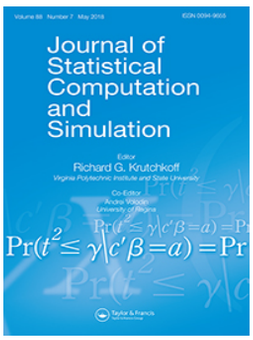

Journal of Statistical Computation and Simulation

ISSN: 0094-9655 (Print) 1563-5163 (Online) Journal homepage: https://www.tandfonline.com/loi/gscs20

A cluster tree based model selection approach for logistic regression classifier

Ozge Tanju & Zeynep Kalaylioglu

To cite this article: Ozge Tanju & Zeynep Kalaylioglu (2018) A cluster tree based model selection approach for logistic regression classifier, Journal of Statistical Computation and Simulation, 88:7, 1394-1414, DOI: 10.1080/00949655.2018.1437442

To link to this article: https://doi.org/10.1080/00949655.2018.1437442

Published online: 18 Feb 2018.

JOURNAL OF STATISTICAL COMPUTATION AND SIMULATION, 2018
VOL. 88, NO. 7, 1394–1414
https://doi.org/10.1080/00949655.2018.1437442

Taylor & Francis
Taylor & Francis Group

A cluster tree based model selection approach for logistic regression classifier

Ozge Tanju[a] and Zeynep Kalaylioglu[b]

[a]Department of Statistics, Ankara University, Ankara, Turkey; [b]Department of Statistics, Middle East Technical University, Ankara, Turkey

ABSTRACT

Model selection methods are important to identify the best approximating model. To identify the best meaningful model, purpose of the model should be clearly pre-stated. The focus of this paper is model selection when the modelling purpose is classification. We propose a new model selection approach designed for logistic regression model selection where main modelling purpose is classification. The method is based on the distance between the two clustering trees. We also question and evaluate the performances of conventional model selection methods based on information theory concepts in determining best logistic regression classifier. An extensive simulation study is used to assess the finite sample performances of the cluster tree based and the information theoretic model selection methods. Simulations are adjusted for whether the true model is in the candidate set or not. Results show that the new approach is highly promising. Finally, they are applied to a real data set to select a binary model as a means of classifying the subjects with respect to their risk of breast cancer.

ARTICLE HISTORY
Received 6 July 2017
Accepted 2 February 2018

KEYWORDS
Model selection; logistic regression; classification; clustering similarity measures

1. Introduction

'All models are wrong, but some are useful' [1]. This famous quote expresses that models are just approximations. Utility of a particular approximating model depends on the specific modelling purpose whether it is for variable selection, prediction, or classification. Therefore model selection methods should take the account of modelling purpose. However, this is not the case with current standard model selection criteria. In this article, we draw attention to this conflict and propose a new family of model selection criteria to be used particularly when modelling purpose is classification.

Current model selection stage is mainly based on (i) hypothesis testing, (ii) residual analysis, (iii) use of information theoretic model selection criteria. Numerous model selection methods based on i–iii are given in Rao and Wu [2]. Of the three categories, the standard ones are the information theoretic model selection criteria such as Akaike Information Criterion (AIC) [3], Bayesian Information Criterion (BIC) [4], Consistent Akaike Information Criterion (CAIC) [5], Information Complexity Criterion (ICOMP) [6] and

CONTACT Zeynep Kalaylioglu kzeynep@metu.edu.tr

Corrected Akaike Information Criterion (AIC_c) [7]. They are based on penalized likelihood functions and the objective function is $-2\log$ (likelihood). These are widely used over a set of nested, non-nested and overlapping linear, generalized linear, time series, nonlinear and mixed effects models. Performances of these criteria depend on many factors including sample size, modelling purpose and the types of the models in the candidate set. Overlapping and/or nonlinear models are perhaps the least addressed model types in the model selection literature. Recent model selection criteria for nonlinear models set include Kim and Cavanaugh [8], Zhang and Wu [9] and Claeskens et al. [10]. For overlapping models (linear or nonlinear), recent developments include Apparicio and Villanua [11] and Marcellino and Rossi [12]. Our focus is model selection in generalized linear models for categorical responses when the modelling purpose is classification. We propose a new family of model selection criteria that assess the models based on their classification performance and denote them by CC, short for cluster tree based criteria. Our aims are to (i) investigate the performances of the standard model selection criteria for different modelling purposes, (ii) compare the proposed purpose-specific criteria with standard criteria, which are *purpose-free*, when the modelling purpose is classification.

The remaining of the paper is organized as follows. Section 2 introduces the cluster tree based criteria developed for selecting the best logistic regression model where the modelling purpose is classification. Section 3 explains the approaches used to compare the new criteria with standard information based criteria. Section 4 is an extensive Monte Carlo simulation study that illustrates performances of cluster and information based model selection criteria. Section 5 presents a real data analysis. Finally, Section 6 compiles the main results.

2. Cluster tree based model selection

Logistic regression is widely used in practice as a classification tool, e.g. see [13–15] for applications in biomedical studies. It is a special case of generalized linear regression models, in which the logit function is used as the link function and

$$\text{logit}(P(Y_i = 1 \mid X_i)) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}, \quad i = 1, 2, \ldots, n \tag{1}$$

where

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right). \tag{2}$$

Classification of the subjects by their observed categorical responses and by their model based expectations can be thought as two different classifying (cluster) trees. Comparison of these cluster trees may give information about the plausibility of a particular categorical regression model when the intention is to use the model as a classifying tool. In particular, a strikingly high level of similarity between these two trees may be an indicator of a good fit. Most of the model diagnostics or goodness of fit testing strategies are based on residuals that measure the distance between observed and fitted values and small residuals are associated with a good fit. Our model selection criteria are based on the distance between the observed and predicted cluster trees. The distance describes the similarity/dissimilarity of the observed classification and the classification based on the predicted model. We adopt

Table 1. Number of pairs of responses classified in different or same cluster by observed and predicted trees.

		Predicted tree	
		Different cluster	Same cluster
Observed tree	Different cluster	A_{00}	A_{01}
	Same cluster	A_{10}	A_{11}

Jaccard (J) index [16] and Fowlkes and Mallows (FM) measure [17] to measure this distance. They measure the level of similarity between the two cluster trees. Our choice of distance measures is motivated by the fact that these two measures are more sensitive to dissimilarities between the clusterings [18]. Objective function in the proposed criteria is the distance between the observed classification and classification based on the predicted model. Objective functions in information theoretic model selection criteria are Kullback–Leibler distance. We think preceding objective function is more appropriate when the modelling purpose is classification. This idea applies to all categorical models such as probit and multinomial regressions which are also model based means of classification. Refer to [19–21] for examples of such models being used as classification tools.

Our approach is explained as follows. As seen in Table 1, let A_{00} be the number of pairs classified in different clusters by both observed and predicted trees, A_{01} be the number of pairs put in the different clusters by the observed tree but in the same cluster by the predicted tree, A_{10} be the number of pairs put in the same cluster by the observed tree but in different clusters by the predicted tree and finally A_{11} be the number of pairs classified in the same cluster by both observed and predicted trees.

J and FM are given below. Higher J and FM or equivalently lower (1-J) and (1-FM) are associated with stronger similarity between the two trees.

$$J = \frac{A_{11}}{A_{11} + A_{10} + A_{01}} \tag{3}$$

$$FM = \frac{A_{11}}{\sqrt{(A_{11} + A_{10})(A_{11} + A_{01})}}. \tag{4}$$

Lower {AIC, AIC$_c$, CAIC, BIC, ICOMP}, {1-J and 1-FM} lead to more plausible models. Below we develop a class of cluster tree based model selection criteria by penalizing 1-J and 1-FM.

