CLUSTER BASED MODEL DIAGNOSTIC FOR LOGISTIC REGRESSION

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

ÖZGE TANJU

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
STATISTICS

JUNE 2016

Approval of the thesis:

CLUSTER BASED MODEL DIAGNOSTIC FOR LOGISTIC REGRESSION

submitted by **ÖZGE TANJU** in partial fulfillment of the requirements for the degree of **Master of Science in Statistics Department, Middle East Technical University** by,

Prof. Dr. Gülbin Dural Ünver _____
Dean, Graduate School of **Natural and Applied Sciences**

Prof. Dr. Ayşen Akkaya _____
Head of Department, **Statistics**

Assoc. Prof. Dr. Zeynep Kalaylıoğlu _____
Supervisor, **Statistics Department, METU**

Examining Committee Members:

Prof. Dr. Ayşen Akkaya
Statistics Department, METU _____

Assoc. Prof. Dr. Zeynep Kalaylıoğlu
Statistics Department, METU _____

Prof. Dr. Meriç Çolak
Health Care Management Department, Başkent University _____

Prof. Dr. Birdal Şenoğlu
Statistics Department, Ankara University _____

Assoc. Prof. Dr. Vilda Purutçuoğlu
Statistics Department, METU _____

Date: _____

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last Name: ÖZGE TANJU

Signature :

ABSTRACT

CLUSTER BASED MODEL DIAGNOSTIC FOR LOGISTIC REGRESSION

Tanju, Özge
M.S., Department of Statistics
Supervisor : Assoc. Prof. Dr. Zeynep Kalaylıoğlu

June 2016, 101 pages

Model selection methods are commonly used to identify the best approximation that explains the data. Existing methods are generally based on the information theory, such as Akaike Information Criterion (AIC), corrected Akaike Information Criterion (AICc), Consistent Akaike Information Criterion (CAIC), and Bayesian Information Criterion (BIC). These criteria do not depend on any modeling purposes. In this thesis, we propose a new method for logistic regression model selection where the modeling purpose is classification. This method is based on a measure of distance between two clusterings. There are many clustering similarity measures in the literature. Our model selection procedure is based on Jaccard index (Downton and Brennan, 1980) and Fowlkes-Mallows Index (Fowlkes and Mallows, 1983). The new model selection approach is compared against the currently used common methods in an extensive simulation study concerned with many different realistic scenarios. Scenarios are divided into two based on modeling purposes. Simulation scenarios are also grouped whether the true model is in the candidate models or not. We consider linear and nonlinear logistic models which are nested and non-nested, random-effects and fixed-effects models as true models. Simulation results show that the new method is highly promising. Apart from the new method, this thesis also provides an extensive comparison of the current methods based on information criteria. Finally, cluster based and information based criteria are applied to a real data set to select a binary model.

Keywords: Model Selection, Logistic Regression, Classification, Clustering Similarity Measures

ÖZ

LOJİSTİK REGRESYONDA KÜMEYE DAYALI MODEL SEÇİMİ

Tanju, Özge
Yüksek Lisans, İstatistik Bölümü
Tez Yöneticisi : Doç. Dr. Zeynep Kalaylıoğlu

Haziran 2016 , 101 sayfa

Model seçim yöntemleri veriyi açıklayan en iyi yaklaşık modeli belirlemek için yaygın olarak kullanılır. Mevcut model seçim metotları genellikle Akaike bilgi kriteri (AIC_c), tutarlı Akaike bilgi kriteri (CAIC), Bayesian bilgi kriteri (BIC) ve bilgi karmaşıklığı kriteri (ICOMP) gibi bilgi teorisi kullanan kriterlere dayalıdır. Bu kriterler herhangi bir modelleme amacına bağlı değildir. Bu tezde, lojistik regresyon için modelleme amacı sınıflandırma olan yeni bir model seçim yöntemi önerilmiştir. Bu yeni metot iki kümeleme arasındaki mesafenin ölçüsüne dayalıdır. Literatürde bir çok kümeleme benzerlik ölçüleri mevcuttur. Bizim model seçim prosedürümüz Jaccard ve Fowlkes-Mallows indekslerini baz almaktadır. Bu yeni model seçim yaklaşımı ile literatürde yaygın olarak kullanılan diğer metotlar bir çok farklı gerçek senaryo için geniş çaplı bir simülasyon çalışması ile karşılaştırılır. Senaryolar modelleme amaçlarına dayalı olarak ikiye ayrılır. Gerçek model olarak iç içe ve iç içe olmayan, rasgele etkili ve sabit etkili lineer ve lineer olmayan lojistik regresyon modelleri incelenmiştir. Simülasyon sonuçları yeni önerilen metodun benzer konuda gelecekte yapılacak çalışmalara temel oluşturacak nitelikte olduğunu göstermiştir. Bu tez çalışmasında yeni bir metot önermenin yanı sıra literatürde var olan bilgi temelli kriterlerin geniş çaplı bir karşılaştırılması da yapılmıştır. Tezin sonunda küme temelli ve bilgi temelli kriterler lojistik model seçimi için gerçek bir veri seti üzerinde uygulanmıştır.

Anahtar Kelimeler: Model Seçim Yöntemleri, Lojistic Regresyon, Sınıflandırma, Kümeleme Benzerlik Ölçümleri

To my family

ACKNOWLEDGMENTS

Firstly, I would like to express my sincere gratitude to my advisor Assoc. Prof. Dr. Zeynep Kalaylıoğlu for her precious guidance and endless patience during this thesis study. She was always helpful and very understanding in every stage of my study. With her immense knowledge, academic experience and elegant stance, she has always been a role model for me in academic career. It was a great fortune to be her student.

I would like to present my grateful thanks to my examining committee members, Prof. Dr. Ayşen Akkaya, Prof Dr. Meriç Çolak, Prof. Dr. Birdal Şenoğlu and Assoc. Prof. Dr. Vilda Purutçuoğlu for their valuable time to review my study.

I would also like to thank to Ezgi Ayyıldız, Tuğba Erdem, Duygu Varol and Çiğdem Güngör for their friendship and kind support. Also, I would like to thank to all the members of METU Statistics Deparment.

My special thanks go to Gürsu Gürer for his endless and loving support during this thesis. I also owe my thanks to Pınar Pekmez and Umur Berberoğlu for their warm frienship.

Finally, my deepest gratitude are for my lovely family, Süreyya Tanju, Ferudun Tanju and Bilge Tanju, for their endless support and sacrifices. They loved me unconditionally through every stage of my life. I would never achieve to be here without their encouragement.

TABLE OF CONTENTS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

LIST OF ABBREVIATIONS

AIC	Akaike Information Criterion
AIC_c	Corrected Akaike Information Criterion
BIC	Bayesian Information Criterion
BMI	Body Mass Index
CAIC	Consistent Akaike Information Criterion
CC	Cluster Based Criteria
ER	Estrogen Receptor
FIC	Focused Information Criterion
FM	Fowlkes Mallows
GLM	Generalized Linear Model
HRT	Hormone Replacement Theory
ICC	Intraclass Correlation Coefficient
ICOMP	Information Complexity
IFIM	Inverse Fisher Information Matrix
KL	Kullback Leibler
MLE	Maximum Likelihood Estimator
OR	Odds Ratio
PR	Progesterone Receptor
ROC	Receiver Operator Characteristic
TCR	True Classification Rate

CHAPTER 1

INTRODUCTION

"All models are wrong, but some are useful" (Box, 1976). This famous quote expresses that models are just approximations. Finding the *useful* model requires an adequate model selection process. It is needed for obtaining the best approximation using the data set. The importance of model selection is well understood by many researchers over the decades. Model selection is carried out for different purposes. These are for variable selection, for prediction, for classification.

For instance in a regression analysis, interest may lie in finding out the explanatory variables with non-zero regression coefficients (variable selection). In some regression, modeling purpose is to construct a prediction model to predict future responses from a given set of covariates (prediction). Modeling for variable selection and prediction can be referred as model fitting. In some areas, especially in archaeometry and medicine, interest lies in classifying the objects/subjects given a set of covariates (classification).

Current model selection procedures are based on i. hypothesis testing, ii. residual analysis, iii. use of information theoretic criteria. Numerous model selection methods based on i-iii are given in Rao and Wu (2001). However, of the three types of procedures, most widely used are the information theoretic model selection criteria (iii) and this is the focus of the thesis. These are based on penalized likelihood. The most commonly used ones are Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). There are many recent studies in literature proposing new criteria in order to overcome the problems related to these criteria. Aparicio and Villanua (2007) proposed a new model selection criterion denoted by C_2 for nested

binary models. It is based on the distance between observed variable and predicted probabilities. Cavanaugh (2004) proposed a new model selection criteria based on symmetric Kullback-Liebler (KL) divergence, whereas AIC type of criteria estimates the directed KL divergence. Claeskens et al. (2006) proposed a new model selection criterion based on focused information criterion (FIC), which is developed by Claeskens and Hjort (2003). They adjusted FIC based on prediction purposes for logistic models. Muller and Welsh (2010) proposed a new model selection methodology, namely model selection curves. All of these indicate that model selection is a complicated problem, it is still a hot topic in statistical research and it is the interest in many ongoing researches.

What we notice is that current model selection criteria do not take the account of modeling purpose. We think modeling purpose should be accounted for in the model selection process. This is implied also by C.R. Rao as a conclusion in his 2001 paper with Wu in that *"We wish to emphasize that the model we use to analyze a data set depends on the specific questions to be answered"*. We think each purpose (variable selection, prediction, classification) is related with a different question. There are only some studies in the literature that take the account of modeling purpose. They are basically the researches led by Celeux resulting publications of which are namely Biernacki et al. (2000) and Bouchard and Celeux (2006). Method in the first one is based on an integrated completed likelihood where the method in the later one relies on Bayesian paradigm and is called Bayesian Entropy Criterion.

In this thesis, our interest lies in model selection in binary regression where the modeling purpose is classification. Information theoretic criteria, such as AIC and BIC, are widely used for many different types of modeling such as linear regression, generalized linear regression (e.g. binary, Poisson), time series models, nonlinear models, and mixed effects models. We here focus on logistic regression models. Logistic regression is one of the classification methods in literature (Lee et. al, 2005). Therefore, we propose new model selection criteria for logistic regression in which the modeling purpose is classification. We will denote them by CC, short for cluster based criteria. They are based on clustering similarity measures existing in literature such as Fowlkes-Mallows (FM) and Jaccard measures. They define similarity between the two cluster trees. If these two trees are true and estimated trees out of a logistic re-

gression, their similarity can be an indicator of *usefulness* when the modeling purpose is classification.

In this thesis, we give information about the conventional model selection criteria in Chapter 2. We explain different logistic regression models and how their adequacy is checked. Chapter 2 also includes the theory behind the comparison of model selection criteria for different modeling purposes. We mainly focus on model fitting (i.e. modeling for variable selection and prediction) and classification purposes and evaluate those criteria accordingly. In Chapter 3, cluster based criteria are presented in detail. The need for a penalty term for those criteria are shown by small simulation studies. The results for Monte Carlo simulations are given in Chapter 4. We evaluated the performances of cluster based and information based criteria for model fitting and classification purposes. Chapter 5, on the other hand, present the outputs for a real data analysis. Finally, in Chapter 6, we sum up our studies and give remarkable points of this thesis.

CHAPTER 2

MODELS AND MODEL SELECTION

2.1 Models

Regression is a very common and useful tool in statistics to form the relation between variables. There is a vast amount of source on linear regression models. We herein only very briefly review the basics of it. The aim is to fit a model that explains the outcome in terms of related factors. Fitted model may be used to predict future values of the outcome variable given a set of effective factors. In general, a linear regression model fitted for n observations with k regressors can be written as the following.

$$y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik} + \epsilon_i, i = 1, 2, ..., n$$

where y_i indicates the outcome, namely the response variable, x_i 's are the factors, which are generally called as regressor variables, and β 's are regression coefficients. They are the parameters of the model to be estimated. They measure the linear effect of each factor on the response. Finally, ϵ_i represents the error term. This equation is in the form of a linear function. However, it is also statistical since it involves the random error term. The least squares estimation method is often used to estimate the parameters. This method is based on minimizing the sum of squares of errors. In order to apply this method and to make inference on the model, there are some assumptions:

1. Errors should follow a normal distribution with zero mean and a constant variance, σ^2.

2. Errors should be uncorrelated. i.e. repeated measurements of response should be independent of each other.

3. The linearity between response and regressors should be satisfied.

2.1.1 Generalized Linear Regression

Linear regression models are based on a very strict assumptions. It is not applicable for non-normal data such as count data, binary data, categorical data, and as such. For this problem, Nelder and Wedderburn (1972) developed a technique named as Generalized Linear Models (GLM).

