Model Identification in Linear Fixed Effects Models

Ondrej Blaha* Julia Volaufova* Lynn R. LaMotte*

Abstract

Our study focuses on numerical investigation of performances of current existing variable selection techniques incorporating statistics like adjusted R^2, AIC, BIC, or SBC for linear models. Specifically, we focus on the ability of these statistics to detect a true model among all possible sub-models. Furthermore, we explore the dependence of the successful true model detection on the parameter setting. Simulation studies were designed to investigate properties of detection of the true model among all possible models. Results provide a new perspective on the current, commonly used techniques. The consequences of the results are discussed as well.

Key Words: Variable selection, Model identification, Information criteria, True model

1. Introduction

Model building and variable selection specifically are prominent parts of statistical modeling, which have been attracting extreme attention over the last few decades. The fact that variable selection in both, fixed as well as mixed effect models is a very hot topic (and probably always will be) with new findings and clarifications can be documented by the amount of recent publications regarding this topic. In this study we focus on identification of the true model (set of covariates) in the multiple linear fixed effects models. We consider several commonly used criteria for variable selection, using optimal all-possible-submodels and investigate performance of these criteria given several parameter setting scenarios.

2. Methods

In this paper we consider standard fixed effects linear regression model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

with $\mathbf{Y}$ being a response vector for all n individual sampling units which are independent of each other. $\mathbf{X}$ is a $n \times p$ design matrix, $\boldsymbol{\beta}$ is a vector of unknown parameters of length p, and $\boldsymbol{\epsilon}$ corresponds to a vector of random errors with the normal distribution $\boldsymbol{\epsilon} \sim N(0, \sigma^2 \boldsymbol{I})$. All objects in the equation are of the dimension n denoting the number of the independent sampling units. We assume that the columns of the matrix $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_p)$ were centered and scaled (have zero mean and variance equals to 1).

Criteria considered for the model identification are adjusted R^2, Akaike information criterion (AIC) first introduced in [1], corrected Akaike information criterion (AIC_c) introduced in [9], Sawa's Bayesian information criterion (BIC) described in [7], Schwarz's Bayesian information criterion (SBC) from [8], and Mallow's C_p introduced in [5]. Just for the clarity purposes we report the definitions of these criteria as they were used in simulation study:

$$R^2_{Adj.} = 1 - \frac{(n-1)\frac{SSE}{SST}}{n-k},$$

*LSU Health Sciences Center, 2020 Gravier Street, New Orleans, LA 70112

$$AIC = n\log\left(\frac{SSE}{n}\right) + 2k,$$

$$AIC_c = AIC + \frac{2k(k+1)}{n-k-1},$$

$$BIC = n\log\left(\frac{SSE}{n}\right) + 2(k+2)\frac{n\hat{\sigma}^2}{SSE} - 2\left(\frac{n\hat{\sigma}^2}{SSE}\right)^2,$$

$$SBC = n\log\left(\frac{SSE}{n}\right) + k\log(n),$$

$$C_p = \frac{SSE}{\hat{\sigma}^2} - n + 2k,$$

where $SSE = \left(\mathbf{Y} - \hat{\mathbf{Y}}\right)'\left(\mathbf{Y} - \hat{\mathbf{Y}}\right)$, $SST = \left(\mathbf{Y} - \bar{Y}\mathbf{1}_n\right)'\left(\mathbf{Y} - \bar{Y}\mathbf{1}_n\right)$, and $\hat{\sigma}^2$ is the variance estimate from the full model (i.e. $k = p$). In the formulas, $\hat{\mathbf{Y}}$ represents fitted values, $\mathbf{1}_n$ is a n-dimensional vector of ones, and $\bar{Y} = \frac{1}{n}\mathbf{1}_n'\mathbf{Y}$.