2.1. Construction of cluster tree based model selection criteria

Our initial simulations showed that 1-J and 1-FM decrease with increasing number of covariates. That is, they have a tendency to select more complex models. In order to avoid favouring overfitting models, they should be penalized for the number of parameters. Penalizing 1-J and 1-FM (i.e. the distance between the predicted and observed clusters) is in spirit similar to penalizing the prediction loss function (i.e. total difference between the predictions and the observations) in Muller and Welsh [22] and Salibian-Barrera and Van Aelst [23].

Desired properties of a penalty are that (i) it should be an increasing function of the number of parameters and (ii) it should also increase by the sample size but with a slower rate. Figure 1 shows the values of the model selection criteria by d, the number of logistic regression coefficients. As seen in the figure, behaviours of AIC, AIC_c, CAIC, BIC and ICOMP are compatible with these properties. The slopes of CAIC and BIC are comparatively steeper implying better handling of overfitting problems. In other words, they choose the true model most of the times, which refers to their consistency based on Definition 3.1.

Let CC_J and CC_{FM} denote the CC based on J and FM respectively. We propose the following:

$$CC_J = (1 - J) + c_n \tag{5}$$

$$CC_{FM} = (1 - FM) + c_n \tag{6}$$

where c_n is a penalty term. Qian and Field [24] showed that a model selection criterion that consists of -2log likelihood and a penalty term is strongly consistent if the penalty term is an increasing function of the model dimension and has an order higher than $O(\log\log n)$. Based on these results, we propose the following two penalty terms, $c_{n1} = (p^u \text{logn})/100$ and $c_{n2} = (p^u \log \text{logn})/100$, where p is the number of regression coefficients (i.e. model dimension) and u is associated with the rate of decrease in 1-FM and 1-Jaccard (we set it at 1 based on a small simulation study). They are divided by 100 to preserve the original interpretation of J and FM measures as a measure between the two trees. Clustering based criteria with c_{n1} are named as CC_{FM1} and CC_{J1} whereas those with c_{n2} are named as CC_{FM2} and CC_{J2}. Figure 2 presents the behaviour of new penalized criteria as the number of parameters increase and are in line with the common criteria given in Figure 1.

3. Comparison of model selection criteria

In this section, we review the basic evaluation criteria to assess and compare the performances of model selection criteria. Consistency and efficiency are used to evaluate the

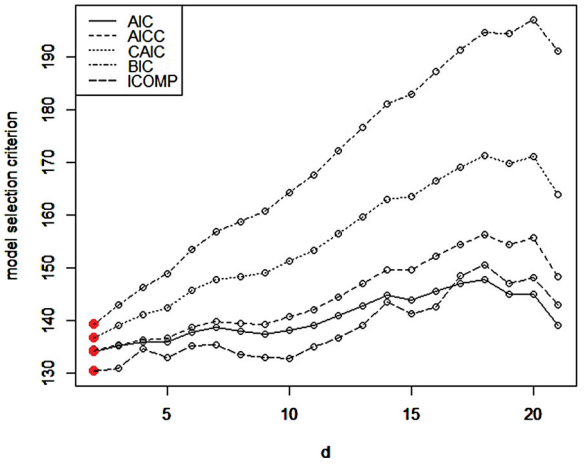

Figure 1. Common criteria versus number of parameters.

overall performance (when modelling purpose is unspecified). Classification performance is evaluated using true classification rate (TCR), sensitivity and specificity. In addition, we define consistency for nonlinear logistic models in Section 3.1.

3.1. Consistency

A model selection method is weakly consistent if the probability of the method selecting the true model in the candidate model set tends to 1 as $n \to \infty$. It is strongly consistent if the method selects the true model in the candidate model set with probability 1. It is very likely in real-life applications that candidate model set may not include the true model. Then the concept of consistency is defined in terms of the model selection method selecting the model in the candidate set that has the minimum Kullback–Leibler (KL) distance to the true model. Below lists the weak and strong consistency definitions used in the following section. First two definitions are found in the literature (e.g. [25]). Definition 3 is a new addition and extends the definition of consistency to nonlinear model selection.

Definition 3.1 (Strong Consistency): Let M_0 be the true model and $M_0 \in \mathcal{C}$, where $\mathcal{C}$ is the set of candidate models. A model selection criterion $R_n(.)$ is consistent if, for any $M_k \in \mathcal{C}$, $R_n(M_k) - R_n(M_0) \geq 0$ almost surely (a.s.) as $n \to \infty$.

Definition 3.2 (Weak Consistency): Let M_0 be the true model and $M_0 \notin \mathcal{C}$. Let $KL(M_1, M_2)$ be the Kullback–Liebler distance between any two models. Let $M_J \in \mathcal{C}$ such

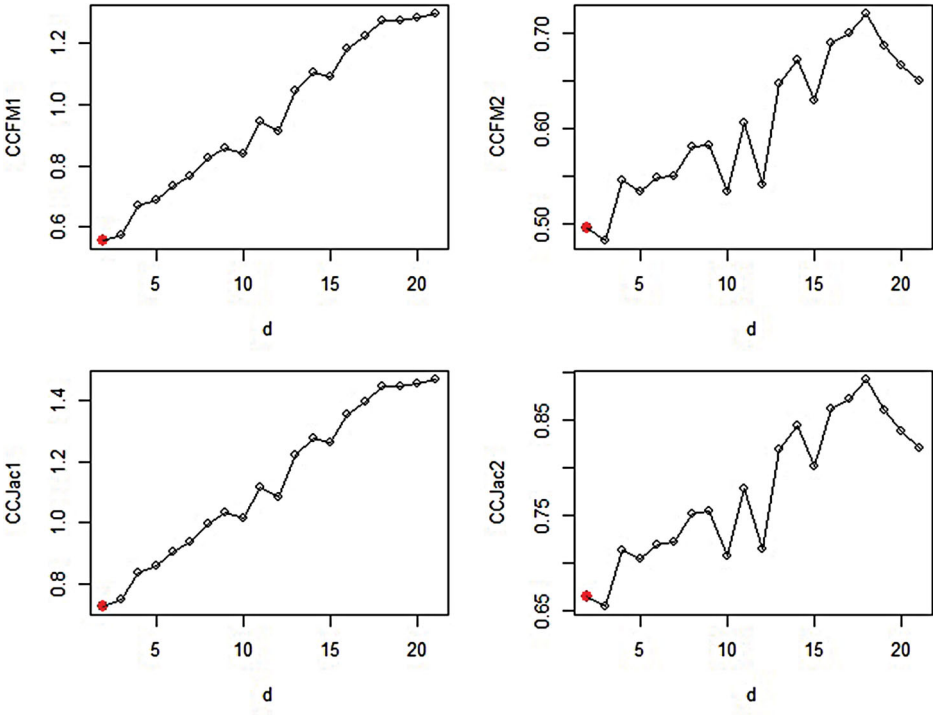

Figure 2. Cluster tree based criteria versus number of parameters.

that $\min_{M_k \in \mathcal{C}} KL(M_k, M_0) = KL(M_J, M_0)$. A model selection criterion $R_n(.)$ is weakly consistent, if the probability of $R_n(.)$ selecting M_J converges to 1 as $n \to \infty$.

AIC type of criteria are proven to be weakly consistent, whereas CAIC and BIC are strongly consistent [11,24,25].

