP_{C1F}$, because it is usually prone to overfitting.

The two recommended *ICOMP* criteria are usually more capable of successfully identifying the best approximating model than other criteria under the simulations of multiple linear regression modeling in this study. And their effectiveness can generally be improved by either increasing sample size or increasing the variability in (**Xβ**).

Future research on *ICOMP* could focus on its application to linear and nonlinear mixed models, which are extensions of the type of linear models covered in this paper. Mixed models consist of both fixed and random components and are capable of analyzing grouped, nested, or hierarchical data structures that are more commonly seen in many fields of study. *ICOMP* would be used to select fixed and/or random components in mixed models. Special *ICOMP* formulas should be developed for mixed models that correspond to formulas for marginal and conditional *AIC* (Vaida & Blanchard, 2005).

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & B. F. Csaki (Eds.), *Second international symposium on information theory* (pp. 267.281). Budapest, Hungary: Akademiai Kiado.

Akaike, H. (1987). Factor analysis and *AIC*. *Psychometrika, 52,* 317-332.

Ando, T. (2010). *Bayesian model selection and statistical modeling*. Boca Raton, FL: Chapman and Hall.

Bozdogan, H. (1987). Model selection and Akaike's Information Criterion (*AIC*): The general theory and its analytical extensions. *Psychometrika, 52,* 345-370.

Bozdogan, H. (2004). Intelligent statistical data mining with information complexity and genetic algorithms. In H. Bozdogan (Ed.), *Statistical data mining and knowledge discovery* (pp. 15-56). Boca Raton, FL: Chapman and Hall.

Bozdogan, H., & Haughton, H. (1998). Information complexity criteria for regression models. *Computational Statistics and Data Analysis, 28,* 51-76.

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and inference: A practical information-theoretic approach* (2nd ed.). New York: Springer.

Claeskens, G., & Hjort, N. L. (2008). *Model selection and model averaging*. Cambridge, UK: Cambridge University Press.

Johnson, R. A., & Wichern, D. W. (1992). *Applied multivariate statistical analysis* (3rd ed.). Englewood Cliffs, NJ: Prentice Hall.

Konishi, S., & Kitagawa, G. (2010). *Information criteria and statistical modeling*. New York: Springer.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics, 22*, 79-86.

Kutner, M., Nachtsheim, C., & Neter, J. (2004). *Applied linear regression models* (4th ed.). New York: McGraw-Hill.

McQuarrie, A. D. R., & Tsai, C. L. (1998). *Regression and time series model selection*. River Edge, NJ: World Scientific Publishing.

Miller, A. (2002). *Subset selection in regression* (2nd ed.). Boca Raton, FL: Chapman and Hall.

Mulaik, S. A. (2009). *Linear causal modeling with structural equations*. Boca Raton, FL: Chapman and Hall.

Schwartz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6,* 461-464.

Sclove, S. L. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika, 52*, 333-343.

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. Boca Raton, FL: Chapman & Hall.

Tamhane, A. C., & Dunlop, D. (1999). *Statistics and data analysis: From elementary to intermediate*. Upper Saddle River, NJ: Prentice Hall.

Vaida, F., & Blanchard, S. (2005). Conditional Akaike information for mixed-effects models. *Biometrika, 92*, 351-370.

Van Emden, M. H. (1971). An analysis of complexity. In *Mathematical Centre Tracts* (Vol. 35). Amsterdam, Netherlands: Mathematisch Centrum.

Send correspondence to: Hongwei Yang
University of Kentucky
Email: hya222@uky.edu