Generalized linear models also relax the assumption of constant variance. A variance function is defined to understand the variance structure. It depends on both a dispersion parameter, ϕ and a function of mean of responses, $\vartheta(\mu)$. Variance function can be expressed by

$$Var(y_i) = \phi\vartheta(\mu)$$

Linearity in generalized linear models may change its usual meaning. Generalized linear models are conducted by using a link function g. The main purpose of using a link function is to change the range of responses into a proper range. The form of a generalized linear model is seen as

$$y_i = g^{-1}(\beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik}) + \epsilon_i, i = 1, 2, ..., n$$

Linear regression models are special cases of generalized linear models, where g is an identity link. As long as the data follows a distribution from exponential family of distributions, it is possible to use the method of generalized linear models. For example, when the data comes from Poisson distribution, the link function becomes the exponential function. For a multinomial data, inverse logit function is again used as the link function.

2.1.2 Logistic Regression

When the dependent variable in a regression model is binary, normality assumption obviously does not hold. The response follows a Bernoulli distribution with parameter p. p is namely the probability of observing the event of interest. Let Y be the response variable following the Bernoulli distribution, and it is either 0 and 1. 1 is for presence of the event of interest, and 0 is for its absence.

Logistic regression is a special case of generalized linear regression models, for which the logit function is used as the link function.

$$logit(P(Y_i = 1|X_i)) = \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik}, i = 1, 2, ..., n$$

Logit function is equivalent to taking natural logarithm of odds that Y=1. Odds that Y=1 means the proportion of P(Y=1|X) to P(Y=0|X). In other words, odds shows how it is likely to have the event of interest to occur for a given level of covariate.

$$log(\frac{P(Y_i = 1|X_i)}{P(Y_i = 0|X_i)}) = \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik}, i = 1, 2, ..., n$$

The above equation is also equivalent to

$$P(Y_i = 1|X_i) = \frac{e^{\beta_0+\beta_1 x_{i1}+...+\beta_k x_{ik}}}{1 + e^{\beta_0+\beta_1 x_{i1}+...+\beta_k x_{ik}}}, i = 1, 2, ..., n$$

For parameter estimation, the most common approach is method of Maximum Likelihood Estimation. Estimated regression coefficients are used in calculating estimated odds ratios ($\hat{OR}$) for two levels of the factor x,

$$\hat{OR} = \frac{\frac{P(Y=1|X=x_1)}{P(Y=0|X=x_1)}}{\frac{P(Y=1|X=x_2)}{P(Y=0|X=x_2)}} = \frac{e^{\beta_0+\beta_1 x_1}}{e^{\beta_0+\beta_1 x_2}} = e^{\beta_1(x_1-x_2)}$$

Odds ratio is the proportion of odds as seen in above equation. It shows how the odds of the event changes in relation to the change in factor x.

Logistic regression is widely used for biological data sets, since in those sets the main interest is generally having a disease or not. Moreover, logistic regression is also named as a classification tool (Lee et. al, 2005). By using a threshold, estimated

probabilities can be grouped into two. With a real data, one can create estimated classes for different characteristics of objects.

Diagnostic checks for logistic regression differ from those for linear regression. Pearson-chi square, and Hosmer-Lemeshow statistics are used to test the significance of the fitted model. In studies for which the main purpose is to group the objects, classification tables are used to calculate the true classification and misclassification rates. Again for the classification studies, ROC curves are used to measure the accuracy of logistic model fit. Sensitivity versus 1-specificity is plotted for all possible cut-off points for obtaining a ROC curve. Sensitivity is the true positive rate and 1-specificity is the false positive rate. The area under this curve measures the fitted model's ability to classify objects in a correct way (Hosmer and Lemeshow, 2000). Another technique is to choose a proper model out of a set of candidate models based on some criteria. This technique is easier to apply and can be used in all generalized linear models. Some commonly used and approved criteria will be explained in detail in the next section.

2.1.3 Random Effects Logistic Regression

Models discussed in previous sections are not applicable i. when the observations are correlated as in a longitudinal or panel study, ii. when there is a common group effect when there are homogeneous clusters in the data sets as in most survey data. A longitudinal study in which responses are repeatedly recorded for each subject in the study also results in a clustered data set, cluster being the longitudinal observations for the same subject. In such data sets, there are two types of effects in consideration: cluster specific effects (random effects) and population effects (fixed effects) (Fitzmaurice et al., 2004).

The model structure can be expressed in two ways. First is the random intercept model. This model includes a group-specific random effect on the response variable. This effect is totally random, and does not depend on any covariate. The form of a random intercept model is given by

$$y_i = b_{0i} + \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik} + \epsilon_i, i = 1, 2, ..., n$$

where b_{0i} represents the random intercept following a normal distribution with mean 0, and a constant unknown variance, σ_0^2. ϵ_i and b_{0i} are independent. For binary responses, logistic random intercept model takes the form of

$$logit(P(Y_i = 1|X_i)) = b_{0i} + \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik}, i = 1, 2, ..., n$$

In a longitudinal study, outcome variable may follow a subject-specific trend over the covariate. In this case, random effects model include both random intercept and random slope. The form of a random slope model is given by

$$y_i = b_{0i} + b_{1i} z_{1i} + \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik} + \epsilon_i, i = 1, 2, ..., n$$

where usually (b_{0i}, b_{1i}) has normal distribution with zero mean vector and a variance-covariance matrix Σ_b, and z_1 is the covariate associated with it which may be a subset of x's. Logistic random slope model can be written as

$$logit(P(Y_i = 1|X_i)) = b_{0i} + b_{1i} z_{1i} + \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik}, i = 1, 2, ..., n$$

Variance of these random terms is the variance between groups. For the random intercept model this variance is σ_0^2. Within group variance is expressed by the variance of the error in the model, which is σ^2. Summation of these two variance terms gives the total variance in the response, $\sigma_0^2 + \sigma^2$. The ratio of within group variance σ_0^2 to total variance in response, $\sigma_0^2 + \sigma^2$ is regarded as intraclass correlation coefficient (ICC). By this, it is understood that how much of the total variance comes from group effects. The formula for ICC is given by

$$ICC = \frac{\sigma_0^2}{\sigma_0^2 + \sigma^2}$$

For logistic models, ICC takes the form of

$$ICC = \frac{\sigma_0^2}{\sigma_0^2 + \pi^2/3}$$

2.2 Model Selection Methods

Fitting the best approximation out of a data set requires a model selection process. One needs to reach out a final model that best explains the data and has the least complexity. This final model is then used for inference. All other candidate models should be eliminated until finding the best one. Model selection criteria serves to selecting the best prediction model, association model and classification model. These criteria include i. information theoretic, ii. Bayesian, iii. information complexity based approaches. There are many methods for model selection. These criteria are mostly based on information theoretic approaches and loss functions. Information based model selection criteria relies on the concept of Kullback-Leibler. Before examining some widely used criteria, the concept of Kullback-Leibler information will be discussed.

KULLBACK-LEIBLER INFORMATION

Kullback and Leibler (1951) proposed a discrepancy measure which is directly related to Fisher's information matrix. It serves as a distance function between two statistical models. However, it should be noted that this is not a symmetrical distance, and triangle inequality does not hold. The distance from the first model to the second is not same as the distance from the second model to the first one. Kullback-Leibler (KL) distance is generally used for model selection purposes.

Let f and g present two probability functions standing for two statistical models. Consider f is fixed as the true model, and g shows the estimated one. Then KL distance from f to g is given by

$$I(f,g) = \int f(x) log(\frac{f(x)}{g(x|\theta)}) dx$$

where x is the data, and θ is the model parameters. KL distance is generally shown by I(f,g), since it is sort of a loss of information when approximating f.

The above equation can also be written as

$$I(f,g) = \int f(x)log(f(x))dx - \int f(x)log(g(x|\theta))dx$$

and it is equivalent to

$$I(f,g) = E_f[log(f(x))] - E_f[log(g(x|\theta))]$$

As it can be easily noticed the first part of the KL distance is not known, and can be considered as a constant. Only $E_f[log(g(x|\theta))]$ will change according to the estimated models. This quantity should be maximized so that $-E_f[log(g(x|\theta))]$ will be minimized in order to choose the best approximating model. Following this result, it can be concluded, model selection is based on the expected log-likelihood of the estimated model.

For binary models, the calculation for KL distance becomes the following.

$$I(f,g) = \sum plog(\frac{p}{\pi})$$

$$I(f,g) = \sum plog(p) - \sum plog(\pi)dx$$

$$I(f,g) = E_p[log(p)] - E_p[log(\pi)]$$

where p is the true probability function and π is the estimated probability function.

2.2.1 Akaike Information Criterion (AIC)

AIC is one of the most common model selection criteria. KL distance is not enough by itself as a model selection criteria, since it depends on an unknown truth. Akaike (1973) used likelihood theory to estimate the KL distance between the true and the candidate models. Akaike's purpose was to minimize this loss as well as to keep the model as simple as possible. It's pointed out that the bias in estimating a model is related to the number of parameters, k. Therefore, they subtracted the bias from the

estimated expected log-likelihood. The more parameters exist in a model, the more bias will occur. In order to eliminate this problem, AIC includes a penalty term depending on the number of estimated parameters. This penalty term is found while minimizing the information loss, namely KL distance.

$$AIC = -2log(L) + 2k$$

where L stands for the likelihood function, and k is the number of estimated parameters. The first term in AIC formula serves as a measure of goodness-of-fit, and 2k is the penalty term. For selecting a better model, one should choose the one with minimum AIC. By this way, the chosen model has the least information loss, so it can be seen as the best approximation.

2.2.2 Bayesian Information Criterion (BIC)

Bayesian Information Criterion (BIC), or namely Schwarz's Criterion is also a widely known model selection criterion (Schwarz, 1978). Its derivation is not based on KL distance, but it is based on selecting a mode which has the highest Bayesian posterior probability. BIC is given by

$$BIC = -2log(L) + klog(n)$$

The penalty term of BIC depends on the sample size, and it is much greater than the one in AIC. This property makes BIC preferable for overfitting problems. BIC is not based on the information theory, although its name indicates so. Smaller BIC indicates better model.

2.2.3 Consistent Akaike Information Criterion (CAIC)

Another extension of AIC is the Consistent Akaike Information Criterion (CAIC) (Bozdogan, 1987). Staying parallel with Akaike's understanding, Bozdogan extended AIC to make it asymptotically consistent, and more strictly penalized. The formula of this criterion is given by

$$CAIC = -2log(L) + k(log(n) + 1)$$

It is seen that in CAIC the penalty term is an increasing function of the sample size, where in AIC it is independent of the sample size.

2.2.4 Information Complexity Criterion (ICOMP)

Bozdogan (1988) suggested a new criterion for model selection. The idea he followed was the same as Akaike. He tried to minimize Kullback-Liebler distance. However, his penalty term was not directly based on the number of parameters, but it was based on the complexity of covariance matrix of the estimated model. Bozdogan (2000), then extended ICOMP, and it is given by

$$ICOMP = -2log(L) + 2C(F^{-1}(\hat{\theta}))$$

The first component of the above equation is the same as in AIC. Second part is the penalty term, where C stands for complexity. Inside the complexity function C, there exist the inverse Fisher information matrix (IFIM), F^{-1}, for the maximum likelihood estimator (MLE) of the model parameters $\hat{\theta}$. By theorem (Behboodian, 1964), variance of parameter estimates given in the IFIM increase as the number of parameters increase, and this is referred as variance inflation phenomenon. Hence, increasing number of parameters decreases the accuracy of a model fit. This means the penalty term is implicitly based on the number of estimated parameters.