Definition 3.3 (Strong Consistency): Let M_0 be the true nonlinear model with a complicated structure and $M_0 \notin \mathcal{C}$. Let $\mathcal{M}_J \subset \mathcal{C}$ be the set of polynomials well approximating M_0 such that $KL(M_j, M_0;\ j \in J) \leq c$, where c is a known constant and $\mathcal{M}_J$ is a subset of 'correct' models. A model selection criterion R_n is consistent if, for any $M_k \notin \mathcal{M}_J$, $R_n(M_k) - R_n(M_j) \geq 0$ almost surely (a.s.) $n \to \infty$.

A differentiable nonlinear function can always be well approximated by a polynomial of order p. Therefore, there is a true polynomial with an order p that is equivalent to the true nonlinear model with a complicated structure (M_0). $\mathcal{M}_J$ is the set of fitted polynomials that are best fitting among all the models in $\mathcal{C}$.

To the best of our knowledge, consistency of model selection methods in nonlinear logistic regression has not been addressed in the literature. Here we extend the consistency theorem of Qian and Field [24] for linear logistic regression to nonlinear logistic regression. We assume that simplest correct polynomial model is the model with minimum KL distance to the true model. Related axiom is given in Appendix C.

Conditions: Let $X = (X_1, \ldots, X_n)^{\mathrm{T}}$ be a single explanatory variable. Let $D = [1\ X\ X^2\ X^3 \cdots X^P]$ be the design matrix in a p-order polynomial logistic regression. Let $h(\eta) = \exp(\eta)/(1 + \exp(\eta))$.

(C.1) Columns of D are linearly independent.
(C.2) $E(DD^{\mathrm{T}})$ is positive definite.
(C.3) $E(\Pi_0(1 - \Pi_0)DD^{\mathrm{T}})$ and $E(\exp(-b\|D\|)\Pi_0(1 - \Pi_0)DD^{\mathrm{T}})$ are positive definite where $\Pi_0 = h(D^{\mathrm{T}}\beta_0)$ with β_0 being the true coefficients of the correct approximating polynomial with minimum order.
(C.4) $E(\|D\|^{2+\kappa}) < \infty$ for some $\kappa > 0$.
(C.5) $sup_k m_k < \infty$ where m_k is the number of parameters in the model.

Theorem: *Suppose conditions* (C.1)–(C.5) *hold. Then, if the order of the penalty term is greater than O(loglogn), then model selection criterion $R_n(.)$ is strongly consistent.*

Proof: Under conditions (C.1)–(C.5), following hold:

(C.1) $\lim_{n\to\infty} \lambda_k(I_n(\beta_0)) = \infty, k = 0, \ldots, p$. Also there exists some constant $d_0 > 0$ such that $0 < \lambda_p(I_n(\beta_0)) \leq d_0\lambda_1(I_n(\beta_0))$.
(C.2) $\delta_n(\mathrm{loglog}\lambda_p(I_n(\beta_0)))^{1/2} = o(1)$.
(C.3) $d_1 n \leq \lambda_p(I_n(\beta_0)) \leq d_2 n$ holds for some positive constants d_1 and d_2.
(C.4) $d_3 n \leq \lambda_p(X_n^t M_n X_n) \leq d_4 n$ for some positive constants d_3 and d_4.
(C.5) Let $b = \frac{1}{2}\min_{1\leq i\leq p_{\alpha_0}} |\beta_0(\alpha_0)|$ where α_0 is the correct model in $\mathcal{C}$ with the minimum dimension and $\beta_0(\alpha_0)_i$ is the ith component of $\beta_0(\alpha_0)$. Also let $Q_n = \mathrm{diag}(m_1\, \mathrm{e}^{-\|x_1\|} \times \pi_{01}(1 - \pi_{01}), \ldots, m_n\, \mathrm{e}^{-\|x_n\|} \times \pi_{0n}(1 - \pi_{0n})$ with $\pi_{0k}(k = 1, \ldots, n)$ being the true value of π_k. Then there exists a constant $d_5 > 0$ such that $\lambda_1(X_n^t M_n X_n) \leq d_5 n$.

Above, $\beta_0(M)$ are the true coefficients in the *true* polynomial that correspond to the terms X^k, $k = 1, \ldots, p$, in the fitted pth order polynomial $M \in \mathcal{C}$ and λ is the vector of eigenvalues of a $p \times p$ symmetric matrix. Then, $0 \leq \log L(\hat{\beta}(M) \mid Y, X) - \log L(\beta_0(m) \mid Y, X) = O(\text{loglogn})$ a.s. by Qian and Field [24]. Hence, $0 \leq R_n(\beta_0(M) - R_n(\hat{\beta}(M)) = m_M(\log L(\hat{\beta}(M) \mid Y, X) - \log L(\beta_0(M) \mid Y, X)) + (C(n, h(X, \beta_0)) - C(n, X^{\mathrm{T}}\hat{\beta})) = O(\text{loglogn}) + O(v_n)$, where $v_n > \text{loglog}n$, $h(X, \beta_0)$ is the true non-linear canonical predictor, $X^{\mathrm{T}}\hat{\beta}$ is fitted estimated canonical polynomial predictor and C(.,.) is a penalty function. ■

3.2. Efficiency

Efficiency is defined in terms of squared loss and given by $\mathcal{L} = \sum E((\hat{Y} - Y_{true})^2 \mid Y_{obs})$. A model selection criteria is said to be efficient if the probability of choosing the model with minimum loss tends to 1 as $n \to \infty$ [25]. The efficiency definition for logistic model selection criteria is given as follows.

Definition: Let $\mathcal{L}_{\min}$ be the minimum loss among the candidate models, and let $\mathcal{L}_0$ be the loss of a model chosen by a particular criterion. The model selection criterion $R_n(.)$ is efficient, if $\mathcal{L}_{\min}/\mathcal{L}_0$ converges to 1 in probability as $n \to \infty$. Based on this definition, Claeskens and Hjort [25] showed that AIC and AIC_c are efficient.

3.3. TCR, sensitivity, specificity

In various different disciplines, logistic regression is used as a classification tool [26–28]. Classification performance of a particular model is assessed by its TCR, sensitivity and specificity. Let n_{ij} $(i = 0, 1,\ j = 0, 1)$ be the number of subjects that are in cluster i based on their observed value and in j based on their predicted output. Then, true classification rate, sensitivity and specificity are

$$TCR = \frac{n_{00} + n_{11}}{n_{00} + n_{01} + n_{10} + n_{11}} \tag{7}$$

$$sensitivity = \frac{n_{11}}{n_{10} + n_{11}} \tag{8}$$

$$specificity = \frac{n_{00}}{n_{00} + n_{01}}. \tag{9}$$

4. Simulation studies

In this section, we present our extensive Monte Carlo simulation studies that investigated the performances of information and cluster tree based criteria in binary logistic regression selection under different scenarios adjusted for sample sizes and effect sizes. The performances are examined in two major settings depending on whether the true model is included in the candidate set (setting A) or not (setting B). For each, we tailored different realistic scenarios (cases). For setting A, we consider three cases that are common in practice. Case 1 is model selection over a candidate set of linear and quadratic models while

case 2 is over a set of main and interaction models. Finally, case 3 is the selection over a set of linear nested models. In setting B, we reconsider the linear nested model case.