3. Simulation study

The purpose of this set of simulations is to explore potency of the currently existing variable selection criteria to detect a true sparsity pattern. Next, our goal is to explore dependency of model identification on the parameter configuration. Such parameters are: number of non-zero coefficients k in the underlying true model, variance σ^2, magnitude of the parameter coefficients $\boldsymbol{\beta}$, and sample size n. In our case we have a set of $p = 10$ potential explanatory variables, from which (after normalization of columns of the design matrix $\mathbf{X}$) the true model is determined setting two or seven ($k \in \{2, 7\}$) of the regression coefficients to given non-zero values. For each k, there were two vectors of parameters selected $\boldsymbol{\beta}_{k;1},\ \boldsymbol{\beta}_{k;2}$ with appropriate number of non-zero coefficients which differed in magnitude, i.e. $\left\|\boldsymbol{\beta}_{k;1}\right\| \ll \left\|\boldsymbol{\beta}_{k;2}\right\|$ (see Table 3.1). Two different values of variance were selected from a logarithmic grid as $\sigma^2 \in \{0.01,\ 1\}$. Sample sizes are $n \in \{20,\ 50,\ 100,\ 300\}$. For each sample size one design matrix $\mathbf{X}$ was generated from multivariate normal distribution with covariance between the columns of $\rho = 0.3$. For all configurations number of the repetition was held fixed at $1,000$.

It is important to mention that all the values were selected from much larger pole of parameters we considered and used for the actual simulation and they represent rather extreme values in order to support our findings in the most informative way. Results from the other parameter settings fall along the way of the presented selection.

For each combination of the parameter configuration (listed in Table 3.1), there were $1,000$ responses $\mathbf{Y}_1, \ldots, \mathbf{Y}_{1,000}$ generated from a normal distribution $N(\mathbf{X}_s\boldsymbol{\beta}_{k;i}, \sigma^2\boldsymbol{I})$, where the notation $\mathbf{X}_s$ explicitly expresses the fact that the centered and scaled design matrix was used for data generation process and $\boldsymbol{\beta}_{k;i}$ represents one of the vector of coefficients listed in the Table 3.1. Finally, each of these responses was fitted to all $2^p - 1 = 1023$ possible models and values of the selected criteria were recorded. For each criterion, minimum/maximum value was found and the model at which the minimum/maximum was attained was identified. Only one of the 1023 possible models (intercept model only is excluded) represents the true model. The true model selection by the given criterion was recorded and the success rate investigated.

All simulations were built and executed in proc IML of SAS 9.4.

k	$k=2$		$k=7$	
$\boldsymbol{\beta}$	$\boldsymbol{\beta}_{2;1}$	$\boldsymbol{\beta}_{2;2}$	$\boldsymbol{\beta}_{7;1}$	$\boldsymbol{\beta}_{7;2}$
	$\begin{pmatrix} -1 \\ 2.2 \\ 0 \\ . \\ . \\ . \\ 0 \end{pmatrix}$	$\begin{pmatrix} -4 \\ 3.5 \\ 0 \\ . \\ . \\ . \\ 0 \end{pmatrix}$	$\begin{pmatrix} -1 \\ 0.85 \\ 1.6 \\ 2.2 \\ 1.25 \\ 0.76 \\ 1.1 \\ 0 \\ 0 \\ 0 \end{pmatrix}$	$\begin{pmatrix} -4 \\ 3.5 \\ 3.2 \\ 4.5 \\ 4.1 \\ 3.1 \\ 3.7 \\ 0 \\ 0 \\ 0 \end{pmatrix}$
Criteria considered: $\sigma^2 \in \{0.01,\ 1\}$				
Sample size: $n \in \{20,\ 50,\ 100,\ 300\}$				
Repetitions: $r = 1,000$				

Table 3.1: All parameters involved in the data generation process.

4. Results:

The most interesting conclusions, which can be drawn from the results of the simulation study are illustrated by graphs. Graphs 4.1 and 4.2 reflect the true model identification as a function of sample size. Each panel shows the frequency of the true model detection by each criterion for a different parameter configuration. Figure 4.1 illustrate the situation when $k = 2$ and Figure 4.2 situation when $k = 7$. Individual panels within each figure differ in values of σ^2 and $\boldsymbol{\beta}$. Finally, each line represents the detection rate of the true set of covariates for one criterion and consists of connected values (dots) for each value of sample size. There are several conclusion which can be drawn from these results. However, only two points can be made generally.

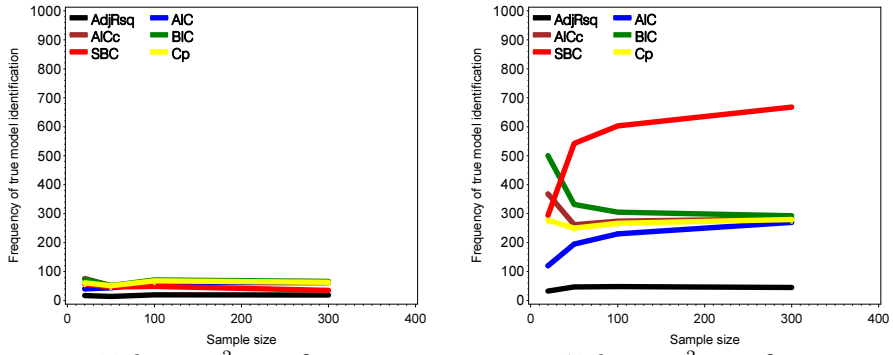
- The detection rate of the true sparsity pattern highly depends on the parameter configuration and is very sensitive to a change in most of them (except for k).
- There is not a single situation, configuration, or setting when one criterion would dominate the others.