2.2.5 Corrected Akaike Information Criterion (AICc)

For small sample sizes AIC is used with a correction term (Hurvish, Tsai, 1989). This corrected AIC is denoted by AIC_c, and it is given by

$$AIC_c = -2log(L) + 2k + \frac{2k(k+1)}{n-k-1}$$

As the sample size gets larger AIC_c gets closer to AIC. In the same way, smaller AIC_c values indicates better models for model selection process.

2.3 Comparison of Model Selection Criteria

Methods that are examined in this thesis are given in the previous sections. They are compared with each other and with the proposed criteria by simulation studies in Chapter 4. This comparison is based on purpose of modeling.

2.3.1 Comparison of Model Selection Criteria When Modeling Purpose is Model Fitting

When modeling purpose is model fitting, comparison of model selection criteria is conducted in terms of two essential terms: consistency and efficiency. Their definitions are provided in the following sections.

2.3.1.1 Consistency

A model selection method is weakly consistent if the probability of the method selecting the true model from the candidate model set tends to 1 as n goes to ∞. It is strong consistency if the method selects the true model from the candidate model set with probability 1.

Sometimes in real data applications, true model may not be included in the candidate model set. In this case, we assume that there is a model in the candidate model set that is closest in KL distance to the true model. Consistency is then related to the model selection method selecting this model.

Below listed the weak and strong consistency definitions for the cases focused in this thesis. First two definitions can be found in the related statistical literature as well. Definition 3 is a new addition by us for model selection where there is nonlinearity.

Definitions

Definition 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 $n \to \infty$.

In other words, probability of $R_n(M_0)$ being the minimum among all $R_n(M_k)$, where $M_k \in \mathcal{C}$ as n goes to ∞ is 1. That is $P(\forall \varepsilon > 0, \exists\, n_0$ such that for all $n \geq n_0$, for any $M_k \in \mathcal{C}$, $R_n(M_k) - R_n(M_0) \geq \varepsilon) = 1$. Strong consistency implies weak consistency.

Definition 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 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 $n \to \infty$.

Definition 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 $\mathcal{C}$.

Main Results

Given previous definitions, there are many studies evaluating conventional model selection criteria. AIC type of criteria are proven to be weakly consistent, whereas CAIC and BIC are strongly consistent. (Qian and Field, 2002; Claeskens and Hjort, 2008; Aparicio and Villanua, 2007).

To the best of our knowledge, consistency of model selection criteria in nonlinear logistic regression models has not been addressed in the literature. Here we extend the

consistency theorem of Qian and Field (2002) for linear logistic regression to non-linear logistic regression. In that paper, they established strong consistency of some important model selection criteria in logistic regression with linear predictor. Here we deal with logistic regression with nonlinear predictor. We assume that simplest correct polynomial model is the model with minimum KL distance to the true model. Related axiom is given in Appendix A.1.

The following conditions are needed:

Conditions: Let $X = (X_1, ..., X_n)^T$ be a single explanatory variable. Let D = [1 X X^2 X^3... 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) $E(DD^T)$ is positive definite.
(C.3) $E(\Pi_0(1-\Pi_0)DD^T)$ and $E(exp(-b||D||)\Pi_0(1-\Pi_0)DD^T)$ are positive definite where $\Pi_0 = h(D^T\beta_0)$ 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) $\lim_{n\to\infty} \lambda_k(I_n(\beta_0)) = \infty$, $k = 0, ..., p$. 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) $\delta_n(loglog\lambda_p(I_n(\beta_0)))^{1/2} = o(1)$.
(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 i^th component of $\beta_0(\alpha_0)$. Also let $Q_n = diag(m_1 e^{-||x_1||} \times \pi_{01}(1-\pi_{01}), ..., m_n e^{-||x_n||} \times \pi_{0n}(1-\pi_{0n})$ with $\pi_{0k}(k = 1, ..., n)$ 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 $X^k, k = 1, ...p$ in the fitted p^{th} order polynomial $M \in \mathcal{C}$ and λ presents the eigenvalues of a $p \times p$ symmetric matrix. Then, $0 \leq logL(\hat{\beta}(M)|Y, X) - logL(\beta_0(m)|Y, X) = O(loglogn)$ a.s. by Qian and Field (2002).

Hence, $0 \leq R_n(\beta_0(M) - R_n(\hat{\beta}(M)) = m_M(logL(\hat{\beta}(M)|Y, X) - logL(\beta_0(M)|Y, X)) + (C(n, h(X, \beta_0)) - C(n, X^T\hat{\beta})) = O(loglogn) + O(v_n)$ where $v_n > loglogn$, where n is the sample size, $h(X, \beta_0)$ is the true nonlinear canonical predictor, $X^T\ \hat{\beta}$ is fitted estimated canonical polynomial predictor, and C(.,.) is a penalty function.

2.3.1.2 Efficiency

Model selection criteria are also evaluated in terms of loss functions. We use the definition of the average squared distance between the observed values and predicted probabilities as the loss function for a logistic model. It is given by

$$\mathcal{L} = \sum((\hat{Y} - Y_{true})^2|Y_{obs})$$

A model selection criteria is efficient if the probability of choosing the model with minimum loss goes to 1 as n goes to ∞. The efficiency definition for logistic model selection criteria 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 chosen model by a criterion. The model selection criterion $R_n(.)$ is efficient, if $\frac{\mathcal{L}_{min}}{\mathcal{L}_0}$ converges to 1 in probability as n$\rightarrow \infty$.

Based on this definition Claeskens and Hjort (2008) showed that AIC and AIC_c are efficient.

2.3.2 Comparison of Model Selection Criteria When Modeling Purpose is Classification

Logistic regression is one of the classification tools. A group of subjects is divided into two groups based on their fitted probabilities by using a proper cut-off value. In

this section, we compare the classification based criteria and the conventional ones in terms of their accuracy in correctly classifying the subjects. To evaluate this, we use true classification rate, sensitivity, and specificity. Before explaining these measures, a general notation used for their definitions are given.

Table 2.1 illustrates the number of subjects put in two clusters, namely 0 and 1. Rows present the classification based on the observed values of 0 and 1 (true classification), whereas columns present the classification based on the predicted probabilities (model based classification).

Table 2.1: Number of subjects classified by two classifications

		Predicted	
		0	1
Observed	0	n_{00}	n_{01}
	1	n_{10}	n_{11}

n_{00} is the number of subjects classified as 0 by both classifications, n_{01} is the number of subjects classified as 0 based on observed values, as 1 based on predicted probabilities, n_{10} is the number of subjects classified as 1 based on observed values, as 0 based on predicted probabilities, n_{11} is the number of subjects classified as 1 by both classifications.

2.3.2.1 True Classification Rate (TCR)

Using the notation given in Table 2.1, true classification rate (TCR) is defined as

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

2.3.2.2 Sensitivity

Sensitivity is the proportion of true positives. Using the notation given in Table 2.1, the formula for sensitivity is given by

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

2.3.2.3 Specificity

Specificity is the proportion of true negatives. Using the notation given in Table 2.1, the formula for specificity is given by

$$specificity = \frac{n_{00}}{n_{00} + n_{01}}$$

CHAPTER 3

CLUSTERING BASED MODEL SELECTION

Penalized likelihood (information) based criteria do not take the account of modeling purpose. We believe that an effective model selection criteria should take the account of that. It is also stated in Baudry et. al (2015). Deriving model selection criteria with such a purpose would lead more adequate and parsimonious models.

In this chapter, our aim is to develop a model evaluating criterion that may be particularly useful when the modeling purpose is classification. Our approach is based on cluster tree similarity measures. Similarity of clusters may be used as an indicator of good fit for logistic regression for the cases in which the modeling purpose is classification. First we give a brief outline of cluster analysis and cluster similarity measures. Then we give our method that is based on similarity measures.

3.1 Cluster Analysis

In cluster analysis, one groups objects based on their similarities and dissimilarities in terms of some of their characteristics. Similarities are usually defined by Euclidean distance. These groups of objects form clusters which compose a cluster tree. Cluster analysis is a useful tool in many areas of research such as biology, psychology, insurance and earthquake studies. There are different types of clusterings. The most common ones are k-means clustering and hierarchical clustering. In k-means clustering, exactly k groups of objects are constructed. In hierarchical clustering, objects are belonged to several sub-clusters. In following sections, we examine some measures that measure similarities between clusterings. They are generally used for evaluating

hierarchical clusterings conducted with complete linkage algorithm.

3.2 Cluster Similarity Measures

Many indexes have been developed in order to measure the level of similarity between two cluster trees. Rand index, adjusted Rand measures, Fowlkes and Mallows measure and Jaccard index are the most common ones in the literature. They are based on the number of pairs exist in clusters. All these measures are based on the general notation explained below.

Table 3.1: Number of pairs classified by two clusterings

		Second clustering	
		Pairs in different clusters	Pairs in the same clusters
First clustering	Pairs in different cluster	A_{00}	A_{01}
	Pairs in the same clusters	A_{10}	A_{11}

where A_{00} is the number of pairs classified in different clusters by both partitions, A_{01} is the number of pairs put in the different clusters by the first clustering but in same cluster by the second clustering, A_{10} is the number of pairs put in the same cluster by the first clustering but in different clusters by the second clustering, and A_{11} is the number of pairs classified in the same cluster by both partitions. Related likelihood is given in Appendix A.2. For instance, the same five objects are put in two different clusterings as in the following trees (Fowlkes and Mallows, 1983).

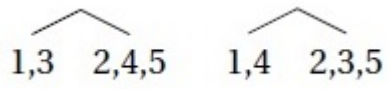

Figure 3.1: Two clustering trees

Trees in 3.1 are used to find A_{11}, A_{10}, A_{01} and A_{00}. Corresponding pairs are given in Table 3.2.

Table 3.2: Existing Pairs

Pairs in the same cluster according to the first clustering	(1,3)(3,1)(2,4)(4,2) (2,5)(5,2)(4,5)(5,4)
Pairs in different clusters according to the first clustering	(1,2)(2,1)(1,4)(4,1)(1,5)(5,1) (3,2)(2,3)(3,4)(4,3)(3,5)(5,3)
Pairs in the same cluster according to the second clustering	(1,4)(4,1)(2,3)(3,2) (2,5)(5,2)(3,5)(5,3)
Pairs in different clusters according to the second clustering	(1,2)(2,1)(1,3)(3,1)(1,5)(5,1) (4,2)(2,4)(4,3)(3,4)(4,5)(5,4)

Then Table 3.1 for this data set is given in Table 3.3.

Table 3.3: Number of pairs classified by two clusterings

		Second clustering	
		Pairs in different clusters	Pairs in the same clusters
First clustering	Pairs in different cluster	A_{00}=6	A_{01}=6
	Pairs in the same clusters	A_{10}=6	A_{11}=2

Rand index is derived with the purpose of evaluating the clustering methods (Rand, 1971). It is a proportion of positive and negative similarities in both clusterings to all cases. The other similarity measures are derived in order to handle some problems related to Rand index. Morey and Agresti (1984) showed that the Rand index fail to reveal the dissimilarities between the two partition. Even for randomly clustered objects, the Rand index tends to take high values. Their reasoning is that Rand counts the similarities even occurred by chance. Morey and Agresti adjusted the Rand index by subtracting a correction factor and eliminated the agreements occurred by chance. Later, it is proved that there has been some incorrect assumptions in calculating the correction factor (Hubert, Arabie, 1985). Hubert and Arabie (1985) also adjusted the Rand index. They suggest that the Rand index lacks a constant expected value. For example, a value of 0 is an indicator of perfect dissimilarity between two clusterings for this type of criteria. However, the Rand index never takes such a value. A perfect independence of two clusterings can only by demonstrated by a value of 0. Therefore, Hubert and Arabie use the assumption of randomness of two cluster trees, and take the generalized hypergeometric distribution as the null hypothesis. Their

adjusted Rand index is shown as the most desirable measure among the others (Steinley, 2004). However, the assumptions that they used seem unrealistic for many cases. Furthermore, from the formula it is assessed that this measure takes values between -1 and 1. Under their assumption of total independence of two clusterings, the index takes a constant value of 0. This should be an indicator for the perfect dissimilarity. It is also known that the higher values stand for higher similarity. Hence, negative values of HA seem uninterpretable. Jaccard and Fowlkes-Mallows measures are proved to be more sensitive to dissimilarities between clusterings. Therefore, we focus on Jaccard and Fowlkes-Mallows measures and explain them in more detail.