Considered sample sizes are $n = 100$,500 or 1000 representing relatively small, moderate and rather large samples respectively, particularly in biological and economical studies. Binary response data Y_i are generated from $Be(p_i)$ where $p_i = P(Y_i = 1)$, for $i = 1, \ldots, n$, as seen below. Each experiment is repeated 1000 times. Performances are evaluated using consistency, efficiency, TCR, sensitivity and specificity. Consistency and TCR results are given here while the others are found in Appendix A.2. Frequency of a model selection criterion selecting the true model converging to 1 with increasing n in a simulation study indicates strong consistency. Consistency and efficiency results provide information about the overall performances of model selection criteria. TCR/sensitivity/specificity on the other hand provide information about their specific performance when the modelling purpose is classification in particular.

Setting A: Candidate Model Set Includes the True Model

Case 1: Linear and Quadratic Models

True model is $\text{logit}(p_i) = \beta_0 + \beta_1 x_i + \beta_2 x_i^2$, where $X_i \sim U(-3, 3)$. Three different true parameter sets are considered for $(\beta_0, \beta_1, \beta_2)$ implying models with different degrees of nonlinearity (see Figure A1 in Section A.1.1). These are (1.138, 1.256, 0.0038) for Model 1, (−2.742, 0.722, 0.391) for Model 2 and (−5.733, −0.275, 1.056) for Model 3. The aim of the current simulation study is twofold; comparison of the performances of the model selection criteria with respect to sample size and level of nonlinearity of the underlying true model. Candidate model set consists of the following models.

1. $\text{logit}(p(Y_i = 1 \mid x_i)) = \beta_0 + \beta_1 x_i + \beta_2 x_i^2$
2. $\text{logit}(p(Y_i = 1 \mid x_i)) = \beta_0 + \beta_1 x_i.$

Finite sample properties of the model selection criteria in terms of consistency and TCR are given in Tables 2 and 3 respectively. In Table 2, notice that if the true underlying model is strictly nonlinear (model 3), all model selection criteria converge to 1 fast. For model 2 where true nonlinearity is less severe, classification based criteria seem to have slower consistency. Overall, among the information based criteria, ICOMP seems to be the best one. AIC and AIC_c perform better than CAIC and BIC. AIC is consistent when the generating model is the extended model [11]. Table A1 indicates similar results in terms of efficiency. According to Table 3, cluster tree based criteria with penalty term c_{n2} seem more preferable

Table 2. Frequency of selecting the true model out of 1000 replicates.

	Model 1		Model 2		Model 3	
Criterion	$n = 100$	$n = 500$	$n = 100$	$n = 500$	$n = 100$	$n = 500$
AIC	161	192	897	1000	1000	1000
AIC_c	151	188	892	1000	1000	1000
CAIC	84	67	859	1000	1000	1000
BIC	42	18	790	1000	1000	1000
ICOMP	218	304	916	1000	1000	1000
CC_{FM1}	82	10	454	371	975	999
CC_{FM2}	186	155	583	684	989	1000
CC_{J1}	100	27	508	473	985	1000
CC_{J2}	198	178	599	732	995	1000

Table 3. Monte Carlo average of TCR.

	Model 1		Model 2		Model 3	
Criterion	$n = 100$	$n = 500$	$n = 100$	$n = 500$	$n = 100$	$n = 500$
AIC	0.717	0.734	0.763	0.771	0.829	0.825
AIC_c	0.716	0.734	0.763	0.771	0.829	0.825
CAIC	0.717	0.736	0.763	0.771	0.829	0.825
BIC	0.716	0.735	0.762	0.771	0.829	0.825
ICOMP	0.717	0.736	0.762	0.771	0.829	0.825
CC_{FM1}	0.716	0.735	0.753	0.740	0.831	0.819
CC_{FM2}	0.718	0.737	0.768	0.770	0.829	0.825
CC_{J1}	0.715	0.735	0.761	0.750	0.834	0.824
CC_{J2}	0.719	0.737	0.768	0.772	0.829	0.825
TCR of true model	0.716	0.734	0.763	0.771	0.829	0.825

in this scenario. c_{n1}, on the other hand, seems to overpenalize the model for large number of parameters (the quadratic models in this case). Overall, CC_{FM2} and CC_{J2} have slightly better classification properties than the standard criteria.

Case 2: Main and Interaction Models

True model is $\text{logit}(p_i) = \beta_0 + \beta_1 x_i + \beta_2 d_i + \beta_3 x d_i$, where $X_i \sim U(-3, 3)$ and $d_i \sim Be(0.5)$, $i = 1, \ldots, n$. Three different true parameter settings are considered for $(\beta_0, \beta_1, \beta_2, \beta_3)$ implying models with different levels of interaction (see Figure A2). These are $(-1.702, 0.135, 0.269, 0.0898)$ for Model 1, $(-1.702, 0.135, 0.693, 0.231)$ for Model 2 and $(-1.702, 0.135, 1.792, 0.597)$ for Model 3. The aim of the current simulation study is twofold; comparison of the performances of the model selection criteria with respect to sample size and true level of interaction. In Figure A2, line $d = 0$ represents the model without d. The further away the line from $d = 0$ line is, the more pronounced effect the interaction has. Three independent simulation experiments are carried out corresponding to the three different true models considered. In each experiment, candidate model set consists of the two models that are with and without interaction. Candidate model set consists of the following models.

1. $\text{logit}(p(Y_i = 1 \mid x_i)) = \beta_0 + \beta_1 x_i + \beta_2 d_i + \beta_3 x d_i$
2. $\text{logit}(p(Y_i = 1 \mid x_i)) = \beta_0 + \beta_1 x_i.$

Finite sample properties of the model selection criteria are given in Tables 4 and 5. According to Table 4, rates of consistency of model selection criteria depend on the true nature of the underlying model. For instance, if the data inherit a profound true interaction, convergence to 1 is much faster. Overall, ICOMP outperforms the rest. One striking result is that BIC is underperforming when used for interaction models. Performances of CC_{FM2} and CC_{J2} are notably better than AIC type criteria particularly for small samples such as $n = 100$. Results regarding efficiency are given in Table A4 and in the similar manner. According to Table 5, CC_{FM2} and CC_{J2} overall outperform the rest. Among the information based criteria, ICOMP outperforms the others (except under Model 3). Cluster tree based criteria have relatively lower sensitivity (except Model 3) and higher specificity (Tables A5 and A6 respectively).

Case 3: Linear Nested Models

Comparison of nested models is common in various many applications. Two regression models are nested if one can be transformed to the other by constraining some of the

Table 4. Frequency of selecting the true model out of 1000 replicates.

	Model 1		Model 2		Model 3	
Criterion	$n = 100$	$n = 500$	$n = 100$	$n = 500$	$n = 100$	$n = 500$
AIC	212	412	504	983	998	1000
AIC_c	185	403	471	983	996	1000
CAIC	73	128	262	877	979	1000
BIC	32	35	146	697	951	1000
ICOMP	375	644	690	996	999	1000
CC_{FM1}	314	164	414	297	783	599
CC_{FM2}	558	580	671	835	958	996
CC_{J1}	380	279	473	407	827	710
CC_{J2}	566	603	678	854	963	997

Table 5. Monte Carlo average of TCR.

