Electronic Journal of Statistics
Vol. 7 (2013) 2762–2793
ISSN: 1935-7524
DOI: 10.1214/13-EJS861

Model selection rates of information based criteria

Ashok Chaurasia and Ofer Harel

Department of Statistics, University of Connecticut, 215 Glenbrook Road, Unit 4120, Storrs, CT 06269-4120
e-mail: achaurasia.uconn@gmail.com; ofer.harel@uconn.edu

Abstract: Model selection criteria proposed over the years have become common procedures in applied research. This article examines the true model selection rates of any model selection criteria; with true model meaning the data generating model. The rate at which model selection criteria select the true model is important because the decision of model selection criteria affects both interpretation and prediction.

This article provides a general functional form for the mean function of the true model selection rates process, for any model selection criteria. Until now, no other article has provided a general form for the mean function of true model selection rate processes. As an illustration of the general form, this article provides the mean function for the true model selection rates of two commonly used model selection criteria, Akaike's Information Criterion (AIC) and Bayesian Information Criterion (BIC). The simulations reveal deeper insight into properties of consistency and efficiency of AIC and BIC. Furthermore, the methodology proposed here for tracking the mean function of model selection procedures, which is based on accuracy of selection, lends itself for determining sufficient sample size in linear models for reliable inference in model selection.

AMS 2000 subject classifications: Model selection, model selection rate, AIC, BIC, discrete process, discrete process mean function, multiple linear regression, linear models, generalized linear models.

Received January 2013.

1. Introduction

The task of model selection is common in most disciplines where the objective is to select a simple model that "best" explains or predicts the data. In the search for the "best" model(s), it is possible for the true data generating model to not be in the collection of models being considered in the analysis phase. Also, there is no guarantee that the subset of predictors deemed as "best" will be unique. However, in practice it is common to assume that the true model is contained in the class of models under consideration.

The most common approach for comparison and selection of models is penalized model selection criteria. In various disciplines, the most commonly used penalized model selection criteria are the Akaike Information Criterion (Akaike, 1974) and Bayesian Information Criterion (Schwarz, 1978). These model selection procedures are well known and documented in statistical literature. The existing literature on AIC and BIC, though extensive in discussing properties of consistency and efficiency, does not address the important question of their true model selection rates, when the true model is contained in the class of model under consideration. The first two papers that discuss the rate of convergence of AIC are Shibata (1981) and Hurvich and Tsai (1995). The major conclusion in these two papers implies that the model selected by AIC will contain the true model with probability 1. However, this conclusions still does not address the probability with which AIC will select the true model.

The main purpose of this article is to describe the true model selection rate for any model selection criterion (MSC), when the true model is contained in the class of model under consideration. When making a decision about a model that explains the data, it is prudent that we understand the reliability and consequence of that decision. In response to such questions, this paper quantifies the reliability of decisions made by commonly used model selection criteria, such as AIC and BIC. The methodology developed in this paper can be extended to other diverse model selection criteria.

The rest of the article is organized as follows. Section 2 provides the model and notation to be used in the succeeding sections and mathematically describes the objective of this article. Section 3 briefly discusses AIC and BIC in terms of objective, assumption and formulation. Section 4 provides the intuition and motivation for the proposed method for tracking the correct of model rate. Section 5 presents the simulation study designed to illustrate the performance of each criteria in selecting the correct model. Section 6 provides an application in sample size determination for reliable model inference by taking advantage of the proposed success rate function. Section 7 provides a discussion on extension of the proposed success rate function in data with outliers and in generalized linear models (GLM). Section 8 provides a summary of findings.

2. Model and notation

This article primarily focuses on the multiple regression model, given as follows:

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{1}$$

where $\boldsymbol{y}$ is a n-dimensional vector, $\mathbf{X}$ is a $n \times p$ matrix, $\boldsymbol{\beta}$ is a p-dimensional vector, and $\boldsymbol{\varepsilon}$ is the n-dimensional error vector with $\boldsymbol{\varepsilon} \sim Normal_n(\mathbf{0}, \sigma^2\mathbf{I})$. Suppose the collection of candidate models under consideration, $\boldsymbol{\mathcal{M}}$, contains the data generating ("true") model $\mathcal{M}_*$ of dimension p_*. Let $\mathscr{D}_{p_*,n,p}$ represent the data generated from $\mathcal{M}_*$ (unknown), with n observations, p potential predictors, and $n > p \geq p_*$. The objective is to perform model selection on $\mathscr{D}_{p_*,n,p}$ for the linear model (1) set-up.

Under model (1), suppose that a model is selected using some model selection criterion (MSC) for the specific purpose of prediction or interpretation. Let $T_{p_*,n,p}$ be the indicator function with success indicating that model selected by MSC is same as the "true" data generating model. Consequently, if p_*, n and p are allowed to vary (increase) under the condition that $n > p \geq p_*$, then the indexed set of random variables given by $\{T_{p_*,n,p}\}_{\{(p_*,n,p):\ n>p\geq p_*\}}$ represents a (Bernoulli) process in p_*, n and p.

The main objective of this article is to study this process and estimate its **mean** function. Estimation of the mean function requires repeats for each setting of (p_*, n, p). Hence, the new collection $\{T^{(i)}_{p_*,n,p}\}_{\{(i,p_*,n,p):\ i\geq 1\ \&\ n>p\geq p_*\}}$ denotes the process with repeats (given by i). Observe that for any (n, p_*, p), $\mathbf{Var}(T_{p_*,n,p}) \leq 0.25$, so by weak law of large numbers for any fixed (n, p_*, p) with $n > p \geq p_*$ and $1 \leq i \leq J$, $J^{-1}\sum_{i=1}^{J} T^{(i)}_{p_*,n,p} \xrightarrow{\mathbb{P}} \mathbf{E}(T_{p_*,n,p})$ as $J \to \infty$.

3. Methods

A model selection procedure intends to select the "best" subset of predictor variables for a specific purpose such as interpretation or prediction. The term "best" usually refers to a balance between the number of explanatory variables and goodness of fit. Earlier model selection criteria centered around (adjusted) residual sums of squares, stepwise methods for selecting significant variables, or information theoretic approaches (Rao and Wu, 2001). The list of model selection selection procedure is long and still growing. While some of the new methods evolve from modifications/improvements to existing procedures like AIC and BIC, other use a different measures (other than Kullback-Leibler) to evaluate discrepancy between candidate models and the supposed true model. Some examples include Divergence Information Criterion (DIC; Spiegelhalter et al., 1998), Residual Information Criterion (RIC; Shi and Tsai, 2002), Akaike Information Criterion with Fisher Information (AICF; Cetin and Erar, 2002), Focused Information Criterion (FIC; Claeskens and Hjort, 2003). An excellent summary of existing model selection procedures from the frequentist, Bayesian and nonparametric perspectives is discussed in Rao and Wu (2001) and Kadane and Lazar (2004). This article explores the true model selection rates of any model selection procedure and provides an illustration using AIC and BIC. The purpose of introducing only AIC and BIC in this section and not other modern model selection methods, was merely because these methods are most familiar to the general audience and are readily available in most software packages. Moreover, the methodology proposed in this paper applies generally to any decision making process, just as those made by model selection procedures.

3.1. Akaike information criterion

Akaike (1974) introduced the Akaike Information Criterion (AIC), an information theoretic approach for model (variable) selection, via Kullback-Leibler divergence. AIC is one of the most common model selection procedures that is available in most statistical software packages. Under the setting of model (1) when $\mathscr{D}_{p_*,n,p}$ is fully observed, AIC for a candidate model (denoted as $\mathcal{M}_c$) is given as follows:

$$AIC = n\ln(\hat{\sigma}^2_c) + 2p_c$$

where $\hat{\sigma}_c^2$ is the maximum likelihood estimate of σ^2 under candidate model $\mathcal{M}_c$ of dimension p_c. In general, AIC for a candidate model is given by

$$AIC = -2\ln(\mathcal{L}_c) + 2p_c \tag{2}$$

where $\mathcal{L}_c$ is the likelihood estimate under $\mathcal{M}_c$. Under this selection criterion, the model with the smallest AIC value is deemed as best. AIC was derived as an asymptotically unbiased estimator of the expected Kullback-Leibler discrepancy between the true and a fitted candidate model. The derivation of (2) requires two important assumptions – (i) the class of models under consideration contains the true model, i.e. $\mathcal{M}_* \in \boldsymbol{\mathcal{M}}$, and (ii) the parameter estimates obtained from maximizing the likelihood of $\mathcal{M}_c$ satisfy the regularity conditions of Maximum Likelihood Estimators (Rao, 1945; Cramér, 1946). The popularity of AIC is due to the fact that its derivation is quite general, it adheres to the concept of parsimony, and is easy to implement in models such as (1). Like any model selection criterion, AIC has its drawbacks. Many researchers have shown that for small samples, AIC is inconsistent and leads to overfitting (e.g., Hurvich and Tsai, 1989). However, AIC has several advantages – it is asymptotically efficient, allows for simultaneous comparison of multiple nested or non-nested models, and allows for model-averaged inference (e.g., Burnham and Anderson, 2004). For details on derivation of AIC refer to Burnham and Anderson (2002) and Cavanaugh (1997).

3.2. Bayesian Information Criterion

Schwarz (1978) proposed the the Bayesian Information Criterion (BIC) as an asymptotic approximation to a transformation of the Bayesian posterior probability of a candidate model. For a candidate model ($\mathcal{M}_c$) the computation of BIC is based on the empirical log-likelihood ($\mathcal{L}_c$) and does not require the specification of priors; it is given as follows:

$$BIC = -2\ln(\mathcal{L}_c) + p_c\ln(n) \tag{3}$$

where p_c is the dimension of $\mathcal{M}_c$. Under the assumption that the true model is contained in the class of models under consideration (i.e. $\mathcal{M}_* \in \boldsymbol{\mathcal{M}}$), it is well known that BIC is consistent (see, e.g., Shibata, 1981; Nishii, 1984). This means that BIC selects the true model with probability one as sample size increases, i.e. $n \to \infty$. A drawback of BIC is inefficiency, which means that asymptotically BIC selects the candidate model which minimizes the mean squared error of prediction. For further details readers are referred to Kass and Raftery (1995) and Neath and Cavanaugh (1997).

4. Intuition & motivation for functional form of rate of correct selection

First, suppose $X \sim \text{Bernoulli}(\phi_{\texttt{naïve}})$ with $\phi_{\texttt{naïve}} = 2^{-1}$, represents the indicator if a model chosen by a naïve model selection criterion (NSC) with success indicating that model selected by NSC is same as the "true" data generating model. In terms of model selection, a NSC with correct model selection rate of 2^{-1}, is far from ideal. In other words, a MSC with correct model selection rate $\phi = 2^{-a}$, where a denotes the rate of the MSC and $a \in [0, 1)$, is preferred over NSC.

Second, suppose p_* denotes the dimension of the "true" model. Then, for any fixed and unknown p_*, as the number of potential predictors ($p \geq p_*$) increases, a decrease in performance by any MSC (i.e. decreasing ϕ), is to be expected. In other words, it is expected that ϕ decreases when the difference $p - p_*$ increases. Hence, it is reasonable to propose the functional form $\phi(p, p_*) = 2^{-a(p-p_*)}$ where $p \geq p_*$ and $a \in [0, 1)$.

Third, suppose model selection criteria MSC_1 and MSC_2 with "true" model selection rates $2^{-a_1(p-p_*)}$ and $2^{-a_2(p-p_*)}$, respectively. Without loss of generality, let $a_1 < a_2$, then for a fixed and unknown p_*, as p increases MSC_1 is preferred over MSC_2. This is because MSC_2 is the first to reach 2^{-1} in comparison to MSC_1, as p increases. For example, if the two model selection criteria are $MSC_1 = BIC$ and $MSC_2 = AIC$, then we expect $a_{_{BIC}} = a_1 < a_2 = a_{_{AIC}}$ because AIC is known to overfit (Hurvich and Tsai, 1989). Hence, it is reasonable to suspect that a MSC with value for a further away from 0 tends to overfit, while a MSC with value for a closer (or equal) to 0 tends to fit smaller dimension models that are closest to the "true" model.