3.2.1 Jaccard

Downton and Brennan (1980) consider the number of pairs put in the same cluster by both of the partitions to measure similarity. They neglect the the number of pairs put in different clusters by both of the partitions. Unlike the Rand index, Jaccard does not produce too high values for misclassifications. The Jaccard index is given by

$$Jaccard = \frac{A_{11}}{A_{11} + A_{10} + A_{01}}$$

Jaccard takes values between 0 and 1, where 0 stands for perfect dissimilarity and 1 is a sign for perfect similarity between clusterings.According to the results of Monte Carlo simulation studies conducted by Milligan and Schilling (1985), Jaccard measure is sensitive to high level of dissimilarities between two clusterings. It also has a greater variability than other measures. This is a sign for its ability to notice even slight differences between clusterings.

3.2.2 Fowlkes-Mallows (FM)

Fowlkes and Mallows (1983) showed that Rand is highly dependent on the number of clusters and developed a new measure. Their simulations studies show that as the number of clusters increase, the Rand index takes values near 1 even for highly dissimilar clusterings. By using the general notation given above, FM is given by

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

FM changes between 0 and 1 in the same way as Jaccard. Milligan and Schilling (1985) showed that FM is very sensitive to severe misspecifications in clusterings.

3.3 Clustering Based Model Selection

Clustering a group of subjects according to their observed values of 0 and 1, and clustering the same group of subjects according to fitted probabilities can be thought as two different clusterings. Comparing these two clusterings may in some situations give information about the accuracy of a logistic regression model. In particular, a strikingly high level of similarity may be an indicator of good fit. Most of the model diagnostics (or goodness of fit testing ideas) are based on residuals, which measure the distance between observed and fitted value in a regression model. In a good fit, residuals are small. For any type of diagnostic method, the main interest is the distance between observed and predicted values. Clustering algorithms are also based on distance between objects. Comparing clustering based on observed values against that based on predicted values, is done by comparing the topographies of the two binary trees (clusterings) in terms of how similar/distant they are. Therefore, clustering similarity measures given in the previous section can be used as model selection criteria for logistic regression. Among the above measures, Jaccard and FM are chosen as the new model selection criteria due to their advantageous properties.

Jaccard and FM are used as model selection criteria in the same way with the existing criteria, namely AIC, AIC_c, CAIC, BIC, and ICOMP. Smaller values of $\{$AIC, AIC_c, CAIC, BIC, ICOMP$\}$ and $\{$1-FM, 1-Jaccard$\}$ indicate better fitted and better classifying models.

We noticed that, the new measures (1-FM) and (1-Jaccard) decrease as the number of parameters in the model increases, thus are refrained from selecting parsimonious models unlike desired, and thus they need to be penalized for the number of parameters. This is all illustrated using the following simulation experiment. In this simulation study, true model setting is given by

$$logit(P(Y_i = 1|X_i)) = 0.5 + 0.3x_i$$

where y is the binary response and x follows U(-3,3). A of 20 candidate models is constructed where the models include 1,2,3,...,20 covariates respectively. 20 different covariates from different distributions are added to the model.

As explained before, the smaller values of 1-FM and 1-Jaccard correspond to better models. In this simulation study, the true model involves two estimated parameters. All the other models in the candidate set have higher number of parameters. A reasonable model selection tool should choose the true model among those models including unnecessary factors. What we expect from a good criteria is to increase as the fitted model gets further from the true model.

Figure 3.2 and Figure 3.3 present the graphs of the value of the criteria versus the number of parameters (d). For this illustration the number of parameters in the true model is 2. Therefore, criteria should be lowest, when d is equal to 2. True model is marked by a red dot in the graphs. Both 1-FM and 1-Jaccard decrease, as more parameters are included in the model. In other words, they tend to select the model consisting redundant variables. Therefore, the proposed criteria should also have a penalty term for the number of parameters.

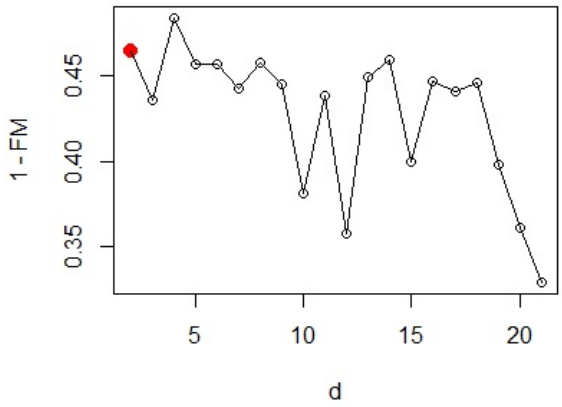

Figure 3.2: 1-FM vs. number of parameters

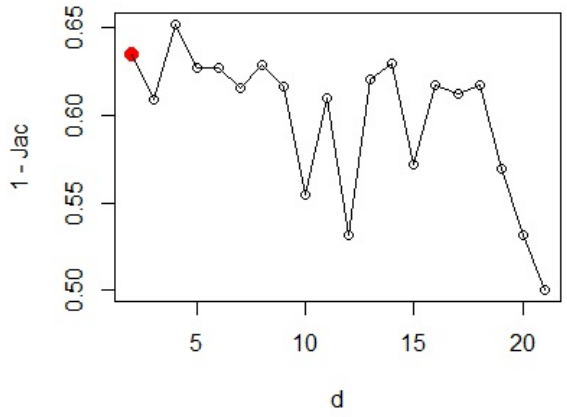

Figure 3.3: 1-Jac vs. number of parameters

The reason behind this situation can be associated with the residual idea. Let y_i^* present the predicted grouping values.

$$y_i^* = 1, if P(Y_i|X_i) > c$$
$$y_i^* = 0, if P(Y_i|X_i) < c$$

where c is the cut point of the clustering tree. Let r_i be the residuals from a logistic regression fit.

$$r_i = y_i - \hat{y}_i$$

If r_i is very close to 0, this means $\hat{y}_i$ is very close to either 1 or 0, since observed values can only take value 1 and 0. This also implies y_i^* should be either 1 or 0. Therefore, small residuals are indicators for a valid clustering. With this information, we should demonstrate that the more estimated parameters exist in a model, the smaller the residuals become. To illustrate, the following model setting is used.

$$P(Y_i = 1|x_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}$$

where y is the binary response variable and x follows U(0,6). Regression coefficients are chosen with the purpose of having an equal proportion of binary groups.

Candidate models are in the following in which all the covariates follow the same uniform distribution with the true model.

$$
\begin{aligned}
1.\quad P(Y_i = 1|x_i) &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} \\
2.\quad P(Y_i = 1|x_i) &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} \\
3.\quad P(Y_i = 1|x_i) &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} \\
4.\quad P(Y_i = 1|x_i) &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} \\
5.\quad P(Y_i = 1|x_i) &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \beta_6 x_{i6} \\
6.\quad P(Y_i = 1|x_i) &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \beta_6 x_{i6} + \beta_7 x_{i7}
\end{aligned}
$$

Data set is generated based on the true model, then the above 6 candidate models are fit once. Their residual sum of squares are given in Table 3.4.

As seen from Table 3.4, residuals get smaller for larger models, even if the true model is the smallest one. Since residuals are smallest for the largest model, 1-FM and 1-Jaccard will also tend to that model. This is actually the definition for overfitting problem of model selection criteria. The need for a penalty term is validated by this point of view.

Table 3.4: Residual sum of squares

Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
388.348	386.156	386.078	385.867	385.382	384.680

AIC, AIC_c, CAIC, BIC, and ICOMP are widely used for model selection. They are proven to be effective in such studies. Their behaviour in the same simulation scenario is examined in order to get an incentive to enhance 1-FM and 1-Jaccard. Figure 3.4 shows the relationship between each criterion and the number of parameters. Red points are again indicators for the true number of parameters. They all tend to rise as the number of parameters increases. In other words, they are able to pick the true model among the other models. The slopes of these plots are steeper for CAIC and BIC. They penalize the number of unnecessary parameters more than others. Their ability to handle overfitting problem is obvious in this study. In other words, they

choose the true model most of the times, which refers to their consistency based on *Definition 1*.

The behaviour of model selection criteria as the number of parameters increases has been investigated in the literature. Seghouane and Amari (2007) illustrated the relation between AIC type of criteria and the order of a polynomial model. They showed that those criteria tend to increase if the number of estimated parameters is more than the true number of parameters.

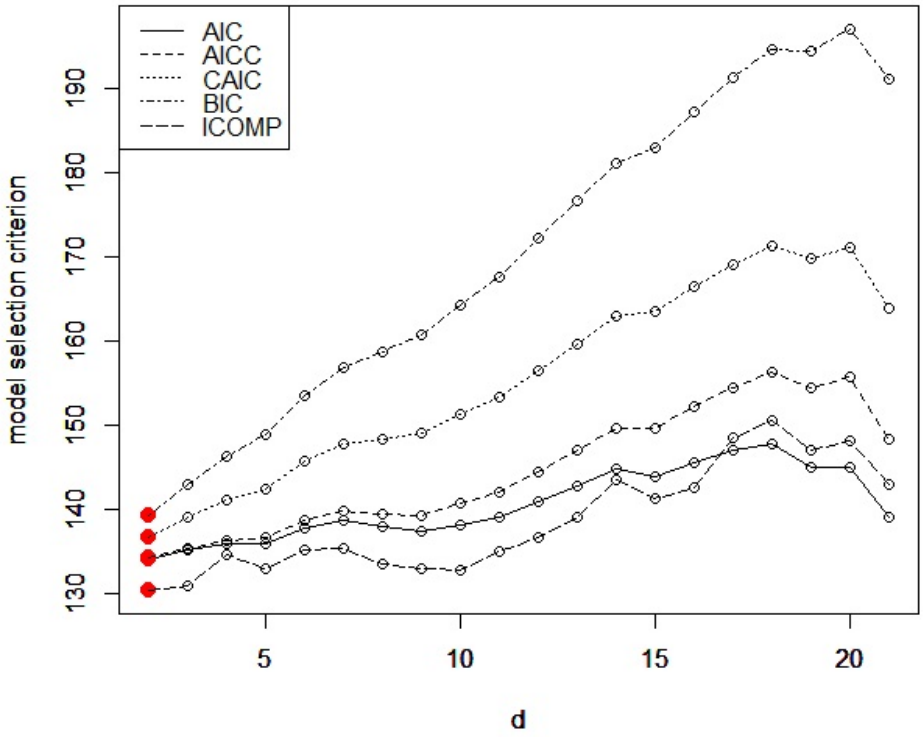

Figure 3.4: Common criteria vs. number of parameters

3.4 Penalty Term

As explained and illustrated in the previous section, we need to develop a penalty term for 1-FM and 1-Jaccard. Desired properties of a penalty are given as follows:

1. It should be an increasing function of the number of parameters.

2. It should also increase by the sample size, but with a smaller rate.

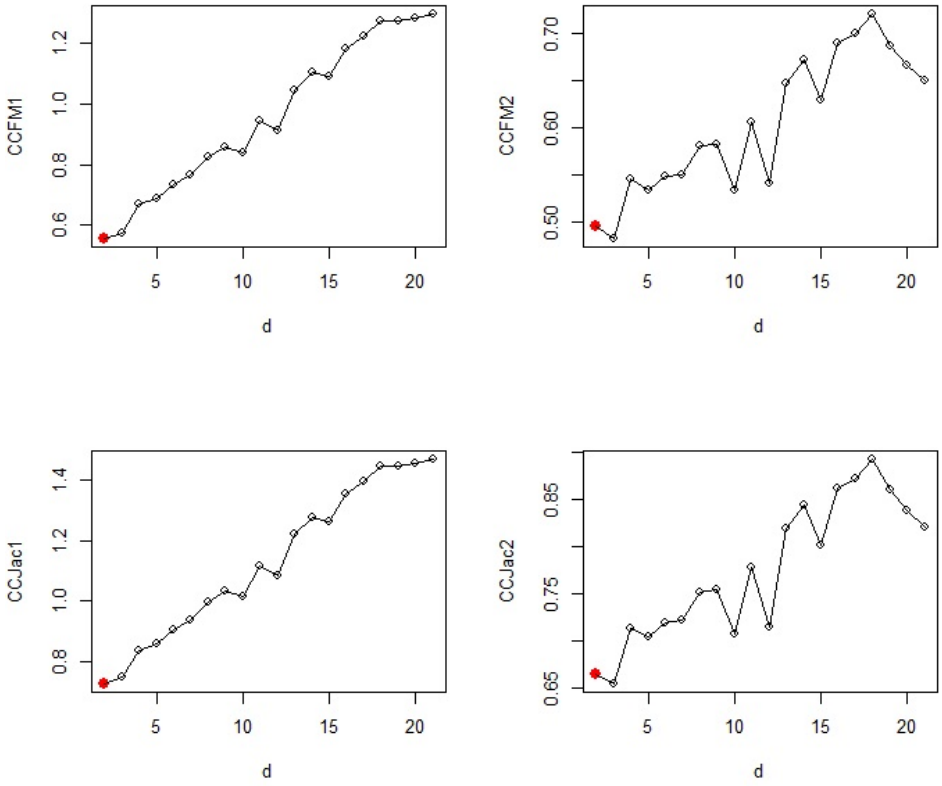

Figure 3.5: Cluster based criteria vs. number of parameters

Let the clustering based model selection criteria be presented by

$$CC_J = (1 - Jaccard) + c_n$$

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

where c_n is the penalty term. There are two penalty terms examined in this thesis. They are based on the existing penalty terms (logn and loglogn) for the common criteria and satisfy the desired properties. Our proposed penalty terms are

$$c_{n1} = (p^u logn)/100$$

$$c_{n2} = (p^u loglogn)/100$$

where p is the number of regression coefficients (i.e. model dimension) and u is the rate of decrease in 1-FM and 1-Jaccard and it is found as 1.

The incentive for these penalty terms is the Qian and Field (2002) result. They 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). Their result is based on law of iterated logarithm. Clustering based criteria with c_{n1} are named as CC_{FM1} and CC_{J1}, and clustering based criteria with c_{n2} are named as CC_{FM2} and CC_{J2}.

Figure 3.5 presents the behaviour of new penalized criteria as the number of parameters increase. These plots are similar to those for the common criteria given in Figure 3.4.

CHAPTER 4

SIMULATION STUDIES

In this chapter, different simulation scenarios are created in order to scrutinize the performances of information based and cluster based criteria. The two distinct families of criteria are compared on the basis of the modeling purpose. Simulation experiments are divided into two: i. if modeling purpose is model fitting (variable selection) (section 4.1), ii. if modeling purpose is classification (section 4.2). For each scenario, data sets including a binary response are generated based on a true model. Logistic regression models are fit on these data sets. For all simulations R programming language version 3.2.1 is used.

4.1 Modeling Purpose is Model Fitting

In this section, modeling purpose is model fitting. Therefore, consistency and efficiency definitions are used to evaluate model selection criteria.

4.1.1 True Model is in the Set of Candidate Models

Firstly, the situations in which the true model is in the set of candidate models are evaluated. In what follows, true models with different characteristics are setup. This enables investigation of performance of model selection criteria under various true settings. For each simulation study, data set is generated based on the true model. This data set is then used to fit different logistic regression models. In this section, true model is fitted as a candidate model along with the misspecified models. The

scenarios here are not applicable to real life cases, but they enable us to observe the exact performances of each criterion. Performance of the model selection criteria is based on strong consistency as in *Definition 1*. Frequency of selecting the true model out of 1000 trials are given in the subsequent tables. For efficiency the same definition given in Chapter 2 is used for all scenarios.

4.1.1.1 Detecting Missing Quadratic Terms

Quadratic models with different levels of lack of linearity are set using the following general form.

$$logit(p(Y_i = 1|x_i)) = \beta_0 + \beta_1 x_i + \beta_2 x_i^2$$

The scenarios used herein are the same scenarios used in Hosmer et al. (1997). In that paper, they investigated the power of various different goodness of fit tests used in logistic regression. There are 8 different true data generation processes under consideration. Vector of true coefficients, $(\beta_0, \beta_1, \beta_2)^T$, are chosen so that following conditional probabilities hold in the logistic model. Resulting response data generating logistic models are given next. These models are used to generate the binary response data. In all settings there is a single covariate and it is generated from U(-3,3). Models are also presented in Figure 4.1.

1. $P(Y = 1|x = -1.5) = 0.05, \quad P(Y = 1|x = 3) = 0.95, \quad P(Y = 1|x = -3) = 0.01$
2. $P(Y = 1|x = -1.5) = 0.05, \quad P(Y = 1|x = 3) = 0.95, \quad P(Y = 1|x = -3) = 0.05$
3. $P(Y = 1|x = -1.5) = 0.05, \quad P(Y = 1|x = 3) = 0.95, \quad P(Y = 1|x = -3) = 0.1$
4. $P(Y = 1|x = -1.5) = 0.05, \quad P(Y = 1|x = 3) = 0.95, \quad P(Y = 1|x = -3) = 0.2$
5. $P(Y = 1|x = -1.5) = 0.05, \quad P(Y = 1|x = 3) = 0.95, \quad P(Y = 1|x = -3) = 0.4$
6. $P(Y = 1|x = -1.5) = 0.05, \quad P(Y = 1|x = 3) = 0.95, \quad P(Y = 1|x = -3) = 0.7$
7. $P(Y = 1|x = -1.5) = 0.05, \quad P(Y = 1|x = 3) = 0.95, \quad P(Y = 1|x = -3) = 0.8$
8. $P(Y = 1|x = -1.5) = 0.05, \quad P(Y = 1|x = 3) = 0.95, \quad P(Y = 1|x = -3) = 0.99$

Under the above settings the correct models are as follows. Each of the eight models

corresponds to a different level of nonlinearity.

$$
\begin{aligned}
1.\quad & logit(p(Y_i = 1|x_i)) &=& -1.138 + 1.256x_i + 0.035x_i^2 \\
2.\quad & logit(p(Y_i = 1|x_i)) &=& -1.963 + 0.981x_i + 0.218x_i^2 \\
3.\quad & logit(p(Y_i = 1|x_i)) &=& -2.336 + 0.857x_i + 0.301x_i^2 \\
4.\quad & logit(p(Y_i = 1|x_i)) &=& -2.742 + 0.722x_i + 0.391x_i^2 \\
5.\quad & logit(p(Y_i = 1|x_i)) &=& -3.232 + 0.558x_i + 0.500x_i^2 \\
6.\quad & logit(p(Y_i = 1|x_i)) &=& -3.858 + 0.349x_i + 0.639x_i^2 \\
7.\quad & logit(p(Y_i = 1|x_i)) &=& -4.128 + 0.2596x_i + 0.699x_i^2 \\
8.\quad & logit(p(Y_i = 1|x_i)) &=& -5.733 - 0.275x_i + 1.056x_i^2
\end{aligned}
$$

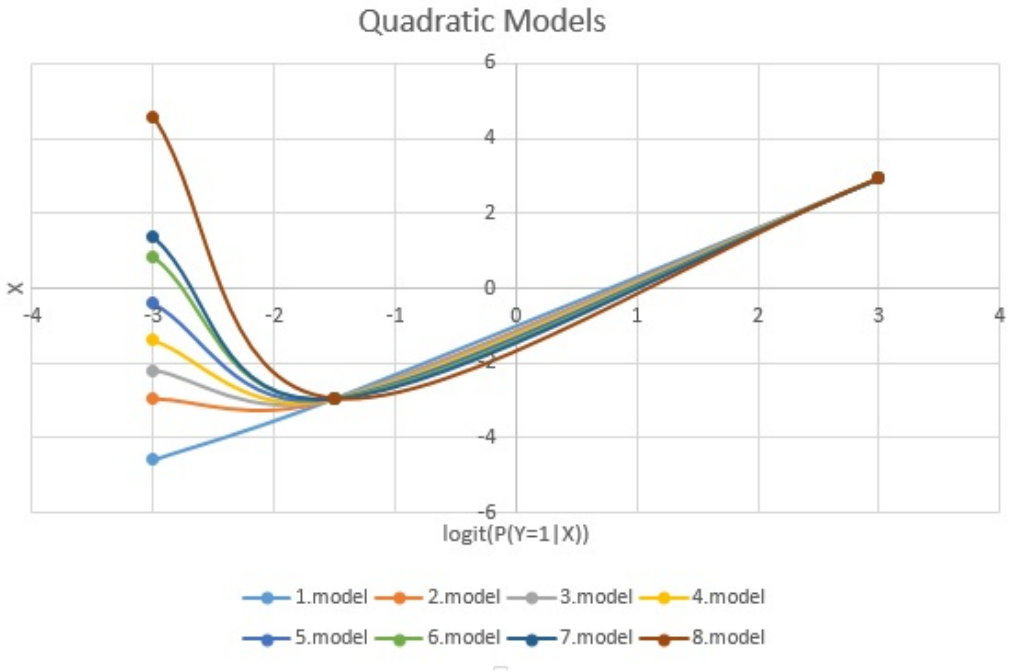

Figure 4.1: The levels of lack of linearity in the logit function

As seen in Figure 4.1, the level of nonlinearity increases from model 1 to model 8. For all of these eight settings, the candidate models are fitted in the following forms. The first one includes a quadratic term as in the generating model, and the second one misses the quadratic term.

$$
\begin{aligned}
1. \quad & logit(p(Y_i = 1|x_i)) &=& \beta_0 + \beta_1 x_i + \beta_2 x_i^2 \\
2. \quad & logit(p(Y_i = 1|x_i)) &=& \beta_0 + \beta_1 x_i
\end{aligned}
$$

Models are compared by using the criteria presented in Chapter 3.

Consistency

Strong consistency *Definition 1* from section 2 is used, since the true model is in the set of candidate models in these simulations. A model selection criterion is strongly consistent if, the selection of the true model happens almost surely, that is, its probability of being minimum for correctly fitted model is 1 for large n. Simulations for eight true models were run 1000 times each with sample sizes of 100 and 500. For each criteria, the frequency of choosing the correct model are shown in Table 4.1.

Table 4.1: Frequency of selecting the true model by each criterion out of 1000 replicates

Tool	Model 1		Model 2		Model 3		Model 4	
	n=100	n=500	n=100	n=500	n=100	n=500	n=100	n=500
AIC	161	192	500	971	729	999	897	1000
AIC_c	151	188	485	970	712	999	892	1000
CAIC	84	67	377	912	621	998	859	1000
BIC	42	18	284	824	525	992	790	1000
ICOMP	218	304	569	983	757	999	916	1000
CC_{FM1}	82	10	239	118	362	243	454	371
CC_{FM2}	186	155	385	435	511	564	583	684
CC_{J1}	100	27	293	195	421	320	508	473
CC_{J2}	198	178	398	470	522	612	599	732
Tool	Model 5		Model 6		Model 7		Model 8	
	n=100	n=500	n=100	n=500	n=100	n=500	n=100	n=500
AIC	984	1000	999	1000	999	1000	1000	1000
AIC_c	983	1000	999	1000	999	1000	1000	1000
CAIC	974	1000	999	1000	999	1000	1000	1000
BIC	962	1000	999	1000	999	1000	1000	1000
ICOMP	985	1000	999	1000	999	1000	1000	1000
CC_{FM1}	636	548	882	872	939	967	975	999
CC_{FM2}	772	827	944	977	969	999	989	1000
CC_{J1}	685	649	904	919	951	983	985	1000
CC_{J2}	784	856	950	984	971	1000	995	1000

In this scenario, true model is the largest model in the candidate set. Therefore, criteria with smaller penalty terms work better here. Cluster based criteria with penalty term c_{n2} are preferable. Cluster based criteria with penalty term c_{n1} penalize the true model too much for the number of estimated parameters. Moreover, its dependence on the sample size is with a higher factor. Therefore, it always performs poorer for larger sample sizes. CC_{FM2} and CC_{J2} are favoured for this case. They can be regarded as consistent in the context of *Definition 1*. However convergence seems to occur at a much slower rate compared to the conventional model selection criteria.