	Model 1		Model 2		Model 3	
Criterion	$n = 100$	$n = 500$	$n = 100$	$n = 500$	$n = 100$	$n = 500$
AIC	0.556	0.556	0.586	0.619	0.680	0.676
AIC_c	0.554	0.555	0.584	0.619	0.680	0.676
CAIC	0.545	0.539	0.569	0.612	0.680	0.676
BIC	0.537	0.529	0.557	0.600	0.679	0.676
ICOMP	0.569	0.573	0.597	0.620	0.680	0.676
CC_{FM1}	0.583	0.553	0.591	0.584	0.664	0.637
CC_{FM2}	0.597	0.588	0.615	0.619	0.678	0.676
CC_{J1}	0.588	0.564	0.597	0.592	0.670	0.648
CC_{J2}	0.596	0.586	0.614	0.621	0.680	0.676
TCR of true model	0.593	0.584	0.611	0.620	0.680	0.676

regression coefficients to zero. When the candidate set consists of less and more saturated models compared to the true model, analyst is faced with potential overfitting and underfitting problems. Some of the existing criteria such as AIC and AIC_c tend to overfit due to moderate penalty terms. Here simulations were run for sample sizes of 500 and 1000 due to convergence problems encountered when $n = 100$. True data generating mechanism is $\text{logit}(P(Y_i = 1 \mid x_i)) = 2.5 + 0.5x_{i1} + 0.8x_{i2} + x_{i3} + 1.2x_{i4} - 4.33x_{i5}$, where x_{ij}s are from $U(0, 6)$. Regression coefficients are set so that cases and controls are distributed equally in the sample. Candidate model set includes the following models.

1. $\text{logit}(P(Y_i = 1 \mid x_i)) = \beta_0 + \beta_1 x_{i1}$
2. $\text{logit}(P(Y_i = 1 \mid x_i)) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}$
3. $\text{logit}(P(Y_i = 1 \mid x_i)) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}$
4. $\text{logit}(P(Y_i = 1 \mid x_i)) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4}$
5. $\text{logit}(P(Y_i = 1 \mid x_i)) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5}$
6. $\text{logit}(P(Y_i = 1 \mid x_i)) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \beta_6 x_{i6}$
7. $\text{logit}(P(Y_i = 1 \mid x_i)) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \beta_6 x_{i6} + \beta_7 x_{i7}$.

The results are given in Tables 6 and 7. According to Table 6, the performances of BIC and CAIC are better than the others. AIC and AIC_c perform rather poorly due to the moderate penalty terms. It is known that AIC tends to overfit [4,5]. ICOMP underperforms relative to CAIC and BIC. Cluster tree based criteria are performing satisfactorily particularly for

Table 6. Frequency of selecting the true model out of 1000 replicates.

Criterion	$n = 500$	$n = 1000$
AIC	771	760
AIC_c	780	763
CAIC	937	967
BIC	982	994
ICOMP	828	828
CC_{FM1}	1000	1000
CC_{FM2}	938	974
CC_{J1}	997	1000
CC_{J2}	833	912

Table 7. Monte Carlo average of TCR.

Criterion	$n = 500$	$n = 1000$
AIC	0.912	0.895
AIC_c	0.913	0.895
CAIC	0.913	0.893
BIC	0.914	0.894
ICOMP	0.913	0.895
CC_{FM1}	0.914	0.894
CC_{FM2}	0.916	0.894
CC_{J1}	0.914	0.894
CC_{J2}	0.916	0.895
TCR of true model	0.914	0.894

large sample. CC with c_{n1} outperform the others. They penalize the *unnecessary* parameters adequately. Finite sample consistency according to Definition 3.1 is observed for all the criteria except AIC and its versions. Also, all criteria seem to be efficient (see Table A7). According to Table 7, in general, cluster tree based criteria have higher TCR especially for moderate sample sizes such as $n = 500$. Monte Carlo average of TCR for each criterion are similar to the TCR of the true model. For the set of overfitted candidate models, cluster tree based criteria give higher TCR than the true model.

Setting B : Candidate Model Set Does not Include the True Model

Case 1: Linear Nested Models

We reconsider Case 3 in Setting A with the true data generating model given therein and the candidate model set that excludes Model 5. In this case, a model selection criterion is consistent if the probability of selecting the model with the smallest Kullback–Leibler distance converges to 1 as n goes to ∞. The results are given in Tables 8 and 9. First, BIC should be neglected in this case as it is based on the assumption that true model is in the candidate set [4]. According to Table 8, among the traditional model selection criteria, CAIC has the best performance. This is again an indication for handling the overfitting problem better than the others. CC_{FM1} and CC_{J1} perform better than all the information based criteria. They guard against overfitting unlike AIC and AIC_c.

Results in Table 9 are similar to those in Table 7, when the candidate set includes both overfitted and underfitted models and when it has only overfitted models. When only underfitted models exist in the candidate set, TCR decreases for each criterion. In that case, common criteria perform better than cluster based criteria.

Table 8. Frequency of selecting the model with minimum KL distance out of 1000 replicates.

Tool	$n = 500$	$n = 1000$
AIC	773	781
CAIC	897	907
AIC_c	782	782
BIC	939	937
ICOMP	813	816
CC_{FM1}	948	924
CC_{FM2}	884	925
CC_{J1}	946	945
CC_{J2}	795	862

Table 9. Monte Carlo average of TCR.

Criterion	$n = 500$	$n = 1000$
AIC	0.913	0.894
AIC_c	0.913	0.894
CAIC	0.913	0.892
BIC	0.913	0.892
ICOMP	0.913	0.894
CC_{FM1}	0.912	0.882
CC_{FM2}	0.914	0.891
CC_{J1}	0.913	0.890
CC_{J2}	0.916	0.892

5. Application

We applied the model selection methods considered herein on data set obtained in a research conducted in Ankara Oncology Research and Education Hospital. The hospital which is in the capital city Ankara was founded in 1956 by Turkish Cancer Research Organization and is the leading national cancer hospital that admits patients from across the country. The data set includes information on disease characteristics, risk factors and adjusting covariates of 249 women with breast cancer and 251 without. It was first analysed by Dogan et al. [29] to investigate the etiologic heterogeneity of breast cancer in Turkish population. Here our aim is to find the best classifying model which eventually can be used by health professionals to classify the Turkish women under risk as high versus low risk given their risk factors.

In what follows, the significant risk factors given in Dogan et al. [29] as well as the general adjusting factors for breast cancer are considered for the modelling. Namely, age (AGE: continuous), body mass index (BMI: continuous), menstrual regularity (MR: categorical, 1: regularity in pre-menopausal period, 2: irregularity in pre-menopausal period, 3: perimenopausal period, 4: post-menopausal period), menstruation age (MA: continuous), age at first birth (AFB: continuous), smoking habit (S: categorical, 0: nonsmoker, 1: smoker), hormone replacement therapy (HRT: categorical, 0: no HRT, 1: HRT with estrogen receptor, 2: HRT with progesterone receptor, 2: both), family history (FH: categorical, 0: no family history, 1: first-order relative, 2: second-order relative), mammography (M: categorical, 0: never had a mammography, 1: twice a year), cystite (CYS: categorical, 0: not have a cyst history, 1: have a cyst history) and biopsy status (BS: categorical, 0: not had a biopsy, 1: had a biopsy). Figure 3 gives the empirical scatter plot of proportion of cases in

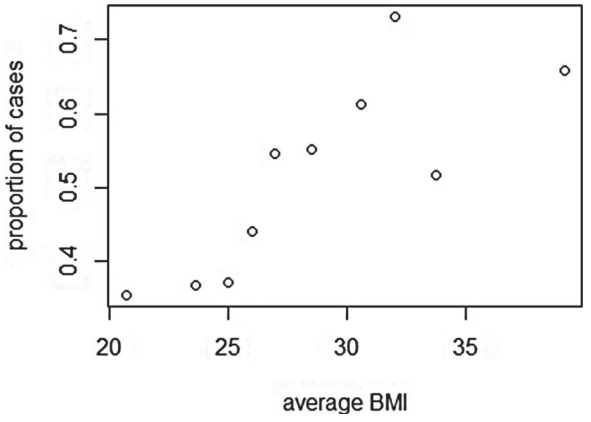

Figure 3. $P(Y = 1)$ versus average BMIs.