Fourth, for $p \ll n \to \infty$, the decreasing sequence $\{\phi(p, p_*)\}_{p \geq p_*}$ for any MSC is a geometric sequence with

$$\sum_{p=p_*}^{\infty} \phi(p, p_*) = \frac{1}{1 - 2^{-a}};\ 0 < a < 1.$$

In the previous statement it is assumed that the rate at which $p \to \infty$ is much smaller than the rate at which $n \to \infty$. Thus, for sufficiently "large" n it is reasonable to propose the following functional form:

$$\phi(p, p_*) = \left\{ \begin{array}{rl} 2^{-a(p-p_*)} & : \ p \geq p_* \text{ and } 0 < a < 1 \\ 0 & : \ \text{otherwise} \end{array} \right. . \tag{4}$$

Finally, when n is not sufficiently "large", then (it is reasonable to assume that) this should have a negative effect on the correct model selection rate of any MSC. Also, for estimation of the error variance in model (1) we require that $n - (p + 1) \geq 0$. Hence, we propose that the correct model selection rate for any MSC is inversely proportional to $(n - (p + 1))^c$ with proportionality constant given by $w(p - p_*)^k$, where constants $w \in (0, 1)$, $k \in (0, 1]$ and $c \in (0, 1]$ are unknown. The reason for such a proportionality constant is two folds – (i) to make the negative effect of not having sufficiently "large" n purely a function of $p - p_*$ (just as it is for when n is "large"), and (ii) the unknown constants w, k, and c allow for the fine tuning of the negative effect of not having sufficiently "large" n. Thus, based these assumptions, for any MSC, the following functional form is proposed for the mean function of the Bernoulli process $\{T_{p_*,n,p}\}_{\{(p_*,n,p):\ n>p\geq p_*\}}$.

$$\phi_n(p,p_*) = \begin{cases} \phi(p,p_*) - \frac{w(p-p_*)^k}{(n-p-1)^c} & : \ p \geq p_* \\ 0 & : \ \text{otherwise} \end{cases} \tag{5}$$

where

$$\phi(p,p_*) = \begin{cases} 2^{-a(p-p_*)} & : \ p \geq p_* \\ 0 & : \ \text{otherwise} \end{cases} . \tag{6}$$

Also note the following relationship in the limit between $\phi_n(p,p_*)$ and $\phi(p,p_*)$

$$\underset{n\to\infty}{lim}\ \phi_n(p,p_*) = \begin{cases} \phi(p,p_*) & : \ p \geq p_* \\ 0 & : \ \text{otherwise} \end{cases} . \tag{7}$$

In equation (5), the entities $\phi_n(p,p_*)$, $\phi(p,p_*)$ and $\frac{w(p-p_*)^k}{(n-p-1)^c}$ represent proportions with $\max\left\{\phi_n(p,p_*), \frac{w(p-p_*)^k}{(n-p_*-1)^c}\right\} < \phi(p,p_*)$. The previous condition imposes the following restrictions on the tuning parameters: $w \in (0,1)$, $k \in (0,1]$, $c \in (0,1]$. Furthermore, $\phi(p,p_*)$ (the entity not dependent on n) is viewed as a measure of consistency of a MSC whereas the difference $\phi(p,p_*) - \phi_n(p,p_*)$ which equals $\frac{w(p-p_*)^k}{(n-p-1)^c}$ is viewed as a measure of efficiency of a MSC. Recall that $\phi(p,p_*)$ is the probability that a MSC will select the true model when sample size is infinite, and $\phi_n(p,p_*)$ is the probability that a MSC will select the true model with a finite sample size. Hence, a MSC that has $\phi_n(p,p_*)$ close or equal to 1 (equivalently, a equal or close to 0) will be consistent. On the other hand, for a given n and $p \geq p_*$, smaller values for the difference $\phi(p,p_*) - \phi_n(p,p_*)$ $\left(\text{i.e. } \frac{w(p-p_*)^k}{(n-p-1)^c}\right)$ means that the model selection criterion is efficient. Hence, the proposed mean function for true model selection rate for any MSC suggests that no model selection criterion can have best of both worlds – consistency and efficiency, i.e. for any MSC, gain in efficiency will come at the cost of loss in consistency, and vice verse. To end this section, it should be pointed out that construction of the proposed mean function for the correct selection process (given by equation (5)) was not built with a particular model selection procedure in mind. It was merely from reflecting on reliability of a decision making process, such as that made by any model selection criterion.

5. Simulations

5.1. Setup

For the simulation study, different data configurations are considered in order to track the process described in section 2. The following data parameters are varied in order to obtain different data configurations. The sample size: n= 20 to 2000 by 20 and true model dimension: $p_* = 2,3,4$. Given a fixed p_*, the potential number of predictors considered: $p = p_*, p_*+1, \ldots, 12$.

For each combination of (n, p_*, p), the matrix $\mathbf{X}$ of dimension $n \times p$ is generated from multivariate distribution $\mathcal{F}$. Given $\mathbf{X}$, if $\boldsymbol{x}_i$ represents the ith column of $\mathbf{X}$ and $\boldsymbol{\varepsilon} \sim Normal_n(\mathbf{0}, \mathbf{I})$, then $\boldsymbol{y}$ is generated as follows: if $p_* = 2$ then $\boldsymbol{y} = 3\boldsymbol{x}_1 + 2\boldsymbol{x}_2 + \varepsilon$, if $p_* = 3$ then $\boldsymbol{y} = 3\boldsymbol{x}_1 + 2\boldsymbol{x}_2 + 1\boldsymbol{x}_3 + \varepsilon$, and if $p_* = 4$ then $\boldsymbol{y} = 3\boldsymbol{x}_1 + 2\boldsymbol{x}_2 + 1\boldsymbol{x}_3 + 3\boldsymbol{x}_4 + \varepsilon$.

For each combination of $(n, p_*, p, \mathcal{F})$, data was simulated $J = 100$ times. In each data set, i.e. each combination of $(n, p_*, p, \mathcal{F}, J)$, model selection is performed with AIC and BIC. Given an $\mathcal{F}$ and a MSC, for each combination of (n, p_*, p), $\hat{\phi}_n(p_*, p) = J^{-1} \sum_{i-1}^{J} T^{(i)}_{n,p_*,p}$ (as discussed in section 4) is recorded.

Finally, for sensitivity analysis of the proposed mean function from section 4, the following distributions are considered for $\mathcal{F}$. First, $\mathcal{F} = Normal_p(\mathbf{0}, \mathbf{I})$, i.e. columns of $\mathbf{X}$ are generated from IID (Independent and Identically Distributed) standard normals. In the sections that follow, this setting will be referred to as **IID Normal**. Second, $\mathcal{F} = Normal_p(\mathbf{0}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma}_{ij} = \left\{ \begin{smallmatrix} 1 & : & i=j \\ 0.9 & : & i<j \end{smallmatrix} \right.$ for $i, j = 1, 2, \ldots, p$. This setting represents highly correlated columns of $\mathbf{X}$ and will be referred to as **Correlated Normal** in the following sections. Third, $\mathcal{F} = Multvariate\ Gamma$ where the marginals (each column of $\mathbf{X}$) have an exponential distribution with scale parameter value of 2. In addition the correlation coefficient between the columns is about 0.85. This kind of structure was chosen to reflect extremely (right) skewed data with high dependence between the columns of $\mathbf{X}$. In the sections that follow, this setting will be referred to as **Correlated Gamma**. In the simulations, the total number of data configurations considered for $p_* = 2, 3, 4$ was 3267, 2970, and 2673, respectively.

5.2. Results

The extensive simulation study results in numerous tables and graphs so, for brevity we only discuss some of the graphs corresponding to the setting where $p_* = 2$, $p = 3, 4, 5, 6, 7, 8$ and $\mathcal{F} =$ **Normal IID**. The graphs corresponding to setting where $p_* = 2$, $p = 3, 4, 5, 6, 7, 8$, $\mathcal{F} =$ **Correlated Normal** and $\mathcal{F} =$ **Correlated Gamma** are provided in appendix A. The graphs corresponding to configuration with $p_* = 3, 4$ are not included since the findings about AIC and BIC were the same across all data configurations considered in the simulation study.

5.2.1. Results for AIC

In keeping track of $\hat{\phi}_n(p, p_*)$, for "large" values of n the true $\phi(p, p_*)$ (as given in equation (5)) is estimated by $\hat{\phi}(p, p_*) = \frac{1}{20} \sum_{n \in \mathfrak{N}} \hat{\phi}_n(p_*, p)$, where $\mathfrak{N} = 1600$ to 2000 by 20. In light of equations (4) and (7), $(1 - 2^a)^{-1} = \hat{\phi}(p, p_*)$ yields that the constant a when using AIC (denoted as $a_{_{AIC}}$) is $\frac{1+\sqrt{2}}{10} \approx 0.2414$. This value of $a_{_{AIC}}$ was constant throughout all data configurations of $(p, p_*, \mathcal{F})$. In light of equation (5), by taking the difference of $\hat{\phi}_n(p_*, p)$ and $\hat{\phi}(p, p_*)$ and equating it to $\frac{w(p-p_*)^k}{(n-p-1)^c}$ we obtain the estimates of w, k and c (using non-linear estimation techniques) to get 2^{-2}, 1, and 1, respectively. These values were constant across all data configurations. Hence, the proposed mean function for true model selection rate by AIC is given as follows:

$$\phi_n^{AIC}(p,p_*) = \begin{cases} \phi_{AIC}(p,p_*) - 0.25\frac{(p-p_*)}{(n-p-1)} & : \ p \geq p_* \\ 0 & : \ \text{otherwise} \end{cases} \tag{8}$$

where

$$\phi_{AIC}(p,p_*) = \begin{cases} 2^{-0.1(1+\sqrt{2})(p-p_*)} & : \ p \geq p_* \\ 0 & : \ \text{otherwise} \end{cases} .$$

The resiliency of the functional form (8) is validated in the graphs shown in subplots of figures 1 (and subplots of figure 4 and 6 provided in appendix A).

In equation (8), as $n \to \infty$ with $n \gg p$, $\phi_{AIC}(p,p_*)$ decreases exponentially as p increases. For illustration of this inconsistency, see figures 1 (a)–(e) which correspond to the simulation setting $p_* = 2$, $p = 3,\ldots,8$, $\mathcal{F} =$ **IID Normal** and MSC=AIC. In figures 1 (a)–(e), as p goes from 3 to 8 (respectively) the horizontal dotted line ($\phi_{AIC}(p,p_*)$) is (approx.) 0.85, 0.72, 0.61, 0.52, 0.44, and 0.37, respectively. These plots illustrates that as p increases (with $p \ll n$) the success rate of AIC in selecting the true model drops exponentially towards zero, and consequently the rate of selecting an over-fitted model by AIC approaches 1. This is not a surprising result because AIC tries to select model that is best for prediction not interpretation. When looking into the derivation of AIC it is clear that its objective is based on minimizing final prediction errors (Hurvich and Tsai, 1995), which AIC successfully accomplishes with probability 1 when $n \gg p \to \infty$.

In equation (8), the difference $\phi_{AIC}(p,p_*) - \phi_n^{AIC}(p,p_*)$ given by $\frac{0.25(p-p_*)}{(n-p-1)}$ indicates the rate of efficiency of AIC. From the plots in figure 1 (a)–(e), the smooth solid curve ($\phi_n^{AIC}(p,p_*)$: success rate at n) catches up with the dotted horizontal line ($\phi(p,p_*)$) fairly early in n. For example observe that in figures 1 (a)–(e) the solid smooth curve is indistinguishable from the horizontal dotted line after sample size of (approximately) $n = 250$. This illustrates the efficiency of AIC.

5.2.2. *Results for BIC*

Similar to the method employed in calculating a_{AIC}, we find that $a_{BIC} = 0$. This value was constant throughout all the data configuration described by combinations of $(p,p_*,\mathcal{F})$. This implies that $\phi_{BIC}(p,p_*) = 1$, which is expected because as $n \to \infty$ BIC will select the generating model with probability one, i.e. $\phi_n^{BIC}(p,p_*) \to \phi_{BIC}(p,p_*) = 1$ (Schwarz, 1978). In light of equation (5), by taking the difference of $\hat{\phi}_n^{BIC}(p_*,p)$ and $\hat{\phi}_{BIC}(p,p_*)$ and equating it to $\frac{w(p-p_*)^k}{(n-p-1)^c}$ we obtain the estimates of w, k and c (using non-linear estimation techniques) to get 0.5,1, and 0.6, respectively. These values were constant across all data configurations. Hence, the mean function for true model selection rate by BIC is given as follows:

$$\phi_n^{BIC}(p,p_*) = \begin{cases} \phi_{BIC}(p,p_*) - 0.5\frac{(p-p_*)}{(n-p-1)^{0.6}} & : \ p \geq p_* \\ 0 & : \ \text{otherwise} \end{cases} \tag{9}$$

where

$$\phi_{BIC}(p) = \begin{cases} 1 & : \ p \geq p_* \\ 0 & : \ \text{otherwise} \end{cases} .$$

The resiliency of the functional form (9) is validated in the graphs shown in subplots in figure 2 (and subplots of figure 5 and 7 provided in appendix A).