AIC is also known to be consistent when the generating model is the extended model (Aparicio, Villanua, 2007). That's what we observe in this situation. AIC and AIC_c perform better than CAIC and BIC. Among the information based criteria ICOMP seems to be the best based on Table 4.1.

If the true model is very mildly nonlinear, the common criteria cannot detect the missing nonlinearity for sample size of 100. For model 1 , AIC, AIC_c, CAIC, and BIC are not able to pick the true model. For such cases, ICOMP, CC_{FM2}, and CC_{J2} have better performances, since they show a higher proportion of selecting the true model. If the underlying true model is more nonlinear (through a more stressed quadratic coefficient), the common tools are successful as expected. However, for relatively small samples as in n=100, for Model 2, their performance is still questionable (AIC=500, AIC_c=485, CAIC=377, BIC=284, ICOMP=569). However, they are able to pick the true model when sample size is increased to 500 (AIC=971, AIC_c=970, CAIC=912, BIC=824, ICOMP=983). The number of selecting the true model for the new methods are less than those.

Efficiency

In terms of efficiency definition given in Chapter 2, average observed efficiency rates are calculated, and are given in Table 4.2. All criteria seem to perform well in terms of efficiency, but there are some remarkable points. Among the common criteria, AIC, AIC_c, and ICOMP perform better than CAIC and BIC. Efficiency of AIC type of criteria is also mentioned in the literature (Claeskens and Hjort, 2007). Their performance gets better from Model 1 to Model 8, and also as the sample size increases. For Model 8, all common criteria managed to choose the model with minimum loss

for all trials.

Table 4.2: Average observed efficiency rates

Tool	Model 1		Model 2		Model 3		Model 4	
	n=100	n=500	n=100	n=500	n=100	n=500	n=100	n=500
AIC	0.996	0.999	0.996	1	0.997	1	0.999	1
AIC_c	0.996	0.999	0.996	1	0.997	1	0.998	1
CAIC	0.994	0.999	0.993	0.999	0.994	1	0.998	1
BIC	0.993	0.998	0.989	0.999	0.991	1	0.995	1
ICOMP	0.996	0.999	0.996	1	0.998	1	0.999	1
CC_{FM1}	0.992	0.998	0.982	0.982	0.971	0.965	0.951	0.939
CC_{FM2}	0.993	0.999	0.986	0.989	0.979	0.980	0.964	0.969
CC_{J1}	0.993	0.998	0.984	0.983	0.974	0.969	0.957	0.949
CC_{J2}	0.994	0.999	0.986	0.990	0.979	0.982	0.965	0.974
Tool	Model 5		Model 6		Model 7		Model 8	
	n=100	n=500	n=100	n=500	n=100	n=500	n=100	n=500
AIC	1	1	1	1	1	1	1	1
AIC_c	1	1	1	1	1	1	1	1
CAIC	0.999	1	1	1	1	1	1	1
BIC	0.999	1	1	1	1	1	1	1
ICOMP	1	1	1	1	1	1	1	1
CC_{FM1}	0.936	0.912	0.966	0.956	0.980	0.987	0.988	0.999
CC_{FM2}	0.962	0.967	0.984	0.993	0.990	0.9996	0.995	**1**
CC_{J1}	0.946	0.932	0.972	0.973	0.984	0.993	0.993	1
CC_{J2}	0.964	0.972	0.986	0.995	0.990	1	0.998	1

When we examine the results for the cluster based criteria in terms of efficiency, it is observed that they are not as successful as the information based criteria. When the true model is close to linear, their efficiency is more or less same with the common criteria. For a slight misspecification, the losses of two candidate models are probably very close to each other. The high efficiency rates are a result of this. This performance falls upto Model 6, and start to rise again after Model 6. When the misspecification is severe, they are able to pick the model minimum loss. It is also observed that cluster based criteria with c_{n2} are more effective for this scenario due to the same reasoning in consistency study.

4.1.1.2 Detecting Missing Interaction Terms

Models involving different level of interaction are constructed in the following form.

$$logit(p(Y_i = 1|x_i)) = \beta_0 + \beta_1 x_i + \beta_2 d_i + \beta_3 x d_i$$

The settings are again based on Hosmer et al. (1997). They also examined the power of goodness of fit tests for detecting a missing interaction. True coefficients $(\beta_0, \beta_1, \beta_2, \beta_3)^T$ are chosen based on the following settings. The generating models are also given next. These models are used to generate the data set that are used to fit logistic regression models. X follows U(-3, 3), whereas the dichotomous variable, d, follows Ber(0.5). The generating models are presented in 4.2.

1. $P(Y = 1|x = -3, d = 0) = 0.1$, $P(Y = 1|x = -3, d = 1) = 0.1$,
   $P(Y = 1|x = 3, d = 0) = 0.2$, $P(Y = 1|x = 3, d = 1) = 0.3$
2. $P(Y = 1|x = -3, d = 0) = 0.1$, $P(Y = 1|x = -3, d = 1) = 0.1$,
   $P(Y = 1|x = 3, d = 0) = 0.2$, $P(Y = 1|x = 3, d = 1) = 0.5$
3. $P(Y = 1|x = -3, d = 0) = 0.1$, $P(Y = 1|x = -3, d = 1) = 0.1$,
   $P(Y = 1|x = 3, d = 0) = 0.2$, $P(Y = 1|x = 3, d = 1) = 0.7$
4. $P(Y = 1|x = -3, d = 0) = 0.1$, $P(Y = 1|x = -3, d = 1) = 0.1$,
   $P(Y = 1|x = 3, d = 0) = 0.2$, $P(Y = 1|x = 3, d = 1) = 0.9$

Under the above settings the correct models are given as follows. Each of these models involves a different level of interaction.

1. $logit(p(Y_i = 1|x_i)) = -1.792 + 0.135x_i + 0.269d_i + 0.0898xd_i$
2. $logit(p(Y_i = 1|x_i)) = -1.792 + 0.135x_i + 0.693d_i + 0.231xd_i$
3. $logit(p(Y_i = 1|x_i)) = -1.792 + 0.135x_i + 1.117d_i + 0.372xd_i$
4. $logit(p(Y_i = 1|x_i)) = -1.792 + 0.135x_i + 1.792d_i + 0.597xd_i$

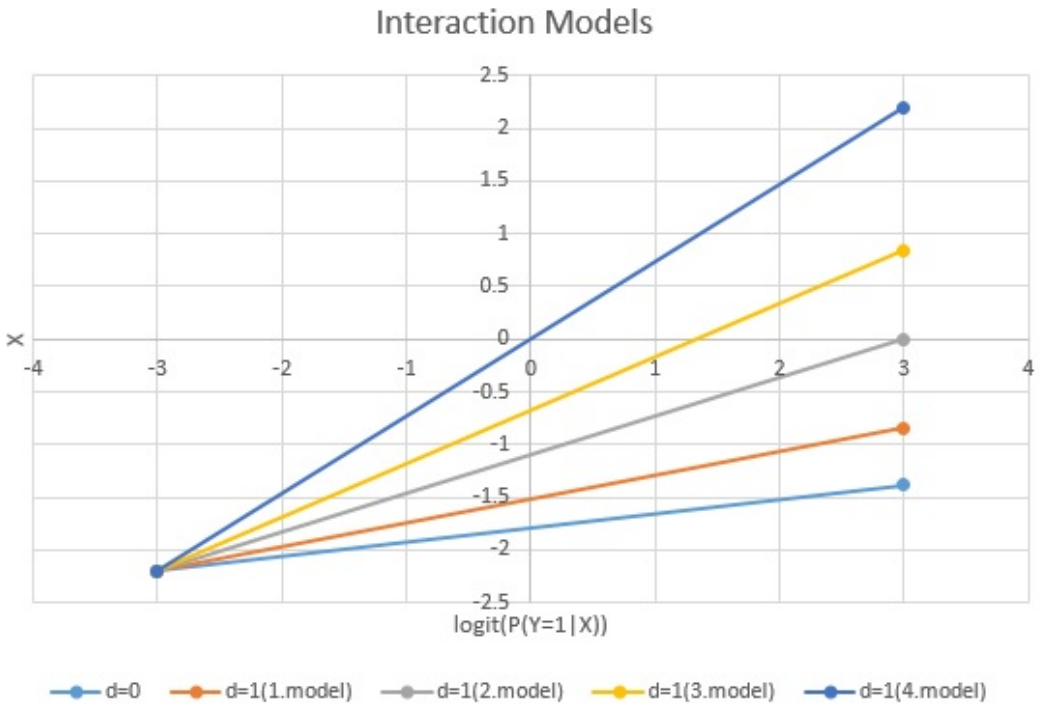

Figure 4.2: The levels of interaction in the logit function

Figure 4.2 shows the interaction levels for four models. The lines are plotted for the dichotomous variable, d, is equal to 0 and 1. If they were parallel for any model setting, it would mean that there is no interaction. In Figure 4.2, all models involve interaction with the dichotomous variable with varying levels. The further away from the parallelism with d=0 line, the more profound interaction a model has. In our setup, the interaction levels increase from model 1 to Model 4. For four generating models, candidate set of fitted models is formed in the following way. The first model is fitted in the same form of the generating model, and the other misses the dichotomous variable and the interaction term. That is, candidate model set consists of the following two model:

$$1. \quad logit(p(Y_i = 1|x_i)) = \beta_0 + \beta_1 x_i + \beta_2 d_i + \beta_3 x d_i$$

$$2. \quad logit(p(Y_i = 1|x_i)) = \beta_0 + \beta_1 x_i$$

Fitted models are compared by using the criteria proposed in Chapter 3.

Consistency

Results are evaluated based on consistency *Definition 1* given in Chapter 2. This defi-

nition suggests that any criterion choosing the true model with probability 1 for large n, will be regarded as a strongly consistent criterion.

Table 4.3: Frequency of selecting the true model by each criterion out of 1000 replicates

Tool	Model 1		Model 2		Model 3		Model 4	
	n=100	n=500	n=100	n=500	n=100	n=500	n=100	n=500
KL	165	559	663	996	907	1000	972	1000
AIC	212	412	504	983	863	1000	998	1000
AIC_c	185	403	471	983	842	1000	996	1000
CAIC	73	128	262	877	657	1000	979	1000
BIC	32	35	146	697	492	994	951	1000
ICOMP	375	644	690	996	947	1000	999	1000
CC_{FM1}	314	164	414	297	561	386	783	599
CC_{FM2}	558	580	671	835	847	932	958	996
CC_{J1}	380	279	473	407	606	504	827	710
CC_{J2}	566	603	678	854	853	941	963	997

Since the generating model is the largest of the candidate set, smaller penalty terms are again more useful. Among the new methods, CC_{FM2} and CC_{J2} are better than others. They are consistent in terms of *Definition1*. Penalty term 1, $plogn/100$, penalizes the model too much, and its performance deteriorates as the sample size increases.

In detecting a missing dichotomous variable and interaction, new methods seem to outperform the common criteria. Only ICOMP performs better than the cluster based criteria. CC_{FM2} and CC_{J2} are successful especially for sample size of 100. AIC and AIC_c are comparable with them for Model 3 (AIC=863, AIC_c=842, CC_{FM2}=847, CC_{J2}=853). CAIC and BIC are comparable even for a more pronounced missingness of interaction(CAIC=979, BIC=951, CC_{FM2}=958, CC_{J2}=963). For sample size of 500, common criteria are better beginning from Model 2. However, new methods also show high rates of choosing the true model. AIC and AIC_c are again better than CAIC and BIC due to smaller penalty terms. Their consistency in such cases is also in the literature. (Aparicio and Villanua, 2007). Their tendency to overfit may also

make them choose the largest model for this scenario.