each BMI group versus averages of grouped BMIs and displays a nonlinear trend. Hence the suitable models of choices are:

$$\begin{aligned} M1 : \text{logit}(P(Y=1))_i = {} & \beta_0 + \beta_1 AGE_i + \beta_2 BMI_i + \beta_3 MR1_i + \beta_4 MR2_i + \beta_5 MR3_i \\ & + \beta_6 MA_i + \beta_7 AFB_i + \beta_8 S_i + \beta_9 HRT1_i + \beta_{10} HRT2_i \\ & + \beta_{11} HRT3_i + \beta_{12} FH1_i + \beta_{13} FH2_i + \beta_{14} M_i + \beta_{15} CYS_i \\ & + \beta_{16} BS_i \\ M2 : \text{logit}(P(Y=1))_i = {} & \beta_0 + \beta_1 AGE_i + \beta_2 BMI_i + \beta_3 BMI_i^2 + \beta_4 MR1_i + \beta_5 MR2_i \\ & + \beta_6 MR3_i + \beta_7 MA_i + \beta_8 AFB_i + \beta_9 S_i + \beta_{10} HRT1_i \\ & + \beta_{11} HRT2_i + \beta_{12} HRT3_i + \beta_{13} FH1_i + \beta_{14} FH2_i + \beta_{15} M_i \\ & + \beta_{16} CYS_i + \beta_{17} BS_i \end{aligned}$$

where M1 is a fully linear model where M2 is quadratic in BMI. Output for these candidate models are given in Tables A10 and A11. * signs for significant covariates at 0.05 significance level. Model selection criteria and p values of the Hosmer and Lemeshow test (for the global significance of the model) are given in Table 10. According to the p values, both models are significant where significance of M2 is stronger. In the table, selected models are marked in bold. Accordingly, information based criteria, CC_{FM2} and CC_{J2} select M2, whereas CC_{FM1} and CC_{J1} select M1. Recalling Table 2, ICOMP outperformed the other criteria in determining the need for a quadratic term whereas CC_{FM1} and CC_{J1} failed to determine nonlinearity in the model. Here all the criteria, except CC_{FM1} and CC_{J1}, are consistent with the choice of ICOMP. Only CC_{FM1} and CC_{J1} pick M1. This result is consistent with Table 2 as CC_{FM1} and CC_{J1} perform worse than the others. We also ran a simulation study to reveal the finite sample classification properties of the criteria for $n = 500$ (size of the observed dataset) when the candidate set of linear and quadratic logistic models excludes the underlying true nonlinear logistic model. Monte Carlo average of the TCR are equal for both models. Monte Carlo average of sensitivity rates of the models selected by information based criteria, CC_{FM2} and CC_{J2}, were slightly better than those of the rest. On

Table 10. Model selection for the breast cancer study.

Criterion	M1	M2
AIC	567.108	**561.596**
AIC_c	568.378	**563.018**
CAIC	602.933	**599.527**
BIC	638.757	**637.459**
ICOMP	638.757	**637.459**
TCR	0.706	0.706
Sensitivity	0.829	**0.928**
Specificity	**0.582**	0.486
CC_{FM1}	**1.460**	1.496
CC_{FM2}	0.715	**0.706**
CC_{J1}	**1.632**	1.669
CC_{J2}	0.886	**0.879**
p-value	0.970	0.191

the other hand, the models selected by CC_{FM1} and CC_{J1} have better Monte Carlo average of specificity rates than others.

6. Conclusion

Our main aim here was to draw attention to the importance of modelling purpose in model selection. We provided a new model selection approach for logistic regression models that is particularly useful when the modelling purpose is classification. We viewed the predicted and observed binary responses as two different cluster trees and employed clustering tree distances to assess the classification performances of logistic regression models. The special cases considered in the simulation study constitute the basis of modelling endeavours in practice. The simulations showed that overall performances of information based criteria depend on the penalty term. They also showed that when the modelling purpose lies in classification, cluster based criteria lead to best classifying model. Specific results are listed below. When the modelling purpose is unspecified (overall performance), results are as follows:

- When the candidate set includes rather parsimonious models (cases 1 and 2), ICOMP outperform all the others. Among cluster tree based criteria, CC_{FM2} and CC_{J2} are better than CC_{FM1} and CC_{J1}.
- When the candidate set includes both parsimonious and saturated models (case 3), among information based criteria CAIC and BIC perform better than AIC, AIC_c and ICOMP.
- When the candidate set includes both parsimonious and saturated models, overall, cluster tree based criteria perform better than information based criteria.

When the modelling purpose is classification in particular, results are as follows:

- CC_{FM2} and CC_{J2} perform at least as good as or better than standard information based criteria.

Also our application section leads to important results. Accordingly, CC with smaller penalty term is more useful for data analysis with many covariates under investigation.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

[1] Box GEP. Science and statistics. J Amer Statist Assoc. 1976;71(356):791–799.
[2] Rao CR, Wu Y. On model selection. In: Lahiri P, editor. Model selection. Beachwood (OH): Institute of Mathematical Statistics; 2001. p. 1–57.
[3] Akaike H. Information theory and an extension of the maximum likelihood principle. In Petrov BN, Csaki F, editors. Second international symposium on information theory. Budapest: Academiai Kiado; 1973. p. 267–281.
[4] Schwarz G. Estimating the dimension of a model. Ann Statist. 1978;6(2):461–464.
[5] Bozdogan H. Model selection and Akaike's information criterion (AIC): the general theory and its analytical extensions. Psychometrika. 1987;52(3):345–370.
[6] Bozdogan H. ICOMP: a new model selection criterion. In: Bock HH, editor. Classification and related methods of data analysis. Amsterdam: North-Holland; 1988. p. 599–608.
[7] Hurvich CM, Tsai CL. Regression and time series model selection in small samples. Biometrika. 1989;76(2):297–307.
[8] Kim HJ, Cavanaugh JE. Model selection criteria based on Kullback information measures for nonlinear regression. J Stat Plan Inference. 2005;134(2):332–349.
[9] Zhang T, Wu WB. Time-varying nonlinear regression models: nonparametric estimation and model selection. Ann Statist. 2015;43(2):741–768.
[10] Claeskens G, Croux C, Kerckhoven JV. Variable selection for logistic regression using a prediction-focused information criterion. Biometrics. 2006;62:972–979.
[11] Aparicio T, Villanua I. Some selection criteria for nested binary choice model: a comparative study. Comput Stat. 2007;22:635–660.
[12] Marcellino M, Rossi B. Model selection for nested and overlapping nonlinear, dynamic and possibly mis-specified models. Oxf Bull Econ Stat. 2008;70(1):867–893.
[13] Du X, Dua S, Acharya RU, et al. Classification of epilepsy using high-order spectra features and principle component analysis. J Med Syst. 2012;36:1731–1743.
[14] Lee JW, Lee JB, Park M, et al. An extensive comparison of recent classification tools applied to microarray data. Comput Stat Data Anal. 2005;48:869–885.
[15] Subasi A, Ercelebi E. Classification of EEG signals using neural network and logistic regression. Comput Methods Programs Biomed. 2005;78(2):87–99.
[16] Downton M, Brennan T. Comparing classifications: an evaluation of several coefficients of partition agreement. Paper presented at the meeting of the Classification Society; 1980; Boulder, CO.
[17] Fowlkes EB, Mallows CL. A method for comparing two hierarchical clusterings. J Am Stat Assoc. 1983;78(383):553–569.
[18] Milligan GW, Schilling DA. Asymptotic and finite sample characteristics of four external criterion measures. Multivar Behav Res. 1985;20:97–109.
[19] Krishnapuram B, Carin L, Figueiredo MAT, et al. Sparse multinomial logistic regression: fast algorithms and generalization bounds. IEEE Trans Pattern Anal Mach Intell. 2005;27(6):957–968.
[20] Ng AY, Jordan MI. On discriminative vs. generative classifiers: a comparison of logistic regression and naive Bayes. In: Dietterich TG, Becker S, Ghahramani Z, editors. NIPS; MA: MIT Press; 2000. p. 841–848.
[21] Balakrishnan S, Madigan D. Algorithms for sparse linear classifiers in the massive data setting. J Mach Learn Res. 2008;9:313–337.
[22] Muller S, Welsch AH. Outlier robust model selection in linear regression. J Am Stat Assoc. 2005;100:1297–1310.