In equation (9), as $n \to \infty$ with $n \gg p$, $\phi_{BIC}(p,p_*)$ remains at 1 as p increases. For illustration of this consistency, observe figures 2 (a)–(e) which correspond to the simulation setting $p_* = 2$, $p = 3,\ldots,8$, $\mathcal{F} =$ **IID Normal** and MSC=BIC. As p increases from 3 to 8 (respectively) the horizontal dotted line ($\phi_{BIC}(p,p_*)$) remains at 1. These plots illustrates that as p increases (with $p \ll n$) the success rate of BIC in selecting the true model remains at 100%, and consequently the rate of selecting an overfitted model by BIC stays at 0. This is not a surprising result because the objective of BIC is based on maximizing the accuracy in selecting a candidate model that is closest to the true model (Schwarz, 1978), which BIC successfully accomplishes with probability 1 when $n \gg p \to \infty$.

In equation (9), the difference $\phi_{BIC}(p,p_*) - \phi_n^{BIC}(p,p_*)$ given by $\frac{0.5(p-p_*)}{(n-p-1)^{0.6}}$ indicates the rate of (in)efficiency of BIC. From figures 2 (a)–(e), it is clear that the smooth solid curve ($\phi_n^{BIC}(p,p_*)$: success rate at n) catches up with the dotted horizontal line ($\phi_{BIC}(p,p_*)$) quite late in n. For example, in figures 2 (a)–(e) the solid smooth curve is easily distinguishable from the horizontal dotted line even when the sample size is 2000; this illustrates the inefficiency of BIC.

5.2.3. *Comparison of results from AIC and BIC*

Now that the functional form of the mean functions for AIC and BIC are established, we note that the rate (in n) at which $\phi_n(p_*,p)$ catches up with $\phi(p_*,p)$ is much faster in AIC than BIC. Specifically, for any p_*, p with $p \geq p_*$ and given a fixed positive constant $\mathbb{B}$ that represents the difference between $\phi_n(p,p_*)$ (i.e. true model selection success rate with a finite sample size n) and $\phi(p,p_*)$ (i.e. the asymptotic (in n) true model selection success rate), we have the following statement.

If $\mathbb{B} = \phi_{BIC}(p,p_*) - \phi_n^{BIC}(p,p_*) = \phi_{AIC}(p,p_*) - \phi_n^{AIC}(p,p_*)$, then

$$n_{BIC} = [2\,(n_{AIC} - p - 1)]^{10/6} + p + 1. \tag{10}$$

Equation (10) implies that given a model, the sample size required for $\hat{\phi}_n(p_*,p)$ to be within a certain (fixed) bound ($\mathbb{B}$) of $\phi(p_*,p)$, with BIC will be (much) larger than the minimum sample size (required for the same fixed bound) with AIC. The catching up rate of AIC reflects the asymptotic efficiency of AIC and inefficiency of BIC. However, the efficiency of AIC comes at a cost because $\phi_{_{AIC}}(p_*, p)$ decreases exponentially while $\phi_{_{BIC}}(p_*, p)$ remains at 1, as p increases. This result is not surprising when observed from the objective standpoint of AIC and BIC; AIC places importance on precision of selection for prediction whereas BIC targets the accuracy of selection. The classic examples of model selection procedures, AIC and BIC, illustrate that the consistency of a MSC can be obtained **only** at the cost of asymptotic efficiency (Schwarz, 1978; Hannan and Quinn, 1979). This statement is also implied by the proposed mean function (5) which states that as $a \to 0^+$ a MSC gains consistency and loses efficiency, whereas $a \to 1^-$ a MSC gains efficiency and loses consistency. A comparison of the proposed means functions for AIC and BIC (in equations (8) and (9)) confirms these well known facts about AIC and BIC. For example, recall in section 4 (third paragraph) we claimed that $a_{_{BIC}} < a_{_{AIC}}$ which was indeed confirmed in the simulation study: $a_{_{BIC}} = 0$ and $a_{_{AIC}} = 0.25$.

Next we compare the effect of increasing the p on the true model selection rates of AIC and BIC, using figures 3(a) and 3(b). First for AIC, we see in figure 3(a) that increasing p from 3 to 5, and then to 8, causes the mean success rate of AIC (represented by the smooth curve) to drop vertically. In contrast, for BIC, we see in figure 3(b) a very different migration of the contours corresponding to $p = 3$, 5 and 8; the contours shift both vertically (downward) and horizontally (to the right) as p increases. These graphs illustrate that for n, p, p_* where $n > p \geq p_*$, the true model selection rate of BIC is always larger than that of AIC, however having $\phi_n^{^{BIC}} > \phi_n^{^{AIC}}$ does not imply that the decision made by BIC is reliable. For example, when $n = 30$, $p_* = 2$, and $p = 8$ (see figures 3(a) and 3(b)), we observe that $\phi_n^{^{AIC}}$ and $\phi_n^{^{BIC}}$ are (approximately) 0.27 and 0.45, respectively. Indeed, $\phi_n^{^{BIC}} > \phi_n^{^{AIC}}$ but with a true model selection rate of a mere 45% (recall ϕ values $\gg 0.5$ are preferred) doesn't make BIC very reliable at that sample size. Based on simulations (results not included), $\phi_n^{^{BIC}} = 0.45$ implies the 55% of the time BIC selected an overfitted model, i.e. a model containing the true model. The issue of reliability at a given sample size is not limited to AIC or BIC, but pertains to any decision making process. The next section describes two situations where one can take advantage of the proposed mean function of AIC and BIC in determining sufficient sample sizes for reliable inference. Similar approach can be implemented with other model selection criteria.

6. Application

6.1. Application 1: Sample size determination for design of experiments

From the perspective of designing experiments, the proposed mean function for any MSC (given by 5) would aid in power and sample size calculation for designing cost effective experiments. A simple application is given by the following hypothetical practical scenario. Suppose a researcher is interested in modeling an outcome $\boldsymbol{y}$ using predictors $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_{10}$. From existing literature or pilot studies, it is well documented that predictors $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ (say, `Gender` and `Race`) serve as good predictors of $\boldsymbol{y}$, so in the designing phase of this study the researcher elects to always measure these two variables. In addition to $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, the researchers wants to investigate if $\boldsymbol{y}$ can be better explained using variables $\boldsymbol{x}_3, \boldsymbol{x}_4, \ldots, \boldsymbol{x}_{10}$, which have shown potential as good predictors from different studies. Specifically, the researcher wants to calculate the sample size required to provide reliable model selection results. Since measurement of variables comes at a cost, our proposed mean function for correct model selection will aid in cost effective ways of choosing the variables of interest in the designing phase. For example, in this practical scenario $p = 10$, $2 \leq p_* \leq p$, which implies $d = p - p_* \in \{0, 1, 2, \ldots, 8\}$. Let us suppose the researcher wants the reliability of the decision (denoted as $\mathscr{R}$) of at least 80%. In the given scenario, the researcher has limited $2 \leq p_* \leq 10$ (usually it would be $1 \leq p_* \leq p$). Suppose the researcher is interested in the accuracy of model selection and chooses BIC as the criterion of choice. Then, the minimum sample size required for the smallest (and most difficult) model to be detected as the best model (if it is indeed the true model) with reliability of at least $\mathscr{R} \in (0, 1)$ (high values preferred, say 80%) is given by

$$\begin{aligned} n_{\min} &= \min\left\{n \in \mathbb{Z}^+ : \phi_n^{BIC}(d_{\max}) \geq \mathscr{R}\right\} \\ &= \left\lceil \left(\frac{0.5}{1-\mathscr{R}} d_{\max}\right)^{\frac{5}{3}} + p + 1 \right\rceil \\ &= \left\lceil \left(\frac{0.5(8)}{1-0.80}\right)^{\frac{5}{3}} + 10 + 1 \right\rceil \\ &= \left\lceil 20^{\frac{5}{3}} + 10 + 1 \right\rceil \\ &= 159 \end{aligned} \tag{11}$$

where $\lceil \cdot \rceil$ is the ceiling function. If $n_{\min} = 159$ requires a budget beyond what is available, then a large enough sample cannot be collected for reliable model inference. In order to meet the budget limitations, compromises have to be made based on the researcher and/or expert's opinion about which variables among $\boldsymbol{x}_3, \boldsymbol{x}_4, \ldots, \boldsymbol{x}_{10}$ are more important to the study objective, while taking into account the costs related to measuring each variable. If the variables of interest are of similar cost per experimental unit, then emphasis should be placed on the smallest model (as was done in equation (11)) that a sample size will allow (as a consequence of fixed budget) to be reliably detected as the true model, if it is indeed the true model.

In the case where a researcher is interested in models for prediction and chooses AIC, then the minimum sample size calculation is not as straightforward and requires some important observations. In the given example, where $d_{\max} = 8$, the smallest possible probability for correct selection by AIC given $p = 10$ is $\phi(d_{\max}) = 0.2622$. Note, this is the probability at infinite sample size which implies that even with infinite sample size, the probability of choosing the (most difficult) model of dimension $p_* = 2$ as best (if it is indeed the true model) is at least 26.22%. This in turn implies that probability of selecting an overfitted model is at most 73.78%. Hence, the researcher in this example cannot expect the reliability of selecting an overfitted model to exceed 73.78%. In a compromise for a parsimonious model let us suppose the researcher limits the probability of overfitting to be at most 60%, and consequently limits the probability of correct selection to at least 40% (denoted as $\mathscr{R}$). With this in mind, the minimum sample size is determined as follows:

$$\begin{aligned} n_{\min} &= \min\left\{n \in \mathbb{Z}^+ : \phi_n^{AIC}(d_{\max}) \geq \mathscr{R}\right\} \\ &= \left\lceil \frac{0.25}{\phi_{AIC}(d_{\max}) - \mathscr{R}} d_{\max} + p + 1 \right\rceil \\ &= \left\lceil \frac{0.25}{2^{-0.1(1+\sqrt{2})d_{\max}} - 0.4} d_{\max} + 10 + 1 \right\rceil \\ &= 49 \end{aligned} \tag{12}$$

where (as per our hypothetical example),

$$d_{\max} = -\left\lceil \frac{\log_2(\mathscr{R})}{a_{AIC}} \right\rceil = -\left\lceil \frac{\log_2(0.4)}{0.1(1+\sqrt{2})} \right\rceil = 5. \tag{13}$$

6.2. Application 2: Reliability and sample size determination in observational studies

In the case of observational studies, where p is already fixed and n observations have been measured, our proposed mean function can again be helpful in many ways. First, it can give a range for the reliably of detecting the smallest or the largest model as best, if one of the respective models is indeed the true model. For example, continuing with the hypothetical example, where $p = 10$ and $2 \leq p_* \leq 10$, suppose that an observational study has already collected 100 observations. In this scenario, the first concern is what are the reliability values for detecting the true model when using BIC. This is simply answered by evaluating minimum reliability, which is given by $\phi^{BIC}(n = 100, d = 8) =$ 72.93%. If the researcher of this study, finds 72.93% to be low and requires reliability of at least 80% then from table 1 (row $p = 10$ and column $\mathscr{R} \geq 80\%$) the sample size required is at least 159. [1]

In the case of AIC, the minimum reliability of detecting the true model is given by $\phi^{AIC}(n = 100, d = 8) = 23.97\%$. If the researcher finds 23.97% to be too low and wants to boost this probability to at least 40%, then from table 2 (row $p = 10$ and column $\mathscr{R} \geq 40\%$) the sample size required is at least 49.

To conclude this section, note that it is the mean function of true model selection that led to application in design of experiments and observational studies for reliable model inference. Hence, the methodology presented in this paper

[1] Such tables for BIC were constructed for different values of p_* and are available on request.

lends itself to developing average selection rate functions for any model selection procedure, and consequently its application for reliable model inference.

7. Discussion: Outliers, GLM, and extensions

Thus far our findings have been in the simple class of linear models with normally distributed errors. It would be interesting to investigate the performance of the proposed success rate function in more general/complex settings. For example, it would be worthwhile to study the effect of outliers on the success rate of a MSC, and the success rate of a MSC in GLM. With focus on these two settings, a discussion follows next on the mean success rate function of AIC and BIC.

With regards to data with outliers, it is expected that the mean function form given by equation (5) should hold for large samples. For example, simulations were conducted with errors from Student's t-distribution with 5 degrees of freedom and Laplace distribution with location parameter 0 and scale parameter 1. Under the t-distribution with 5 degrees of freedom, the probability of observing extreme values (less than -3 or greater than 3) is 0.03 which is about 11 times the probability of observing extreme values under the standard normal distribution. Under the standard Laplace distribution, the probability of observing extreme values (less than -3 or greater than 3) is 0.05 which is about 18 times the probability of observing extreme values under the standard normal distribution. The results of these supplemental scenarios for AIC are given in figures 8 (a)–(f) and 9 (a)–(f), and for BIC are given in figures 10 (a)–(f) and 11 (a)–(f). These figures again illustrate the resiliency of the functional form proposed for AIC and BIC in the multiple regression setting.