Although the exact behavior of criteria with respect to different parametrization is certainly not well known, the first finding in such a general form seems to be anticipated. The punchline here is that there really cannot be concluded more on the very general level and the findings need to be broken down by every parameter. While the first claim might not be so surprising or unexpected to most people based on the knowledge of the statistical modeling the second finding certainly is. One of the misunderstandings is that Akaike information criterion tends to under-fit the model and therefore should perform better in the situation when k is small while Sawa's Bayesian information criterion (BIC) should perform better for larger values of k. This common misunderstanding is shown not correct by our simulations. However, one recommendation is clear from the results and further supports the theoretical reasoning of Burnham and Anderson in [3]. Small sample correction in corrected Akaike's criterion (AIC_c) truly improves the performance for situation when sample size is low and with increasing sample size the effect of this correction term diminishes and AIC_c converges to the uncorrected AIC. We often fail to recognize this simple fact and the use of uncorrected AIC over the corrected version is therefore unjustified.

Number of covariates k: One can see that the model identification does not really depend on the number of true covariates considered. The profiles for each criterion hardly changed when increasing value of k from 2 to 7.

Sample size n: There is a noticeable effect of sample size where for some criteria, in certain situations the recognition is actually smaller using larger sample size (AIC_c, BIC). The only monotonically increasing function is Schwarz Bayesian Criterion (SBC). This finding is consistent with the literature, specifically with the proven consistency property of this criterion (for further details see [6]). However, even this consistency property fails for some 'unfavorable' cases such as show in the top right panel of Figure 4.2. The key to this point is well explained in [3] which states that Schwarz Bayesian criterion converges to a so called quasi-true model which, apparently, does not have to be the set of preselected variables (true set of covariates).

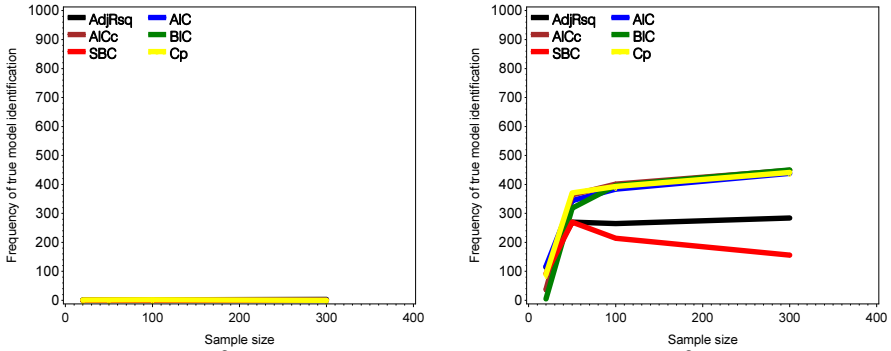
Variance σ^2 and Magnitude of β: It is expected that with increase in $\|\boldsymbol{\beta}\|$ and/or decrease in variance σ^2 any given criterion will have an easier way to detect the true set of covariates. These two parameters truly work together and therefore it makes more sense and investigate them as the signal/noise ratio $\frac{\|\boldsymbol{\beta}\|}{\sigma}$. The higher this value the better recognition we get. However, this is true only to some extent. Improvement beyond certain point is not possible and therefore, for each individual setting (and finite sample size), the detection profiles converge to the situation similar to the ones in the bottom right panels of Figures 4.1 and 4.2. Further increase of $\|\boldsymbol{\beta}\|$ nor decrease in σ^2 will not improve the performance of these criteria.