Efficiency

Efficiency definition is again the same as given in Chapter 2. Average observed efficiency rates are given in Table 4.4. AIC, AIC_c, and ICOMP outperform CAIC and BIC for this scenario, too. CC_{FM2} and CC_{J2} are comparable to those, whereas cluster based criteria with penalty term 1 do not show such performance.

Table 4.4: Average observed efficiency rates

Tool	Model 1		Model 2		Model 3		Model 4	
	n=100	n=500	n=100	n=500	n=100	n=500	n=100	n=500
AIC	0.999	0.998	0.989	0.999	0.996	1	0.999	1
AIC_c	0.999	0.998	0.987	0.999	0.995	1	0.999	1
CAIC	0.999	0.994	0.975	0.998	0.983	1	0.999	1
BIC	0.998	0.992	0.965	0.994	0.968	0.999	0.996	1
ICOMP	0.999	0.999	0.995	0.999	0.999	1	0.999	1
CC_{FM1}	0.998	0.993	0.975	0.973	0.961	0.939	0.960	0.912
CC_{FM2}	0.999	0.997	0.988	0.995	0.989	0.994	0.994	0.999
CC_{J1}	0.998	0.994	0.979	0.978	0.966	0.954	0.969	0.937
CC_{J2}	0.999	0.997	0.988	0.995	0.990	0.995	0.995	0.999

4.1.1.3 Detecting Misspecified Link Function

In this section, the performances of the criteria of our interest are examined in choosing the correct link function. Scenarios are motivated by the settings given in Hosmer et al. (1997) for the power studies of goodness-of-fit tests. Covariates are generated from U(-3, 3). Three different true response data generating mechanisms are considered. These are 1. probit, 2. complementary log-log, and 3. log-log links. The same linear predictor is used for all three models as given in Hosmer et al. (1997).

$$1 \quad . \quad probit(P(Y_i = 1|x_i)) = 0.8x_i$$

$$2 \quad . \quad cloglog(P(Y_i = 1|x_i)) = 0.8x_i$$

$$3 \quad . \quad loglog(P(Y_i = 1|x_i)) = 0.8x_i$$

Candidate model set includes two models: 1. GLM with the true link g and 2. Logistic regression (GLM with logit link). Model selection criteria of Chapter 2 are used and their performances in determining the correct link are assessed. This way, sensitivities of these model selection tools to the correct underlying link are investigated.

$$\begin{aligned} 1. \quad g(p(Y_i = 1|x_i)) &= \beta_1 x_i \\ 2. \quad logit(p(Y_i = 1|x_i)) &= \beta_1 x_i \end{aligned}$$

where g corresponds to probit, complementary log-log, and log-log link function, as the true model.

Consistency

The number of times selecting the true model for four different model settings are given in Table 4.5.

Table 4.5: Frequency of selecting the true model by each criterion out of 1000 replicates

Tool	Model 1(probit)		Model 2(c-loglog)		Model 3(loglog)	
	n=100	n=500	n=100	n=500	n=100	n=500
KL	652	932	850	981	902	999
AIC	612	664	953	1000	925	1000
AIC_c	612	664	953	1000	925	1000
CAIC	612	664	953	1000	925	1000
BIC	612	664	953	1000	925	1000
ICOMP	612	664	953	1000	925	1000
CC_{FM1}	216	365	374	551	381	543
CC_{FM2}	216	365	374	551	381	543
CC_{J1}	216	363	373	551	381	543
CC_{J2}	216	363	373	551	381	543

The performances of the information based criteria are all comparable. They are not very likely to detect the misspecifed link function when the true link is probit. This is also seen in Hosmer et al. (1997)'s paper. On the other hand, their rate of choosing the true link is above 90% for complementary log-log and log-log links. When the true link is complementary log-log as in model 2, all common criteria choose the true

model 953 times out of 1000. When log-log link the true link as in model 3, this number is 925. They are also consistent in the context of *Definition 1*, since they perform better as the sample size increases.

Performances of conventional information based criteria are much superior than those of tree distance based ones.

Efficiency

In the efficiency context, it can be said that all common criteria perform well as seen from Table 4.6. Cluster based criteria seem to be same with them when the true model is probit. This high performance may be again due the similar losses of probit and logit models. They are said to be indistinguishable by model selection tools in literature (Hosmer et. al, 1997). For model 2 and model 3, CC_{FM1} and CC_{J1} outperform the criteria with penalty term 2. However, all of them can still be considered as efficient criteria based on the definition given in Chapter 2.

Table 4.6: Average observed efficiency rates

Tool	Model 1(probit)		Model 2(c-loglog)		Model 3(loglog)	
	n=100	n=500	n=100	n=500	n=100	n=500
AIC	0.999	0.999	0.999	0.999	0.999	0.999
AIC_c	0.999	0.999	0.999	0.999	0.999	0.999
CAIC	0.999	0.999	0.999	0.999	0.999	0.999
BIC	0.999	0.999	0.999	0.999	0.999	0.999
ICOMP	0.999	0.999	0.999	0.999	0.999	0.999
CC_{FM1}	0.998	0.999	0.992	0.999	0.994	0.999
CC_{FM2}	0.998	0.999	0.985	0.992	0.989	0.992
CC_{J1}	0.998	0.999	0.991	0.999	0.993	0.999
CC_{J2}	0.998	0.999	0.984	0.991	0.988	0.991

4.1.1.4 Nested Models

Two models are nested if one can be extended or reduced to the other by changing the number of parameters. These models are widely used and form an important field of modelling. Comparing nested models is an essential problem. It is not so easy to

see if adding one more parameter really contributes to the model or not. Including these models in simulation studies is necessary in order to observe the performances of our criteria. When the candidate set consists of smaller and larger models than the true model, we face with potential overfitting and underfitting problems. Based on the adequacy of penalty term, models can be overfitted or underfitted. Some of the existing criteria tend to overfit due to small penalty terms, such as AIC and AIC_c. These common problems in literature should also be investigated for the clustering based criteria suggested in this thesis. Here simulations were run for sample sizes of 500 and 1000 due to convergence problems.

i. Candidate set of underfitted and overfitted nested models

Firstly, a candidate set including both underfitted and overfitted models along with the true model is constructed. True model has five regressors. True data generating mechanism is given below

$$logit(P(Y_i = 1|x_i)) = 2.5 + 0.5x_{i1} + 0.8x_{i2} + x_{i3} + 1.2x_{i4} - 4.33x_{i5}$$

where y is the binary response variable, and x_1, x_2, x_3, x_4 and x_5 are the covariates generated from U(0,6). Regression coefficients are chosen with the purpose of having an equal proportion of binary groups.

Candidate model set is constructed in the following form.

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

Consistency

The true model is one of the fitted models. The performance of model selection tools is determined by the number of times they select the true model. Model selection tool

is consistent if it chooses the true model almost surely (see *Definition 1*, Chapter 2).

Table 4.7: Frequency of selecting the true model by each criterion out of 1000 replicates

Tool	n=500	n=1000
KL	933	957
AIC	771	760
CAIC	937	967
AIC_c	780	763
BIC	982	994
ICOMP	828	828
CC_{FM1}	1000	1000
CC_{FM2}	938	974
CC_{J1}	997	1000
CC_{J2}	833	912

In Table 4.7, it's seen that BIC and CAIC are good at selecting the true model among both underfitted and overfitted models. Moreover, their selecting probabilities get higher as the sample size increases. AIC and AICc perform poorly in selecting true model probably due to their small penalty terms. AIC's tendency to overfit is mentioned in the literature. (Schwarz, 1978; Bozdogan, 1987). ICOMP's performance is not as good as CAIC and BIC, but it still has a high rate of picking the true model.

Cluster based criteria are very promising for this scenario. They are all good at picking the correct model out of underfitted and overfitted models. Criteria with penalty term 1 outperform the others. It penalizes the unnecessary parameters adequately. If this penalty term were too high, it would also penalize the true parameters and select a smaller model. CC_{FM1} and CC_{J1} are even better than information based criteria. Criteria with penalty term 2 show also higher performance than AIC, AIC_c, and ICOMP. They are comparable with CAIC and BIC. These cluster based criteria become even better with an increase in the sample size. As a result, it can be said that they are consistent according to *Definition 1*.

Efficiency

Efficiency rates are given in Table 4.8. AIC, AIC_c, and ICOMP outperform CAIC and

BIC for this scenario, too. Cluster based criteria show a slightly worse performance than the common criteria. Penalty term 1 and penalty term 2 work more or less in the same way. As the sample size increases, each criterion gets better. Based on the efficiency definition, all criteria can be shown as efficient.

Table 4.8: 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

ii. Candidate set of underfitted nested models

This scenario handles the problem of underfitting. If the penalty terms is too high, that model selection criterion tend to choose smaller models. Therefore, this simulation is helpful to evaluate the model selection criteria for such a case. True model is the same as the previous scenario. Generating model is given by

$$logit(P(Y_i = 1|x_i)) = 2.5 + 0.5x_{i1} + 0.8x_{i2} + x_{i3} + 1.2x_{i4} - 4.33x_{i5}$$

where y is the binary response variable, and x_1, x_2, x_3, x_4 and x_5 are generated from U(0,6). Regression coefficients are again chosen with the purpose of having an equal proportion of binary groups.

Candidate models are given by

1. $logit(P(Y_i = 1|x_i)) = \beta_0 + \beta_1 x_{i1}$
2. $logit(P(Y_i = 1|x_i)) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}$
3. $logit(P(Y_i = 1|x_i)) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}$
4. $logit(P(Y_i = 1|x_i)) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4}$
5. $logit(P(Y_i = 1|x_i)) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5}$

Consistency

For this case candidate model set includes the true model along with the smaller models. Consistency of a model selection tool is again based on *Definition 1*.

Table 4.9: Frequency of selecting the true model by each criterion out of 1000 replicates

Tool	n=500	n=1000
KL	1000	1000
AIC	1000	1000
CAIC	1000	1000
AIC_c	1000	1000
BIC	1000	1000
ICOMP	1000	1000
CC_{FM1}	1000	1000
CC_{FM2}	1000	1000
CC_{J1}	1000	1000
CC_{J2}	1000	1000

As seen in Table 4.9, all tools select the true model in each trial for both the sample size of 500 and 1000. This result is in line with the proposition 1 in Bozdogan and Haughton (1998)'s article. That is when the true model is included in candidate model set of underfitted models, both AIC and BIC are consistent. Furthermore, Aparicio and Villanua (2007) also suggest that both AIC and BIC are consistent when the generating model is largest of the candidate models.

When the candidate set includes smaller models, and the true model, cluster based criteria show the same performance as the common criteria. For both sample sizes, they are able to choose the true model for all trials. If the penalty term 1 was higher than necessary, it would penalize the true model too much, and it would choose a smaller model. The penalty terms for the cluster based criteria seem to be very useful for nested models, when the true model is in the candidate set. Their consistency in the context of *Definition 1* is observable for this case, too.

Efficiency

As seen from Table 4.10, all criteria of interest are able to pick the model with minimum loss for all of the trials.

Table 4.10: Average observed efficiency rates

Tool	n=500	n=1000
AIC	1	1
CAIC	1	1
AIC_c	1	1
BIC	1	1
ICOMP	1	1
CC_{FM1}	1	1
CC_{FM2}	1	1
CC_{J1}	1	1
CC_{J2}	1	1

iii. Candidate set of overfitted nested models

Overfitting is one of the biggest problems for prediction and model selection. Adding more parameters results in variance inflation without necessarily improving the model. This is the rationale behind using penalty terms in model selection criteria. These terms penalize the model selection criterion for including more parameters. However, penalty terms may still fail in handling overfitting problems. The objective of the current investigation here is to evaluate the behaviour of similarity measures as well as the common methods in such cases. In our setup, true model includes two uniform covariates, and the candidate set consists of larger models in addition to true model. Generating model is given by

$$logit(P(Y_i = 1|x_i)) = 2.5 + 0.5x_{i1} - 1.33x_{i2}$$

where y is the binary response variable, and x_1 and x_2 follows U(0,6). Regression coefficients are chosen with the same manner as the previous nested models.