Lastly, in the addressing the mean function for correct selection in GLM, a final simulation study was conducted to observe the behavior of the proposed mean function of BIC in logistic regression where data contains outliers from Student's t-distribution (with 5 degrees of freedom) and standard Laplace distribution. Figures 12 and 13 confirm the validity of the general mean functional form given by equation (5) in section 4. [2]

Additionally, the logistic regression setting also suggests modification to equation (5) for an even more general form given by

$$\phi_n(p, p_*) = \begin{cases} \phi(p, p_*) - \frac{w(p-p_*)^k}{(n-p-r)^c} & : \ p \geq p_* \\ 0 & : \ \text{otherwise} \end{cases} \tag{14}$$

where

$$\phi(p, p_*) = \begin{cases} 2^{-a(p-p_*)} & : \ p \geq p_* \\ 0 & : \ \text{otherwise} \end{cases} . \tag{15}$$

where "r" is added to list of tuning parameters with $p + r > n$. The significance of "r" is to allow for sample sizes where the MLE properties of estimators in GLM will hold true.

[2] Simulations were also done for AIC and since the results were similar, they have been excluded for brevity.

In the simulation study considered in this paper, larger discrepancies were observed between the proposed mean function and observed mean in data configurations with sample size. So, for small samples there is definitely room for improvement and this is currently being explored in a forthcoming paper. Another natural extension includes studying the mean function in GLM where the response is a polytomous (ordinal or nominal) variable. A general extension is to study the mean function in GLM where the response variable are counts from (for example) Binomial, Poisson, Negative Binomial, etc. Other areas of future research include studying the behavior of the proposed method when $\mathbf{X}$ contains different types of variables: continuous, binary, proportions, ordinal, nominal. The methodology presented in this paper could be employed with other model selection criteria. With access to the success rate of a model selection procedure (using the functional form given by equation (14)), we can (i) learn about about their properties such as consistency and efficiency, and (ii) develop rules for sample size calculation for reliable model inference.

8. Concluding remarks

Shibata (1981), and later Hurvich and Tsai (1990), were the first to discuss and formalize model selection rates of AIC. Their conclusions pointed to the well known characteristics of AIC – inconsistency and efficiency. Their conclusions, under certain assumptions, state that AIC will either select the true model or an overfitted model, with probability 1. This paper provides a finer result to compliment their conclusion. This paper provides the probability with which AIC selects the true model in linear models. Consequently, our results imply that when using AIC the probability of overfitting increases exponentially with number of predictors under consideration.

In addition, this paper also provides a general functional form for the mean function of the true model selection rate (discrete) process for any model selection criteria, in the generalized linear models. For validation of the general functional form of the mean function, this paper provides the mean functional forms of AIC and BIC, and illustrates their exceptional performance in an extensive simulation study. The simplicity of the proposed mean function (merely a function of the data dimension) and its resiliency (to data configuration) are some of many strong suits of the proposed mean function. The mean functions of the true model selection rates process of AIC and BIC further confirm their well known properties of consistency and efficiency. The methodology presented in this paper can be used in deriving mean functions for process relating to other diverse model selection criteria and learn about their consistency and efficiency.

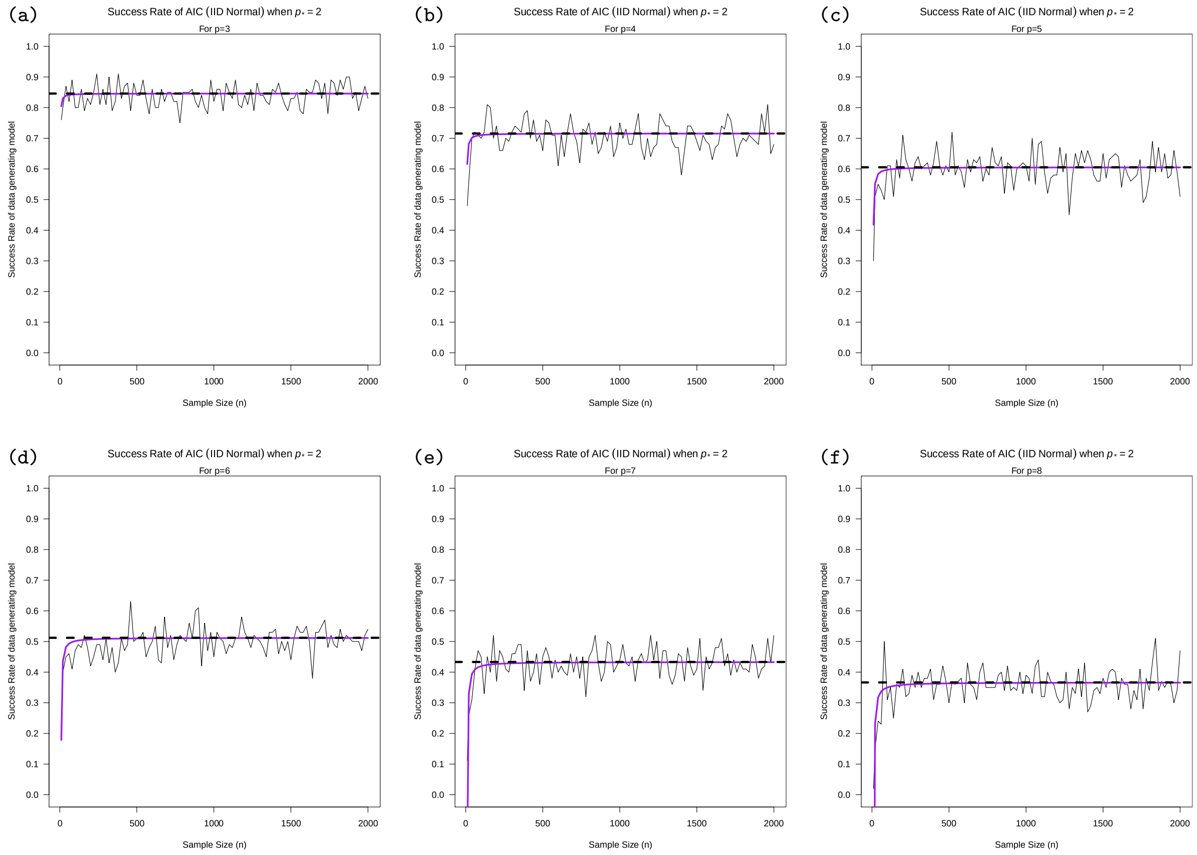
Furthermore, in the multiple linear regression setup, the notion of what quantifies as “large” sample size relative to the number of predictor variables, has always remained unclear with respect to model selection. For example, the derivation of AIC in the elegant and insightful paper by Cavanaugh (1997) requires that $p^2 \ll n$ so that $p^2/n \to 0$, i.e. the rate at which p^2 increases is much smaller than the rate at which n is increases. In practical terms, what ratio of p^2 to n

can be considered as close enough to zero has remained unanswered, until now. With the proposed mean function, the notion of having a “large” enough sample size can be quantified in terms of a pre-specified minimum reliability (detection probability) of correct selection.

In closing, the purpose of this paper was to study and pique interest in evaluating the mean function forms for different model selection procedures. The use of AIC and BIC was mainly for illustrative purpose. Further research in model selection criteria as processes is needed in order to study their model selection rate, which is greatly hindered by the dimension of the data. Focus should be placed in developing the mean function for the true model selection rate of model selection procedures, and to consequently determine their (decision) reliability threshold at a given sample size, and vice versa. Furthermore, the relationship between the reliability of a model selection procedure and sample size will give a unique quantification of $n \gg p$ (where required) for different model selection procedures based on their model detection rates (power). To our knowledge, this paper is the first to quantify $n \gg p$ for the purpose of reliable model selection in generalized linear models. Just as several authors (eg. Cohen and Cohen, 1975; Cohen, 1988, 1992; Green, 1991; Aron and Aron, 2003; Kelley and Maxwell, 2003) have provided unique sample size rules for different statistics, this paper provides a simple but effective methodology (based on accuracy of selection) for determining unique sample size rules for reliable inference in model selection.

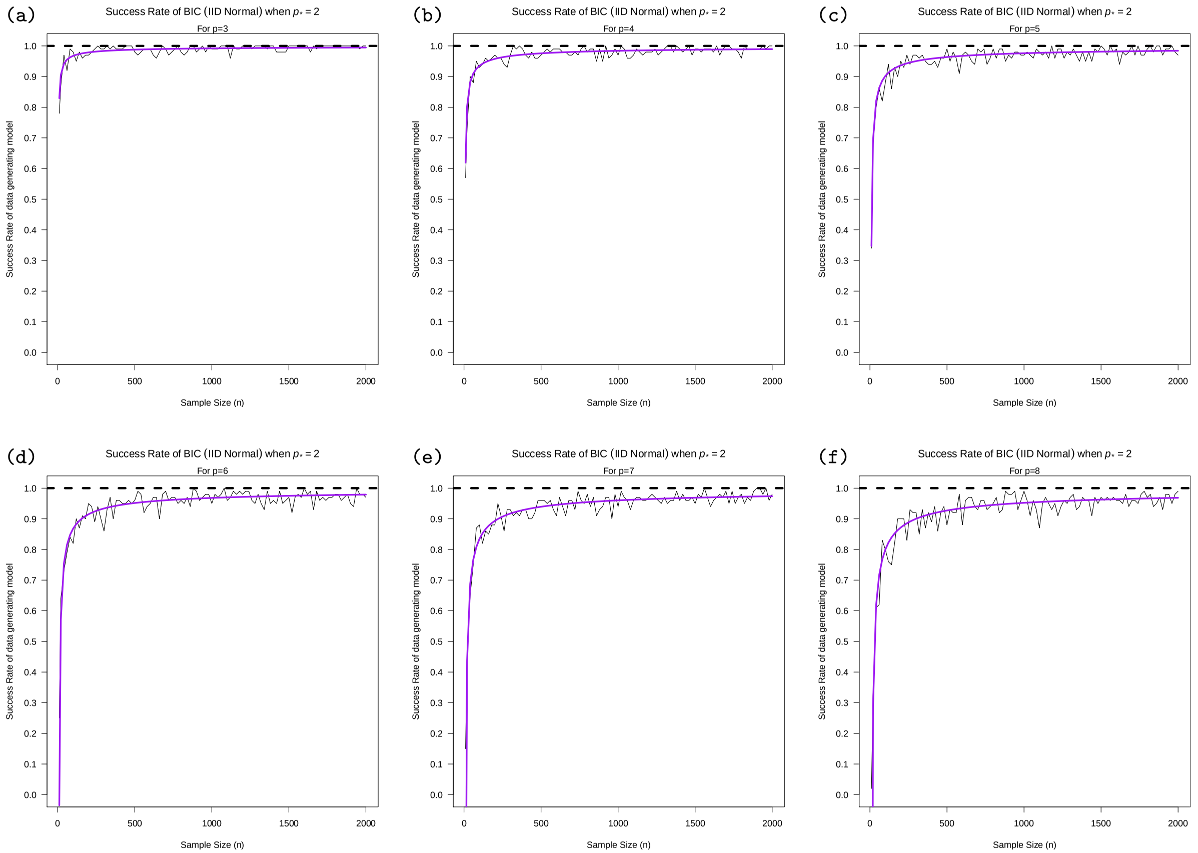
Appendix A: Supporting graphs of configuration considered in the simulation study

This appendix provides graphs corresponding to some additional settings considered in the simulation study. The graphs in figures 4–7 correspond to simulations where $p_* = 2$, $p = 3, \ldots, 8$, variance structure of **Correlated Normal** and **Correlated Gamma**, for AIC and BIC.