[23] Salibian-Barrera M, Van Aelst S. Robust model selection using fast and robust bootstrap. Comput Stat Data Anal. 2008;52(12):5121–5135.
[24] Qian G, Field C. Law of iterated logarithm and consistent model selection criterion in logistic regression. Stat Probab Lett. 2002;56:101–112.
[25] Claeskens G, Hjort NL. Model selection and model averaging. Cambridge: Cambridge University Press; 2008.
[26] Zhu J, Hastie T. Classification of gene microarrays by penalized logistic regression. Biostatistics. 2004;5(3):427–443.
[27] Pernek M, Lackovic N, Matosevic D. Biology and natural enemies of spotted ash looper, *Abraxas pantaria* (Lepidoptera, Geometridae) in Krka National Park. Period Biol. 2013;115:371–377.
[28] Del Canto JG, Gonzalez IS. A resource-based analysis of the factors determining a firm's R & D activities. Res Policy. 1999;28(8):891–905.
[29] Dogan L, Kalaylioglu Z, Karaman N, et al. Relationships between epidemiological features and tumor characteristics of breast cancer. Asian Pac J Cancer Prev. 2011;12:3375–3380.
[30] Seghouane A-K, Amari S-I. The AIC criterion and symmetrizing the Kullback–Leibler divergence. IEEE Trans Neural Networks. 2007;18:97–106.
[31] Seghouane A. Asymptotic bootstrap corrections of AIC for linear regression models. Signal Processing. 2010;1:217–224.

Appendices

Appendix 1. Additional output for Section 4

A.1 Data generating models for Section 4

A.1.1 Models for Case 1

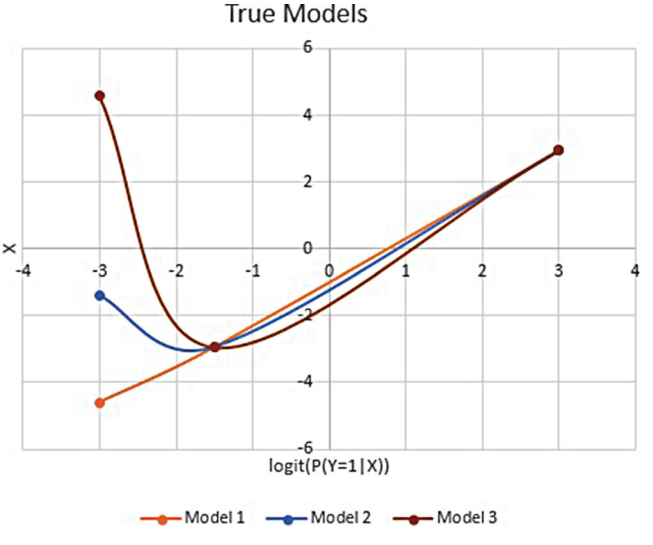

Figure A1. The levels of lack of linearity in the logit function.

A.1.2 Models for Case 2

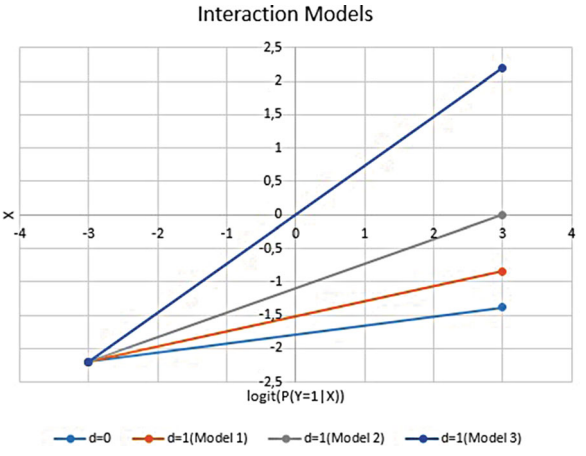

Figure A2. The levels of interaction in the logit function.

A.2 Simulation results

A.2.1 Results for Case 1

Table A1. Average observed efficiency rates.

	Model 1		Model 2		Model 3	
Tool	$n = 100$	$n = 500$	$n = 100$	$n = 500$	$n = 100$	$n = 500$
AIC	0.996	0.999	0.999	1	1	1
AIC_c	0.996	0.999	0.998	1	1	1
CAIC	0.994	0.999	0.998	1	1	1
BIC	0.993	0.998	0.995	1	1	1
ICOMP	0.996	0.999	0.999	1	1	1
CC_{FM1}	0.992	0.998	0.951	0.939	0.988	0.999
CC_{FM2}	0.993	0.999	0.964	0.969	0.995	1
CC_{J1}	0.993	0.998	0.957	0.949	0.993	1
CC_{J2}	0.994	0.999	0.965	0.974	0.998	1

Table A2. Monte Carlo average of sensitivity.

	Model 1		Model 2		Model 3	
Tool	$n = 100$	$n = 500$	$n = 100$	$n = 500$	$n = 100$	$n = 500$
Sensitivity of true model	0.676	0.703	0.582	0.586	0.738	0.734
AIC	0.685	0.703	0.585	0.586	0.738	0.734
AIC_c	0.684	0.703	0.586	0.586	0.738	0.734
CAIC	0.686	0.705	0.585	0.586	0.738	0.734
BIC	0.684	0.704	0.592	0.586	0.738	0.734
ICOMP	0.681	0.703	0.583	0.586	0.738	0.734
CC_{FM1}	0.685	0.704	0.624	0.653	0.738	0.728
CC_{FM2}	0.678	0.701	0.600	0.610	0.738	0.734
CC_{J1}	0.683	0.704	0.615	0.641	0.741	0.732
CC_{J2}	0.680	0.700	0.596	0.602	0.737	0.734

Table A3. Monte Carlo average of specificity.

	Model 1		Model 2		Model 3	
Tool	$n = 100$	$n = 500$	$n = 100$	$n = 500$	$n = 100$	$n = 500$
Specificity of true model	0.737	0.744	0.822	0.819	0.847	0.839
AIC	0.734	0.747	0.820	0.819	0.847	0.839
AIC_c	0.733	0.747	0.819	0.819	0.847	0.839
CAIC	0.733	0.749	0.819	0.819	0.847	0.839
BIC	0.732	0.748	0.815	0.819	0.847	0.839
ICOMP	0.736	0.748	0.820	0.819	0.847	0.839
CC_{FM1}	0.732	0.748	0.791	0.759	0.847	0.834
CC_{FM2}	0.738	0.752	0.819	0.811	0.848	0.839
CC_{J1}	0.731	0.748	0.805	0.775	0.850	0.839
CC_{FM2}	0.739	0.752	0.821	0.816	0.848	0.839

A.2.2 Results for Case 2

Table A4. Average observed efficiency rates.