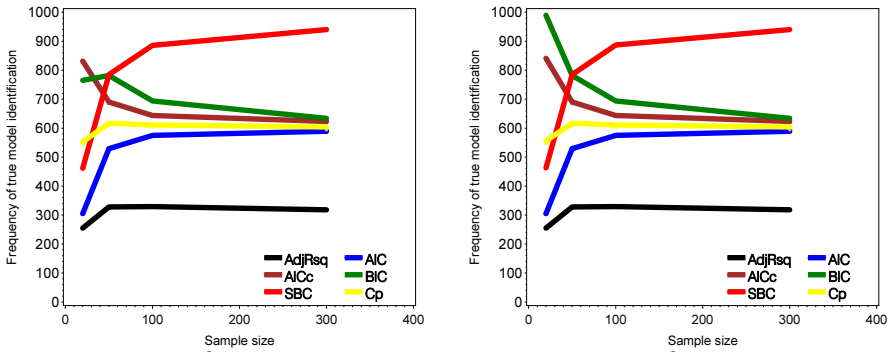
(a) $k = 2,\ \sigma^2 = 1,\ \boldsymbol{\beta}_{2;1}$ (b) $k = 2,\ \sigma^2 = 1,\ \boldsymbol{\beta}_{2;2}$

(c) $k = 2,\ \sigma^2 = 0.01,\ \boldsymbol{\beta}_{2;1}$ (d) $k = 2,\ \sigma^2 = 0.01,\ \boldsymbol{\beta}_{2;2}$

Figure 4.1: Graphs representing the detection rate of the true set of covariates for $k = 2$. Each line corresponds to one of the criteria and consists of connected values of detection for preselected sample sizes.



(a) $k = 7,\ \sigma^2 = 1,\ \beta_{7;1}$ (b) $k = 7,\ \sigma^2 = 1,\ \beta_{7;2}$



(c) $k = 7,\ \sigma^2 = 0.01,\ \beta_{7;1}$ (d) $k = 7,\ \sigma^2 = 0.01,\ \beta_{7;2}$

Figure 4.2: Graphs representing the detection rate of the true set of covariates for $k = 7$. Each line corresponds to one of the criteria and consists of connected values of detection for preselected sample sizes.

5. Conclusion

In the presented simulation study we have shown that the quality of model identification in the linear fixed effect models highly differ with the criterion used. The performance of each criterion further depends on the actual parameter setting including the value of σ^2, $\|\boldsymbol{\beta}\|$, and sample size. Detection rate of a true model by each criterion is very sensitive to a change in any of the mentioned parameters. There is no situation or parameter setting known to us where one criterion always outperforms the others and therefore no recommendation regarding the use of the criteria in specific situations can be made.

REFERENCES

Akaike, H. (1973), "Information Theory and an Extension of the Maximum Likelihood Principle," in *Second International Symposium on Information Theory*, eds. B. N. Petrov and F. Càski, Budapest: Akademiai Kiàdo, pp. 267-281.

Beal, D. J. (2007) "Information Criteria Methods in SAS for Multiple Linear Regression Models," in *SESUG 2007: The Proceedings of the SouthEast SAS Users Group*, Hilton Head, SC, USA, pp. SA05:1-10.

Burnham, K. P., and Anderson, D. R. (2004), "Multimodel Inference - Understanding AIC and BIC in Model Selection" *Sociological method & Research*, Vol. 33, No. 2, p. 261-304.

Imori, S., Katayama, S., and Wakaki H. (2014), "Screening and Selection Methods in High-Dimensional Linear Regression Model" *Preprint provided by autor at: http://www.me.titech.ac.jp/ miyalab/katayama/ paper.html*, Submitted to Electronic Journal of Statistics, p. 1-21.

Mallows, C. L. (1973), "Some Comments on C_p" *Technometrics*, Vol. 15, No. 4, p. 661-675.

Nishii, R. (1984), "Asymptotic properties of criteria for selection of variables in multiple regression" *The Annals of Statistics*, Vol. 12, No. 2, p. 758-765.

Sawa, T. (1978), "Information Criteria for Discriminating Among Alternative Regression Models" *Econometrica*, Vol. 46, No. 6, p. 1273-1291.

Schwarz, G. (1978), "Estimating the Dimension of a Model" *The Annals of Statistics*, Vol. 6, No. 2, p. 461-464.

Segiura, N. (1978), "Further Analysis of the Data by Akaike's Information Criterion and the Finite Corrections" *Communications in Statistics - Theory and Methods*, Vol. 7, No. 1, p. 13-26.

Wasserman, L., and Roeder, K. (2009), "High-Dimensional Variable Selestion" *The Annals of Statistics*, Vol. 37, No. 5A, p. 2178-2201.