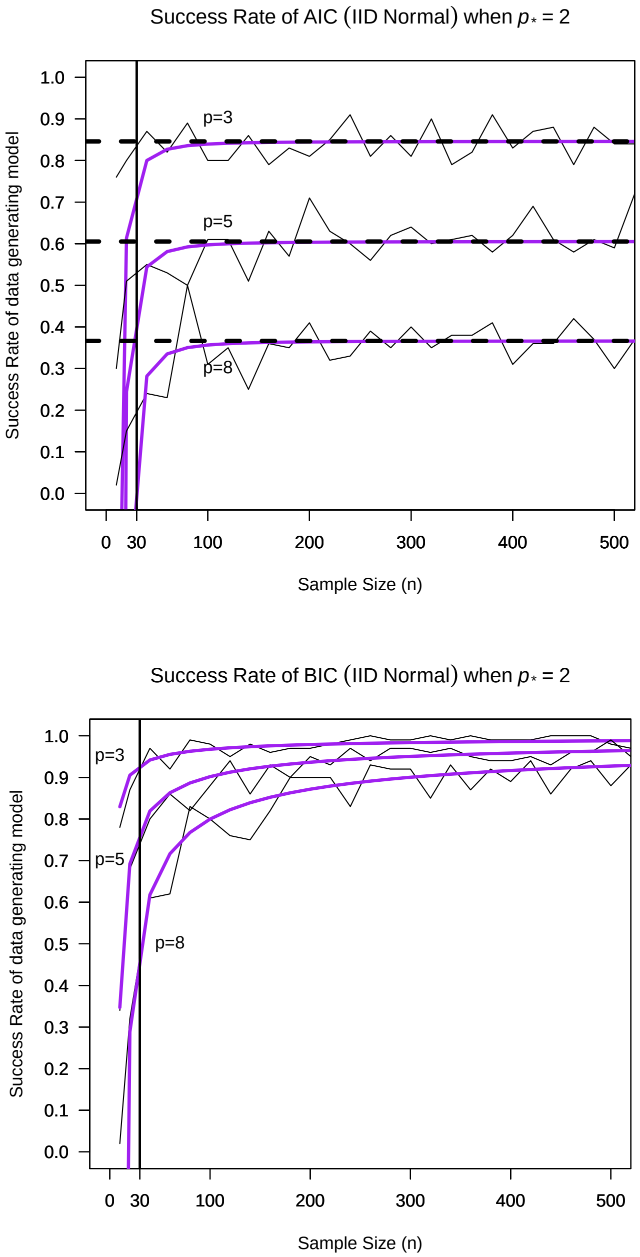
Legend: $\phi_{AIC}(p, p_*)$ – `Horizontal dotted line`; (Observed) $\hat{\phi}_n^{AIC}(p, p_*)$ – `Choppy solid curve`; (Proposed Mean Function) $\phi_n^{AIC}(p, p_*)$ – `Smooth solid curve`

FIG 1. *Observed and proposed mean AIC success rates when $p_* = 2$, $p = 3, 4, 5, 6, 7, 8$, and $\mathcal{F} =$* ***IID Normal***.



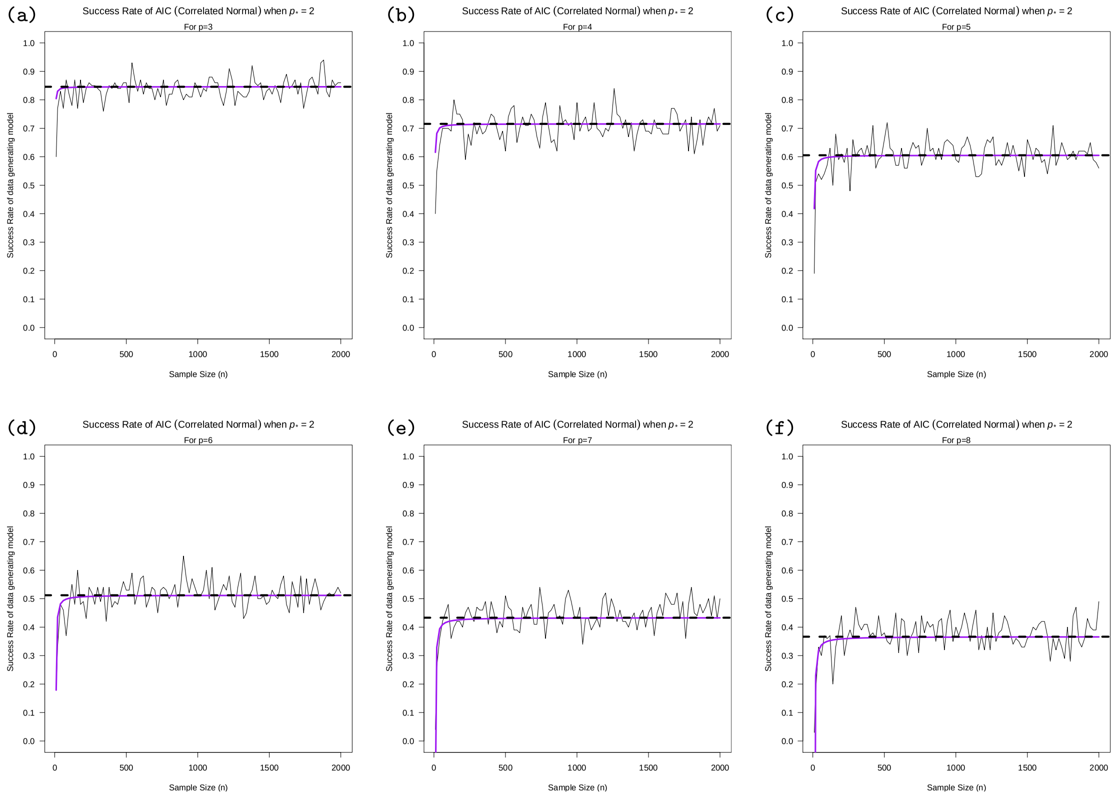
Legend: $\phi_{BIC}(p, p_*)$ – **Horizontal dotted line**; (Observed) $\hat{\phi}_n^{BIC}(p, p_*)$ – **Choppy solid curve**; (Proposed Mean Function) $\phi_n^{BIC}(p, p_*)$ – **Smooth solid curve**

FIG 2. *Observed and proposed mean BIC success rates when $p_* = 2$, $p = 3, 4, 5, 6, 7, 8$, and $\mathcal{F} =$* ***IID Normal****.*



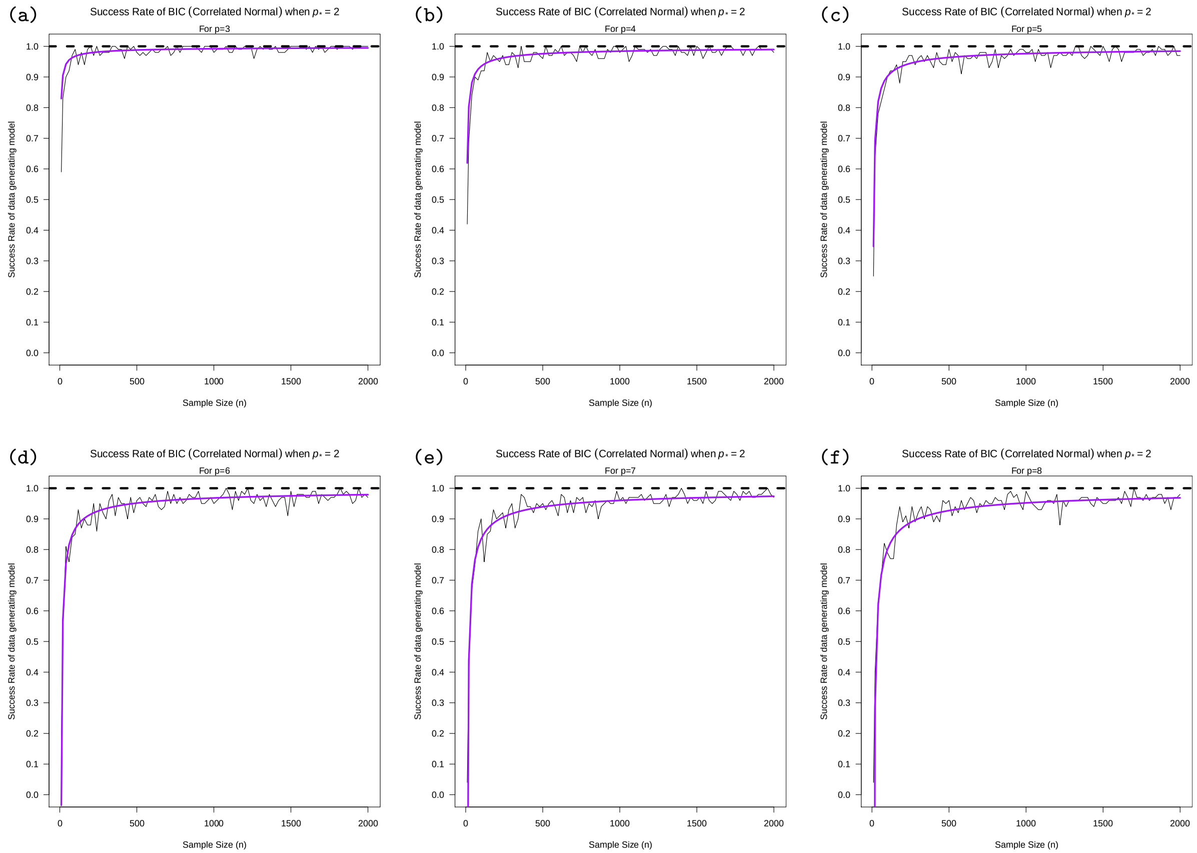
`Legend:` $\phi(p,p_*)$ – `Horizontal dotted line`; (Observed) $\hat{\phi}_n(p,p_*)$ – `Choppy solid curve`; (Proposed Mean Function) $\phi_n(p,p_*)$ – `Smooth solid curve`

FIG 3. *Observed and proposed mean (function) of success rates for AIC and BIC when* $p_* = 2$*,* $p = 3, 5, 8$*, and* $\mathcal{F} =$ ***IID Normal****.*



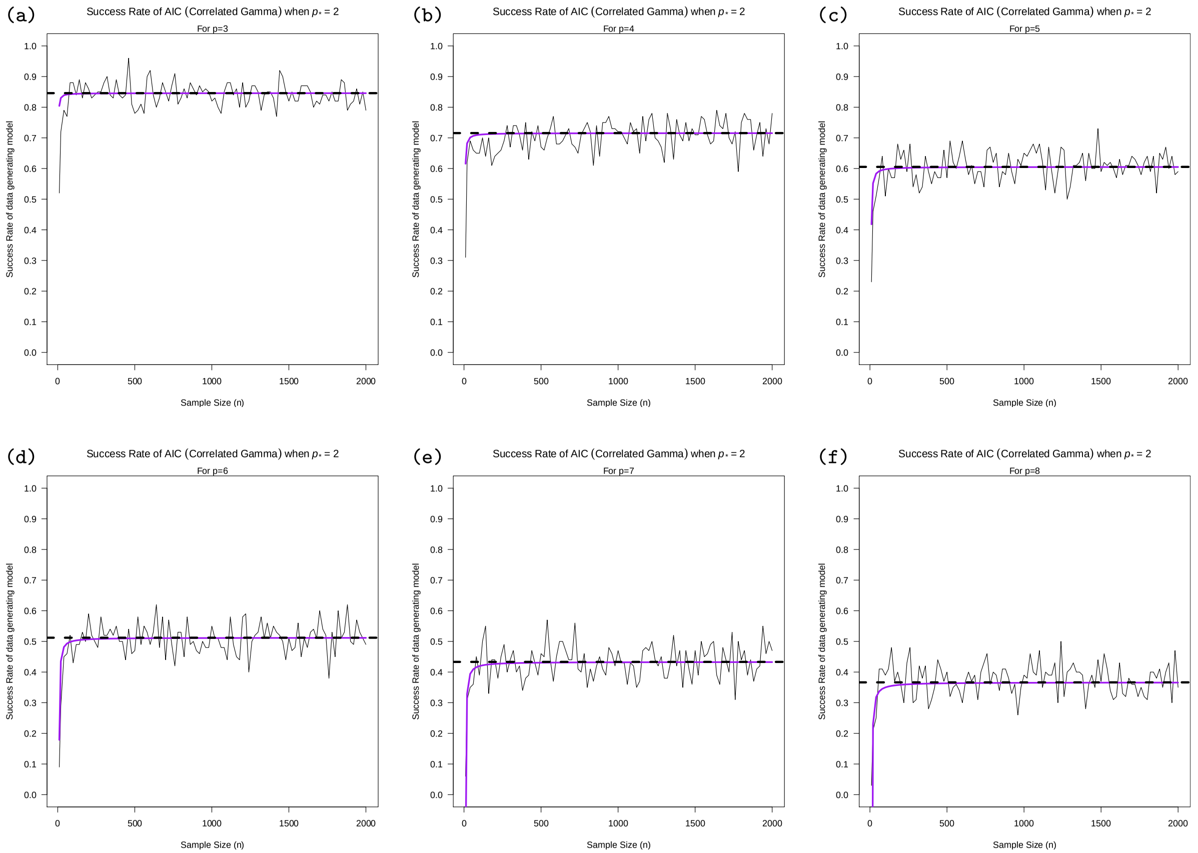
`Legend:` $\phi_{AIC}(p,p_*)$ – `Horizontal dotted line`; (Observed) $\hat{\phi}_n^{AIC}(p,p_*)$ – `Choppy solid curve`; (Proposed Mean Function) $\phi_n^{AIC}(p,p_*)$ – `Smooth solid curve`

FIG 4. *Observed & proposed mean AIC success rates when $p_* = 2$, $p = 3, 4, 5, 6, 7, 8$, and $\mathcal{F} =$ **Correlated Normal**.*