	Model 1		Model 2		Model 3	
Tool	$n = 100$	$n = 500$	$n = 100$	$n = 500$	$n = 100$	$n = 500$
AIC	0.999	0.998	0.989	0.999	0.999	1
AIC_c	0.999	0.998	0.987	0.999	0.999	1
CAIC	0.999	0.994	0.975	0.998	0.999	1
BIC	0.998	0.992	0.965	0.994	0.996	1
ICOMP	0.999	0.999	0.995	0.999	0.999	1
CC_{FM1}	0.998	0.993	0.975	0.973	0.960	0.912
CC_{FM2}	0.999	0.997	0.988	0.995	0.994	0.999
CC_{J1}	0.998	0.994	0.979	0.978	0.969	0.937
CC_{J2}	0.999	0.997	0.988	0.995	0.995	0.999

Table A5. Monte Carlo average of sensitivity.

	Model 1		Model 2		Model 3	
Tool	$n = 100$	$n = 500$	$n = 100$	$n = 500$	$n = 100$	$n = 500$
Sensitivity of true model	0.420	0.416	0.441	0.421	0.568	0.558
AIC	0.471	0.453	0.467	0.421	0.569	0.558
AIC_c	0.472	0.454	0.468	0.421	0.571	0.558
CAIC	0.474	0.475	0.478	0.429	0.569	0.558
BIC	0.478	0.482	0.480	0.440	0.573	0.558
ICOMP	0.456	0.434	0.453	0.421	0.568	0.558
CC_{FM1}	0.429	0.451	0.438	0.448	0.549	0.529
CC_{FM2}	0.420	0.411	0.430	0.418	0.564	0.558
CC_{J1}	0.426	0.434	0.435	0.439	0.554	0.534
CC_{J2}	0.419	0.411	0.429	0.418	0.565	0.558

Table A6. Monte Carlo average of specificity.

	Model 1		Model 2		Model 3	
Tool	$n = 100$	$n = 500$	$n = 100$	$n = 500$	$n = 100$	$n = 500$
Specificity of true model	0.647	0.638	0.673	0.690	0.725	0.724
AIC	0.579	0.591	0.628	0.689	0.724	0.724
AIC_c	0.578	0.589	0.626	0.689	0.723	0.724
CAIC	0.566	0.589	0.604	0.676	0.722	0.724
BIC	0.556	0.548	0.588	0.658	0.719	0.724
ICOMP	0.602	0.618	0.651	0.689	0.724	0.724
CC_{FM1}	0.627	0.586	0.652	0.629	0.712	0.683
CC_{FM2}	0.648	0.640	0.682	0.690	0.724	0.724
CC_{J1}	0.634	0.607	0.660	0.644	0.718	0.697
CC_{FM2}	0.648	0.640	0.683	0.691	0.726	0.724

A.2.3 Results for Case 3

Table A7. Average observed efficiency rates.

Tool	$n = 500$	$n = 1000$
AIC	0.990	0.995
CAIC	0.984	0.991
AIC_c	0.990	0.995
BIC	0.982	0.991
ICOMP	0.988	0.994
CC_{FM1}	0.980	0.990
CC_{FM2}	0.981	0.991
CC_{J1}	0.980	0.990
CC_{J2}	0.983	0.991

Table A8. Monte Carlo average of sensitivity.

Tool	$n = 500$	$n = 1000$
Sensitivity of true model	0.913	0.893
AIC	0.913	0.895
AIC_c	0.914	0.895
CAIC	0.913	0.892
BIC	0.914	0.893
ICOMP	0.913	0.895
CC_{FM1}	0.913	0.893
CC_{FM2}	0.915	0.894
CC_{J1}	0.914	0.893
CC_{J2}	0.916	0.895

Table A9. Monte Carlo average of specificity.

Tool	$n = 500$	$n = 1000$
Specificity of true model	0.914	0.895
AIC	0.912	0.896
AIC_c	0.913	0.896
CAIC	0.913	0.894
BIC	0.914	0.895
ICOMP	0.913	0.896
CC_{FM1}	0.914	0.895
CC_{FM2}	0.916	0.895
CC_{J1}	0.914	0.895
CC_{J2}	0.917	0.895

Appendix 2. Inference based on candidate models in Section 5

Table A10. Candidate model 1.

Factor	OR	95% CI	*p* value
Age	1.027	(0.998,1.057)	0.068
BMI	1.030	(0.987,1.075)	0.180
Menstrual reg1	2.617	(1.253,5.465)	0.010*
Menstrual reg2	13.872	(6.014,34.767)	0.000*
Menstrual reg3	5.103	(2.707,9.846)	0.000*
Menstruation age	0.957	(0.822,1.112)	0.565
Age at the first birth	1.006	(0.978,1.034)	0.680
Smoking	0.789	(0.495,1.257)	0.317
HRT1	0.419	(0.156,1.078)	0.075
HRT2	0.493	(0.227,1.044)	0.069
HRT3	4.941	(1.562,18.009)	0.009*
Family history1	1.321	(0.755,2.321)	0.330
Family history2	2.095	(0.750,6.044)	0.162
Mammography	0.315	(0.190,0.515)	0.000*
Cyst	0.318	(0.147,0.643)	0.002*
Pathology	3.044	(1.285,7.470)	0.013*

Table A11. Candidate model 2.

Factor	OR	95% CI	*p* value
Age	1.020	(0.992,1.051)	0.171
BMI	1.637	(1.171,2.322)	0.004*
BMI^2	0.992	(0.987,0.998)	0.007*
Menstrual reg1	2.497	(1.186,5.254)	0.015*
Menstrual reg2	13.549	(5.847,34.060)	0.000*
Menstrual reg3	5.438	(2.862,10.592)	0.000*
Menstruation age	0.962	(0.825,1.121)	0.622
Age at the first birth	0.999	(0.971,1.029)	0.976
Smoking	0.783	(0.490,1.250)	0.304
HRT1	0.402	(0.149,1.043)	0.064
HRT2	0.466	(0.212,0.995)	0.052
HRT3	5.160	(1.598,19.127)	0.008*
Family history1	1.301	(0.739,2.300)	0.363
Family history2	2.357	(0.836,6.888)	0.109
Mammography	0.306	(0.184,0.502)	0.000*
Cyst	0.312	(0.145,0.636)	0.002*
Pathology	3.059	(1.283,7.559)	0.013*

Appendix 3. Axiom for 'Simplest correct polynomial has the smallest KL divergence from the true nonlinear model'

Let correct model is the true model with a complicated nonlinear structure. An equivalent model is infinite order polynomial model (from the Taylor series expansion of the true model) which we call full model.

Let set of fitted models is a subset of polynomials with different finite orders. Some of the models in this set are wrong models, some are correct.

Note that KL(f,g), i.e. KL distance between the true f (the pdf under the logistic regression with complicated non-linear predictor) and g (the pdf under the logistic regression with a predictor that is a polynomial of order k) is a function of k and it has a unique minimum as illustrated in Figures 2 and 3 of [30] and Figure 2 of [31]. Let $k_{\min}$ be the solution of $(\mathrm{d}/\mathrm{d}k)KL = 0$. Then, polynomials of order k where $k \geq k_{\min}$ are *correct* models and the polynomial of order $k = k_{\min}$ (which is the simplest polynomial) has the lowest KL distance.