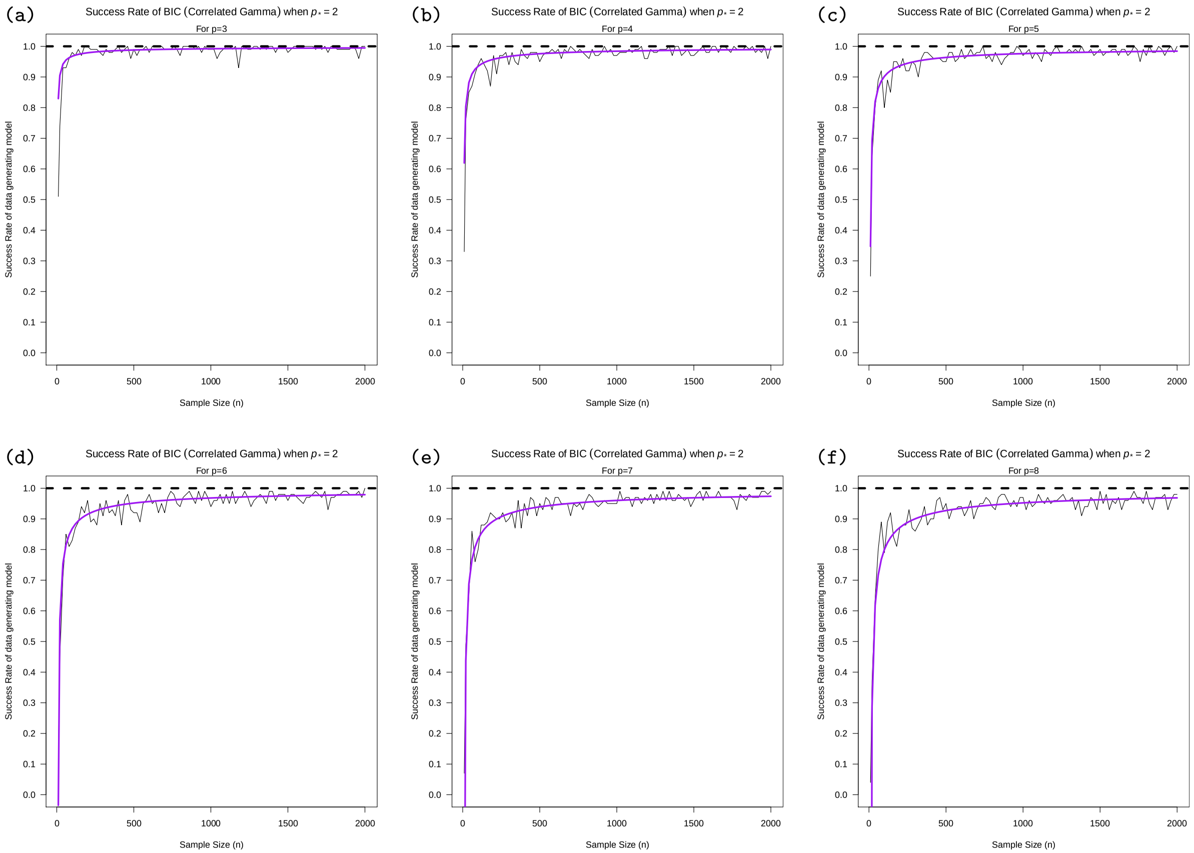
Legend: $\phi_{BIC}(p, p_*)$ – **Horizontal dotted line**; (Observed) $\hat{\phi}_n^{BIC}(p, p_*)$ – **Choppy solid curve**; (Proposed Mean Function) $\phi_n^{BIC}(p, p_*)$ – **Smooth solid curve**

FIG 5. *Observed & proposed mean BIC success rates when $p_* = 2$, $p = 3, 4, 5, 6, 7, 8$, and $\mathcal{F} =$* ***Correlated Normal****.*



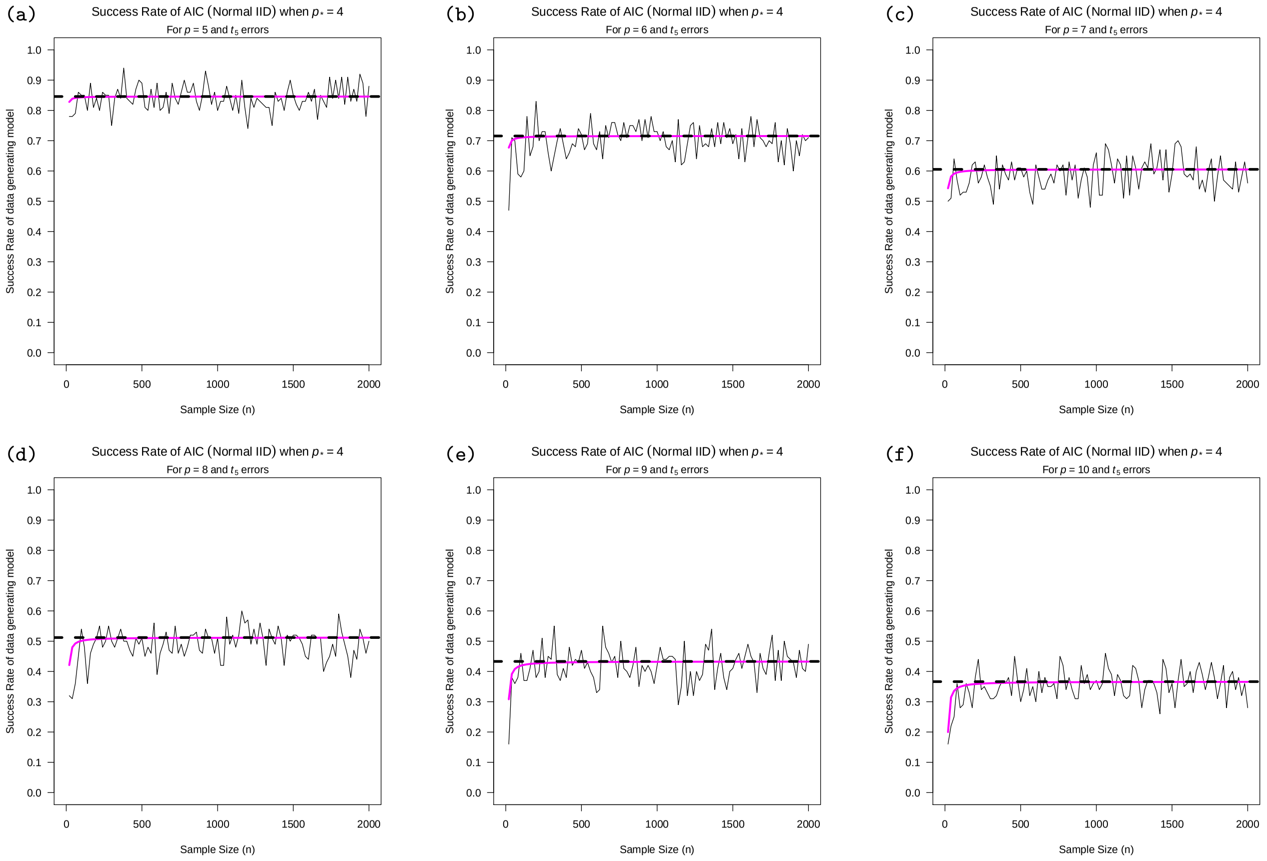
Legend: $\phi_{AIC}(p, p_*)$ – Horizontal dotted line; (Observed) $\hat{\phi}_n^{AIC}(p, p_*)$ – Choppy solid curve; (Proposed Mean Function) $\phi_n^{AIC}(p, p_*)$ – Smooth solid curve

FIG 6. *Observed & proposed mean AIC success rates when $p_* = 2$, $p = 3, 4, 5, 6, 7, 8$, and $\mathcal{F} =$ **Correlated Gamma**.*



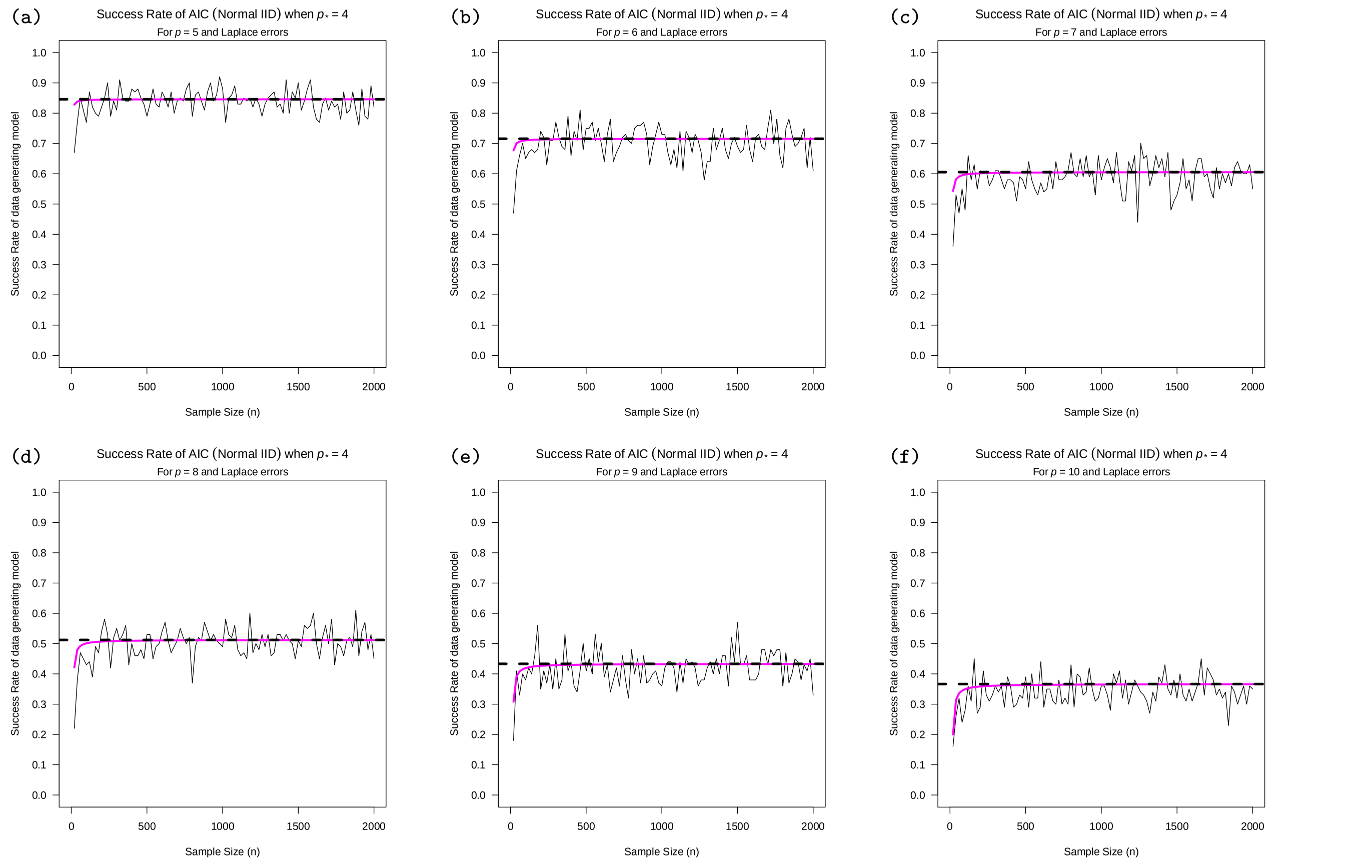
Legend: $\phi_{BIC}(p, p_*)$ – `Horizontal dotted line`; (Observed) $\hat{\phi}_n^{BIC}(p, p_*)$ – `Choppy solid curve`; (Proposed Mean Function) $\phi_n^{BIC}(p, p_*)$ – `Smooth solid curve`

FIG 7. *Observed & proposed mean BIC success rates when $p_* = 2$, $p = 3, 4, 5, 6, 7, 8$, and $\mathcal{F} =$ **Correlated Gamma**.*



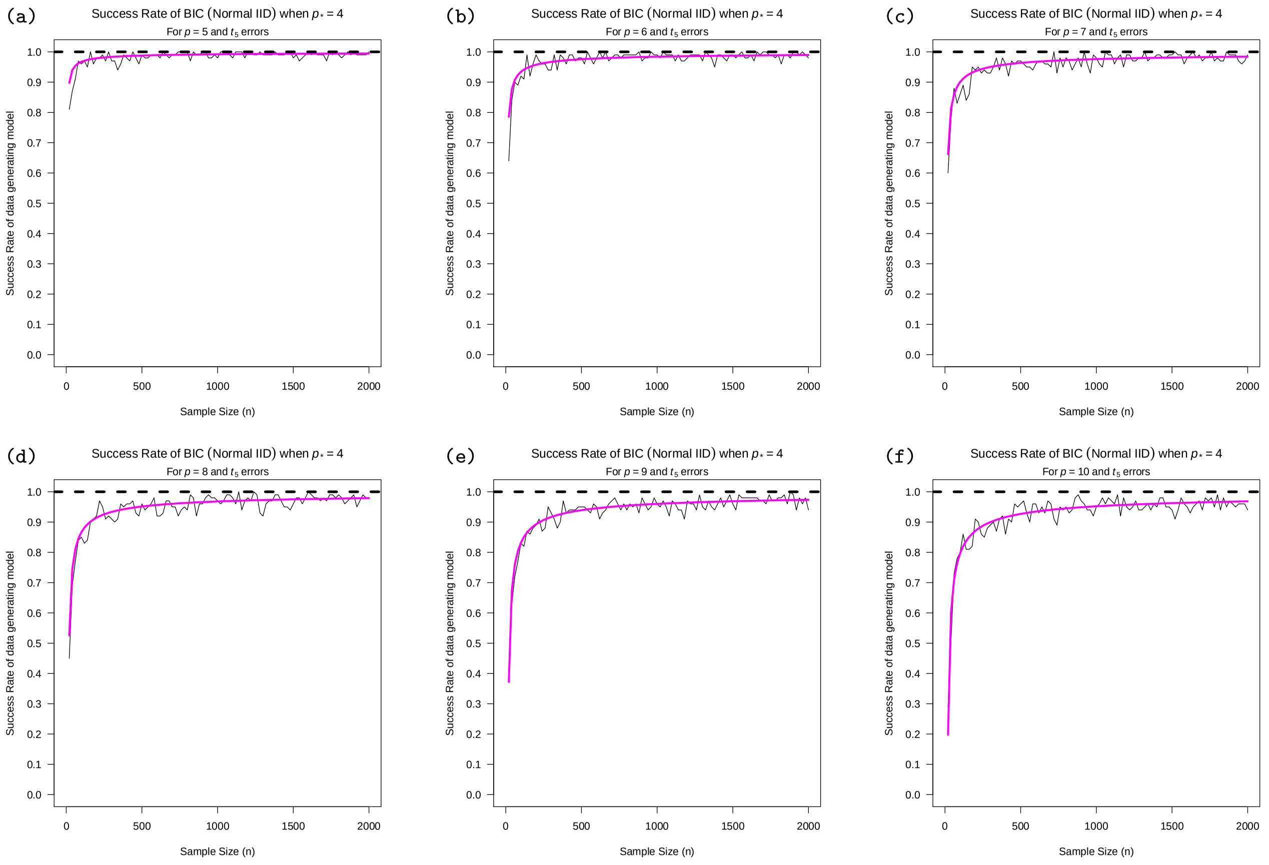
Legend: $\phi_{AIC}(p, p_*)$ – Horizontal dotted line; (Observed) $\hat{\phi}_n^{AIC}(p, p_*)$ – Choppy solid curve; (Proposed Mean Function) $\phi_n^{AIC}(p, p_*)$ – Smooth solid curve

Fig 8. *Observed & proposed mean AIC success rates when $p_* = 4$, $p = 5, \ldots, 10$, $\mathcal{F} =$ **IID Normal**, & errors from t_5.*



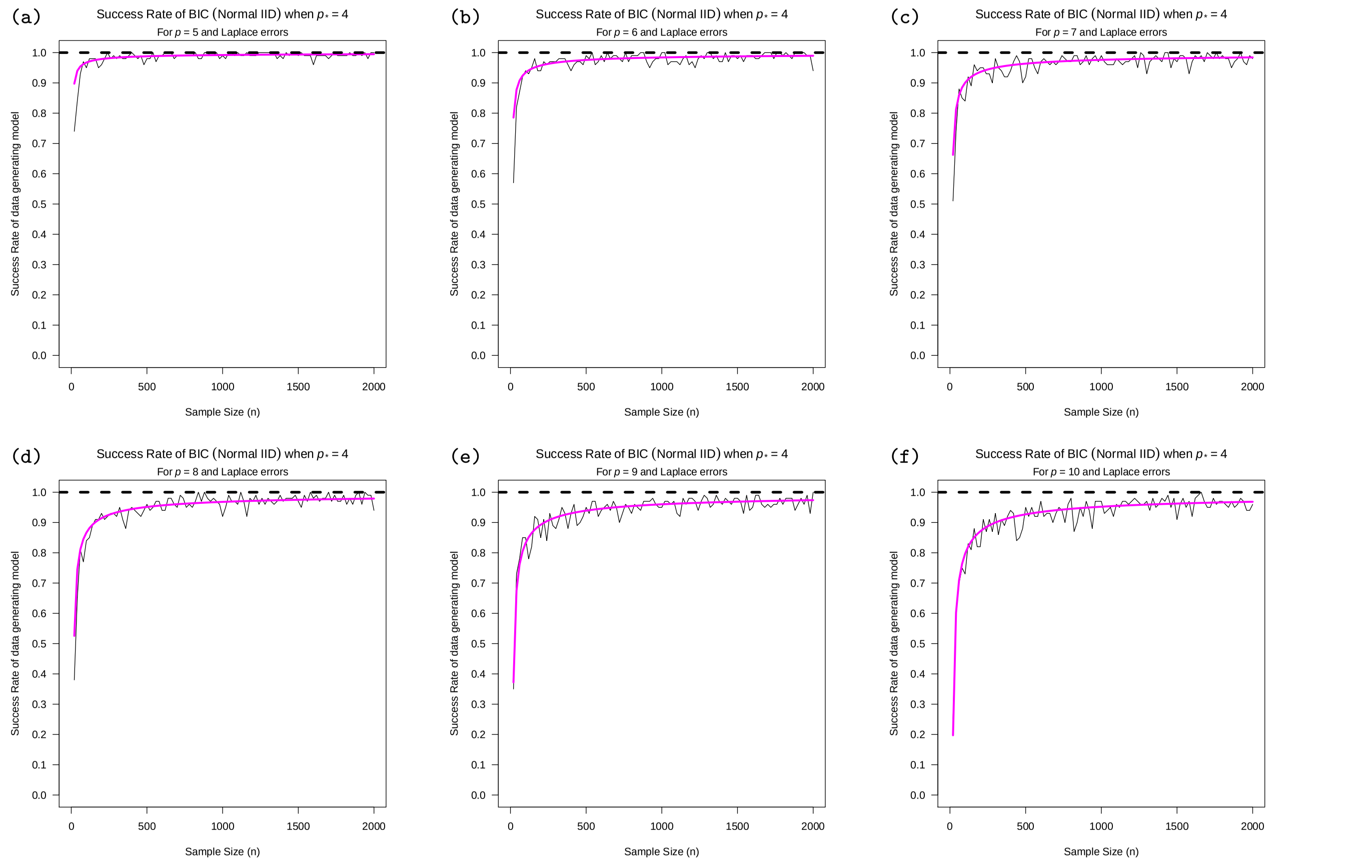
`Legend`: $\phi_{AIC}(p, p_*)$ – `Horizontal dotted line`; (Observed) $\hat{\phi}_n^{AIC}(p, p_*)$ – `Choppy solid curve`; (Proposed Mean Function) $\phi_n^{AIC}(p, p_*)$ – `Smooth solid curve`

FIG 9. *Observed and proposed mean AIC success rates when $p_* = 4$, $p = 5, \ldots, 10$, $\mathcal{F} =$ **IID Normal**, and errors from Laplace(0,1).*



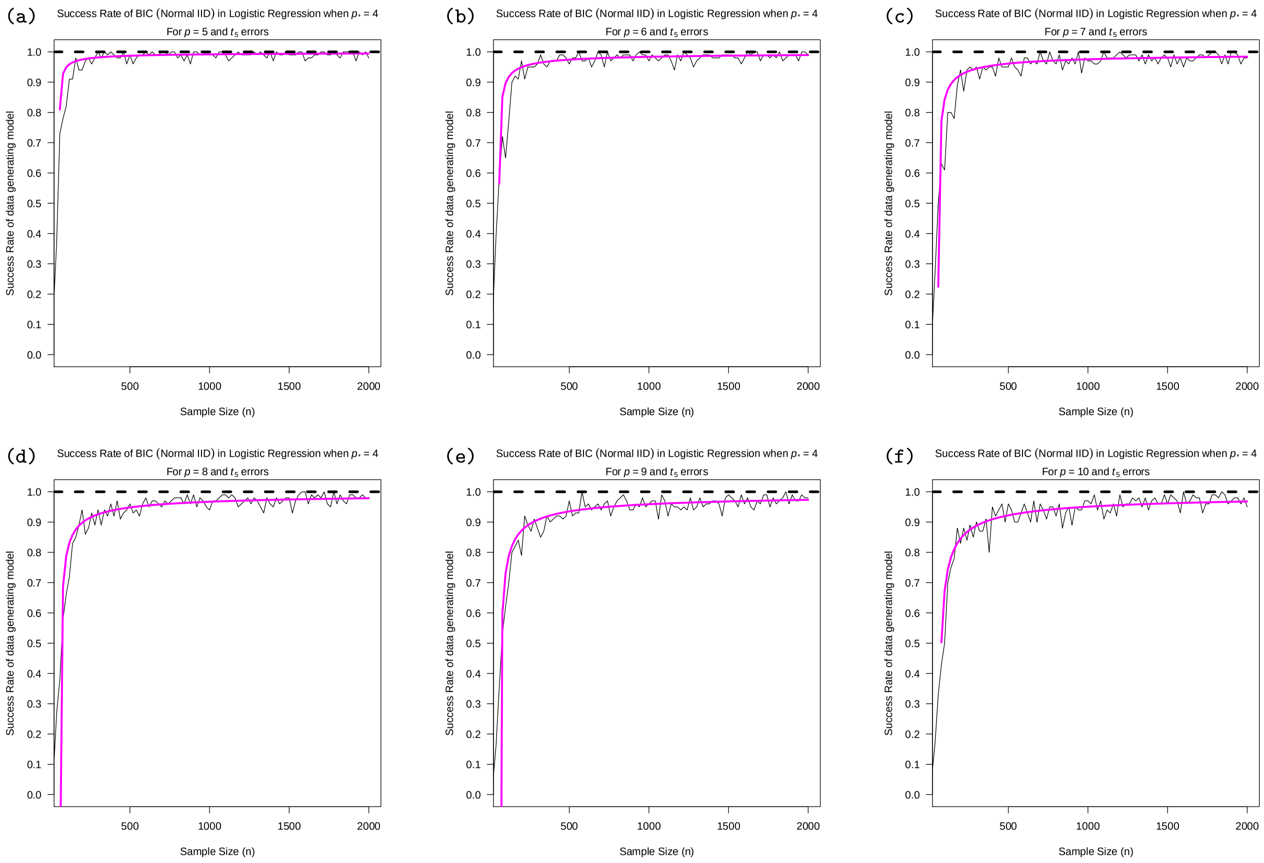
Legend: $\phi_{BIC}(p, p_*)$ – Horizontal dotted line; (Observed) $\hat{\phi}_n^{BIC}(p, p_*)$ – Choppy solid curve; (Proposed Mean Function) $\phi_n^{BIC}(p, p_*)$ – Smooth solid curve

Fig 10. *Observed and proposed mean BIC success rates when $p_* = 4$, $p = 5, \ldots, 10$, $\mathcal{F} =$ **IID Normal**, and errors from t_5.*



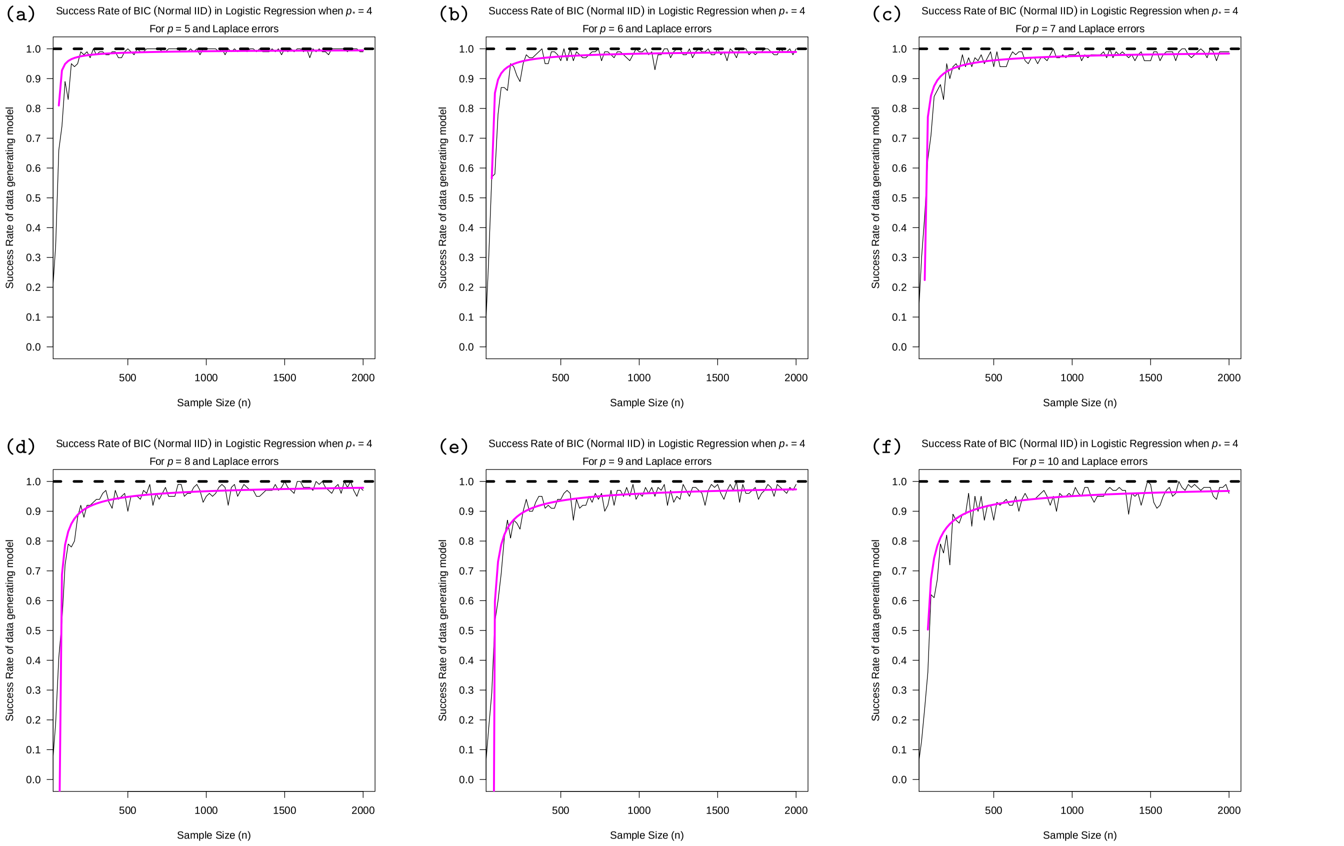
Legend: $\phi_{BIC}(p, p_*)$ – Horizontal dotted line; (Observed) $\hat{\phi}_n^{BIC}(p, p_*)$ – Choppy solid curve; (Proposed Mean Function) $\phi_n^{BIC}(p, p_*)$ – Smooth solid curve

FIG 11. *Observed and proposed mean BIC success rates when $p_* = 4$, $p = 5, \ldots, 10$, $\mathcal{F} =$ **IID Normal**, and errors from Laplace(0,1).*



Legend: $\phi_{BIC}(p, p_*)$ – `Horizontal dotted line`; (Observed) $\hat{\phi}_n^{BIC}(p, p_*)$ – `Choppy solid curve`; (Proposed Mean Function) $\phi_n^{BIC}(p, p_*)$ – `Smooth solid curve`

FIG 12. *Observed & proposed mean BIC success rates when $p_* = 4$, $p = 5, \ldots, 10$, $\mathcal{F} =$ **IID Normal** in Logistic regression with outliers under t_5.*



Legend: $\phi_{BIC}(p, p_*)$ – `Horizontal dotted line`; (Observed) $\hat{\phi}_n^{BIC}(p, p_*)$ – `Choppy solid curve`; (Proposed Mean Function) $\phi_n^{BIC}(p, p_*)$ – `Smooth solid curve`

FIG 13. *Observed & proposed mean BIC success rates for* $p_* = 4$, $p = 5, \ldots, 10$, $\mathcal{F} =$ ***IID Normal*** *in Logistic regression with Laplace(0,1) outliers.*

Appendix B: Sample size and reliability table for application in design of experiments and observational studies

Included in this appendix are tables 1 and 2 for different reliability values ($\mathscr{R}$), the *minimum sample size* required for the furthest (and most difficult) model to be detected as the best model by BIC and AIC, respectively, if the respective model is indeed the true model. Other such tables for AIC and BIC were constructed for different values of p_* and are available on request.

Table 1

For different reliability values ($\mathscr{R}$), the ***minimum sample size*** *required for the furthest (and most difficult) model $\mathcal{M}_*$ (of dimension $p_* = 2$) to be detected as the best model by BIC, if it is indeed the true model*

p	Reliability of at least $\mathscr{R}$							
	0.70	0.75	0.80	0.85	0.90	0.95	0.98	0.99
3	7	8	9	12	19	51	218	683
4	13	16	20	29	52	153	684	2160
5	21	26	35	53	98	296	1340	4241
6	31	40	54	82	155	475	2162	6847
7	43	55	76	117	222	687	3133	9930
8	56	72	101	157	299	929	4244	13454
9	71	92	128	201	385	1199	5486	17393
10	86	113	159	250	479	1497	6851	21727
11	104	136	192	302	582	1820	8336	26438
12	122	161	227	359	692	2168	9935	31512
13	142	187	265	419	810	2540	11644	36935
14	163	215	305	483	935	2935	13460	42698
15	185	245	347	551	1067	3353	15379	48791
16	208	276	392	622	1206	3792	17400	55204
17	232	308	439	697	1352	4253	19519	61930
18	258	342	487	775	1505	4735	21735	68961
19	284	377	538	857	1664	5237	24045	76292
20	311	414	591	941	1829	5760	26447	83917
21	339	452	646	1029	2000	6302	28940	91829
22	369	491	702	1120	2178	6863	31522	100023
23	399	532	761	1213	2361	7444	34191	108496
24	430	574	821	1310	2551	8043	36946	117242
25	462	617	883	1410	2746	8661	39787	126257
26	495	661	947	1513	2947	9296	42710	135537
27	529	707	1013	1618	3153	9950	45716	145078
28	564	754	1080	1727	3366	10621	48804	154877
29	600	802	1150	1838	3583	11310	51971	164931
30	636	851	1220	1952	3806	12015	55218	175237
31	674	902	1293	2069	4035	12738	58542	185790
32	712	953	1367	2188	4268	13478	61945	196589
33	751	1006	1443	2310	4507	14234	65423	207631
34	791	1060	1521	2435	4751	15006	68977	218912
35	832	1114	1600	2562	5000	15795	72606	230432

Table 2

For different reliability values ($\mathscr{R}$), the ***minimum sample size*** *required for the furthest (and most difficult) model $\mathcal{M}_*$ (of dimension $p - d_{max}$) to be detected as the best model by AIC, if it is indeed the true model*

p	Reliability of at least $\mathscr{R}$								
	$(d_{\max}$ at $\mathscr{R})$								
	0.40	0.45	0.50	0.55	0.60	0.65	0.70	0.75	0.80
	(5)	(4)	(4)	(3)	(3)	(2)	(2)	(1)	(1)
2	—	—	—	—	—	—	—	6	9
3	—	—	—	—	—	12	37	7	10
4	—	—	—	19	147	13	38	8	11
5	—	23	90	20	148	14	39	9	12
6	45	24	91	21	149	15	40	10	13
7	46	25	92	22	150	16	41	11	14
8	47	26	93	23	151	17	42	12	15
9	48	27	94	24	152	18	43	13	16
10	49	28	95	25	153	19	44	14	17
11	50	29	96	26	154	20	45	15	18
12	51	30	97	27	155	21	46	16	19
13	52	31	98	28	156	22	47	17	20
14	53	32	99	29	157	23	48	18	21
15	54	33	100	30	158	24	49	19	22
16	55	34	101	31	159	25	50	20	23
17	56	35	102	32	160	26	51	21	24
18	57	36	103	33	161	27	52	22	25
19	58	37	104	34	162	28	53	23	26
20	59	38	105	35	163	29	54	24	27
21	60	39	106	36	164	30	55	25	28
22	61	40	107	37	165	31	56	26	29
23	62	41	108	38	166	32	57	27	30
24	63	42	109	39	167	33	58	28	31
25	64	43	110	40	168	34	59	29	32
26	65	44	111	41	169	35	60	30	33
27	66	45	112	42	170	36	61	31	34
28	67	46	113	43	171	37	62	32	35
29	68	47	114	44	172	38	63	33	36
30	69	48	115	45	173	39	64	34	37
31	70	49	116	46	174	40	65	35	38
32	71	50	117	47	175	41	66	36	39
33	72	51	118	48	176	42	67	37	40
34	73	52	119	49	177	43	68	38	41
35	74	53	120	50	178	44	69	39	42

Note: Since $p \nleq d_{\max}$, the corresponding cells are denoted with "—".

Acknowledgements

We would like to thank the editor, associate editor, and referees for their positive comments and suggestions that greatly improved the quality of this paper. This project was partially supported by Award Number K01MH087219 from the

National Institute of Mental Health. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of Mental Health or the National Institutes of Health.

References

Akaike, H. (1974). A new look at statistical model identification. *IEEE Transactions on Automatic Control 19* **6** 716–723. MR0423716

Aron, A. and Aron, E. N. (2003). *Statistics for Psychology*. Prentice Hall/Pearson Education.

Burnham, K. P. and Anderson, D. R. (2002). *Model Selection and Multimodel Inference: A Practical-Theoretic Approach*, 2 ed. Springer, New York. MR1919620

Burnham, K. P. and Anderson, D. R. (2004). Multimodel inference: understanding AIC and BIC in model selection. *Sociological Methods and Research* **33** 261–304. MR2086350

Cavanaugh, J. (1997). Unifying the derivations of the Akaike and corrected Akaike information criteria. *Statistics and Probability Letters* **33** 201–208. MR1458291

Cetin, M. C. and Erar, A. (2002). Variable selection with Akaike information criteria: a comparative study. *Hacettepe Journal of Mathematics and Statistics* **31** 89–97. MR1987060

Claeskens, G. and Hjort, N. L. (2003). The focused information criterion. *Journal of the American Statistical Association* **98** 900–916. MR2041482

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciencies*. Routledge.

Cohen, J. (1992). A power primer. *Psychological Bulletin* **112** 155.

Cohen, J. and Cohen, P. (1975). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*. Lawrence Erlbaum.

Cramér, H. (1946). A contribution to the theory of statistical estimation. *Scandinavian Actuarial Journal* **1946** 85–94. MR0017505

Green, S. B. (1991). How many subjects does it take to do a regression analysis? *Multivariate Behavioral Research* **26** 499–510.

Hannan, E. J. and Quinn, B. G. (1979). The determination of the order of an autoregression. *Journal of the Royal Statistical Society. Series B (Methodological)* **41** 190–195. MR0547244

Hurvich, C. M. and Tsai, C. L. (1989). Regression and time series model selection in small samples. *Biometrika* **76** 297–307. MR1016020

Hurvich, C. M. and Tsai, C. L. (1990). The impact of model selection on inference in linear regression. *The American Statistician* **44** 214–217.

Hurvich, C. M. and Tsai, C.-L. (1995). Relative rate of convergence for efficinet model selection criteria in linear regression. *Biometrika* **82** 418–425. MR1354238

Kadane, J. B. and Lazar, N. A. (2004). Methods and criteria for model selection. *Journal of the American Statistical Association* **99** 279–290. MR2061890

Kass, R. E. and Raftery, A. E. (1995). Bayes factor. *Journal of the American Statistical Association* **90** 773–795.

Kelley, K. and Maxwell, S. E. (2003). Sample size for multiple regression: obtaining regression coefficients that are accurate, not simply significant. *Psychological Methods* **8** 305.

Neath, A. A. and Cavanaugh, J. E. (1997). Regression and time series model selection using variants of the Schwarz information criterion. *Communications in Statistics* **26** 559–580. MR1436288

Nishii, R. (1984). Asymptotic properties of criteria for selection of variables in multiple regression. *Annals of Statistics* **12** 758–765. MR0740928

Rao, C. R. (1945). Information and accuracy attainable in the estimation of statistical parameters. *Bulletin of Cal. Math. Soc.* **37** 81–91. MR0015748

Rao, C. R. and Wu, Y. (2001). Model selection. *Lecture Notes-Monograph Series* **38** 1–64. MR2000751

Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics* **6** 461–464. MR0468014

Shi, P. and Tsai, C.-L. (2002). Regression model selection—a residual likelihood approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **64** 237–252. MR1904703

Shibata, R. (1981). An optimal selection of regression variables. *Biometrika* **68** 45–54. MR0614940

Spiegelhalter, D. J., Best, N., Carlin, B. P. and Van der Linde, A. (1998). Bayesian deviance, the effective number of parameters, and the comparison of arbitrarily complex models Technical Report, Research Report, 98-009.