Set of candidate models include

1. $logit(P(Y_i = 1|x_i)) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}$
2. $logit(P(Y_i = 1|x_i)) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}$
3. $logit(P(Y_i = 1|x_i)) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4}$
4. $logit(P(Y_i = 1|x_i)) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5}$
5. $logit(P(Y_i = 1|x_i)) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \beta_6 x_{i6}$
6. $logit(P(Y_i = 1|x_i)) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_4 + \beta_5 x_{i5} + \beta_6 x_{i6} + \beta_7 x_{i7}$

Consistency

In this scenario, the generating model is the smallest fitted model in the candidate models. All the other fitted models have more covariates. Since true model is in the set of candidate set, *Definiton 1* for consistency is used.

Table 4.11: Frequency of selecting the true model by each criterion out of 1000 replicates

Tool	n=500	n=1000
KL	974	985
AIC	709	719
CAIC	937	953
AIC_c	724	722
BIC	983	989
ICOMP	777	784
CC_{FM1}	998	1000
CC_{FM2}	868	925
CC_{J1}	997	1000
CC_{J2}	822	892

Results in Table 4.11 indicate only BIC and CAIC handles overfitting problem. AIC is known to be inconsistent for such cases. Bozdogan and Haughton (1998) and Aparicio and Villanua (2007) show that when the candidate model set includes true model and the overfitted models, BIC is consistent while AIC is not.

CC_{FM1} and CC_{J1} are better than all other criteria for this case. CC_{FM2} and CC_{J2} also show remarkable performances. They outperform AIC, AIC_c, and ICOMP. When the true model is the smallest among the candidate set, all the cluster based

criteria managed to pick the true model. Penalty terms seem to handle the overfitting problem in nested models.

Efficiency

Table 4.12: Average observed efficiency rates

Tool	n=500	n=1000
AIC	0.990	0.995
CAIC	0.986	0.993
AIC_c	0.990	0.995
BIC	0.985	0.992
ICOMP	0.988	0.994
CC_{FM1}	0.984	0.992
CC_{FM2}	0.985	0.992
CC_{J1}	0.985	0.992
CC_{J2}	0.984	0.992

Results for the efficiency are given in Table 4.12. When the true model is the smallest of the candidate set, all model selection methods perform well in terms of efficiency. AIC, AIC_c and ICOMP are barely better than BIC and CAIC. Cluster based criteria are slightly worse than the common criteria. Their performances do not differ for penalty term 1 and penalty term 2. Moreover, all gets better with an increase in the sample size.

4.1.1.5 Random Effects Models

As the last part for scenarios in which the true model is included in the candidate models, random effects models are conducted. It is essential to investigate the performances of the model selection criteria in random effects models.

i. Candidate set of random effects models

The abilities of information based criteria and cluster based criteria to recognize the true random relation in a logistic model are evaluated in this part. Generating random intercept model is given by

$$logit(P(Y_{ij}=1|x_{ij})) = b_{0i} + 2.5 + 0.5x_{ij1} - 1.33x_{ij2}$$

where i=1,...,10, j=1,...,n_i, $n_i = \frac{n}{10}$. Candidate model set includes the true model and a random slope model as given by

$$\begin{aligned} 1. \quad logit(P(Y_{ij}=1|x_{ij})) &= b_{0i} + \beta_0 + \beta_1 x_{ij1} + \beta_2 x_{ij2} \\ 2. \quad logit(P(Y_{ij}=1|x_{ij})) &= b_{0i} + b_{1i}x_{i1} + \beta_0 + \beta_1 x_{ij1} + \beta_2 x_{ij2} \end{aligned}$$

where b_0 is the random intercept, and b_1 is the random slope. Random intercept follows a normal distribution with zero mean and a constant variance, σ_0^2, for the true model. Random slope model is fitted with the assumption that (b_{0i}, b_{1i}) follows a multivariate normal distribution. Regression coefficients are chosen as in the nested model settings.

The true model is generated under three different scenarios for ICC's. These three scenarios correspond to ICC=0.3 (low intra-class correlation), 0.5 (mild intra-class correlation), and 0.8 (high intra-class correlation) respectively. Given ICC's, true σ_0^2 are computed as 1.408, 1.972, and 4.929 respectively. The frequency of each model selection criteria selecting the true model is given in Table 4.13.

Consistency

Table 4.13: Frequency of selecting the true model by each criterion out of 1000 replicates

Tool	ICC=0.3		ICC=0.5		ICC=0.8	
	n=500	n=1000	n=500	n=1000	n=500	n=1000
KL	842	888	865	887	850	882
AIC	932	944	946	943	937	933
AICc	935	945	950	944	940	934
CAIC	991	994	996	996	994	997
BIC	999	1000	1000	1000	1000	1000
ICOMP	165	209	176	173	159	116
CC_{FM1}	1000	1000	1000	1000	1000	1000
CC_{FM2}	965	992	979	990	972	985
CC_{J1}	1000	1000	1000	1000	1000	1000
CC_{J2}	940	978	938	979	928	968

Results given in Table 4.13 are evaluated according to the *Definition 1*, as the true model is in the set of candidate models. It is observed from Table 4.13, CAIC and BIC are consistent for all ICC levels. (Their performances improve as sample size increases). Also they perform better than AIC and AIC_c. AIC and AIC_c seem to be consistent for low intra-class correlation (namely ICC=0.3). ICOMP is not able distinguish between two random effects models. Its performance even worsen as the ICC level increases.

Cluster based criteria perform well for all ICC. Both with penalty term 1 and penalty term 2 are successful. However, since the true model is the smallest of the candidates, one with a larger penalty term outperforms the other. CC_{FM1} and CC_{J1} are able to pick the true model for all of the trials. The performances of CC_{FM2} and CC_{J2} get better as the sample size increases. Hence, we consider cluster based criteria as consistent based on *Definition2*.

Efficiency

As seen from Table 4.14, ICOMP performs better than all other criteria in terms of efficiency. Performances of any criteria do not change with ICC. AIC and AIC_c outperform CAIC and BIC. Cluster based criteria are more or less the same with CAIC and BIC. In overall context, they are all efficient based on the definition given in section 2.3.2.

Table 4.14: Average observed efficiency rates

Tool	ICC=0.3		ICC=0.5		ICC=0.8	
	n=500	n=1000	n=500	n=1000	n=500	n=1000
AIC	0.993	0.997	0.993	0.997	0.993	0.996
AICc	0.993	0.997	0.993	0.997	0.993	0.996
CAIC	0.991	0.996	0.991	0.996	0.991	0.995
BIC	0.991	0.996	0.991	0.996	0.990	0.995
ICOMP	0.999	0.999	0.999	0.999	0.999	0.999
CC_{FM1}	0.991	0.996	0.991	0.996	0.990	0.995
CC_{FM2}	0.991	0.996	0.991	0.996	0.992	0.995
CC_{J1}	0.991	0.996	0.991	0.996	0.990	0.995
CC_{J2}	0.991	0.996	0.992	0.996	0.992	0.995

ii. Candidate set of random effects and fixed effects models

As another scenario, candidate set of fitted models involves a fixed effect model along with the true random intercept model. Criteria are evaluated for detecting a missing random term in logistic models. The generating model is the same.

$$logit(P(Y_{ij} = 1|x_{ij})) = b_{0i} + 2.5 + 0.5x_{ij1} - 1.33x_{ij2}$$

Candidate set of models is given by

$$\begin{aligned} 1. \quad logit(P(Y_{ij} = 1|x_{ij})) &= b_{0i} + \beta_0 + \beta_1 x_{ij1} + \beta_2 x_{ij2} \\ 2. \quad logit(P(Y_{ij} = 1|x_{ij})) &= \beta_0 + \beta_1 x_{ij1} + \beta_2 x_{ij2} \end{aligned}$$

Consistency

Table 4.15: Frequency of selecting the true model by each criterion out of 1000 replicates

Tool	ICC=0.3		ICC=0.5		ICC=0.8	
	n=500	n=1000	n=500	n=1000	n=500	n=1000
KL	996	1000	999	1000	1000	1000
AIC	989	1000	995	1000	1000	1000
AICc	989	1000	995	1000	1000	1000
CAIC	978	997	989	999	999	1000
BIC	963	995	985	998	999	1000
ICOMP	999	1000	1000	1000	1000	1000
CC_{FM1}	279	172	492	357	884	852
CC_{FM2}	862	848	915	939	994	997
CC_{J1}	423	302	607	498	924	916
CC_{J2}	889	883	933	954	997	998

According to Table 4.15, among the common criteria, ICOMP seems to the best. AIC, and AIC_c outperform CAIC, and BIC. This may be due to their tendency for overfitting. CAIC and BIC are also good at detecting the missingness of a random

term. The performances of all common criteria gets better as ICC and the sample size increase. They are consistent in terms of consistency *Definition 1*.

CC_{FM1} and CC_{J1} stay far behind CC_{FM2} and CC_{J2}. Penalty term 1 is too high for such cases. It is also observed that they get worse as the sample size increases. However, when ICC gets high for the true model, their ability to picking the random intercept model increases as obviously seen from Table 4.15. Cluster based criteria with penalty term 2 is much better than those. Their performance falls with an increase in the sample size for a true ICC of 0.3, but they get better for larger ICC.

Efficiency

Average observed efficiency rates for this scenario are given in Table 4.16. ICOMP outperforms others as in the previous scenario. Each information based criterion has very high efficiency rates. AIC and AIC_c are a little better than CAIC and BIC. Cluster based criteria are not as successful as information criteria for this case. Their efficiency rates increase as ICC increases. CC_{FM2} and CC_{J2} perform better than CC_{FM1} and CC_{J1}. This may be due to a higher penalty term. The true model is the largest in the candidate set and CC_{FM1} and CC_{J1} penalize it more than necessary. This is due a higher penalty term, $plogn/100$.

Table 4.16: Average observed efficiency rates

Tool	ICC=0.3		ICC=0.5		ICC=0.8	
	n=500	n=1000	n=500	n=1000	n=500	n=1000
AIC	0.999	1	0.999	1	1	1
AICc	0.999	1	0.999	1	1	1
CAIC	0.999	0.999	0.999	0.999	0.999	1
BIC	0.998	0.999	0.999	0.999	0.999	1
ICOMP	1	1	1	1	1	1
CC_{FM1}	0.887	0.870	0.910	0.883	0.974	0.965
CC_{FM2}	0.986	0.984	0.990	0.993	0.999	0.999
CC_{J1}	0.916	0.898	0.934	0.916	0.984	0.982
CC_{J2}	0.989	0.988	0.992	0.996	0.999	0.999

4.1.2 True Model is not in the Set of Candidate Models

For real life cases it is not possible to know the true model. For the simulations in this section, the data set is generated from a true model. However, this true model is not included in the candidate model set. The evaluation of model selection criteria is now based on consistency *Definition 2*. Kullback-Liebler distance is used as the main reference. Criteria that chooses the model with the smallest KL distance to the truth are regarded as successful. For efficiency, the definition is the same for all cases.

4.1.2.1 Nonlinear Model Study

The model which perfectly explains the relation between the response and the covariate is generally something very complicated. The best approach for such cases is to obtain the closest approximation. It is essential to observe the performances of the model selection criteria in order to obtain the best approximation.

We used the same true model given in (Kalaylioglu and Ozturk, 2013)

$$logit(P(Y_i = 1|x_i)) = 3 - 1.5x_i + 5(0.5 - 1/(1 + (x_i + 1)^4)$$

where y is the binary response variable, and x follows U(0,6). Regression coefficients are chosen with the purpose of having an equal proportion of binary groups.

Candidate models are given by

$$\begin{aligned}
1.\quad & logit(P(Y_i = 1|x_i)) &=& \beta_0 + \beta_1 x_i \\
2.\quad & logit(P(Y_i = 1|x_i)) &=& \beta_0 + \beta_1 x_i + \beta_2 x_i^2 \\
3.\quad & logit(P(Y_i = 1|x_i)) &=& \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x^3
\end{aligned}$$

These candidate models are very close to each other as from Figure 4.3. Simulations are replicated for 1000 times. Regression coefficients obtained from the first trial are used to visualize the shapes of candidate models and the true model.

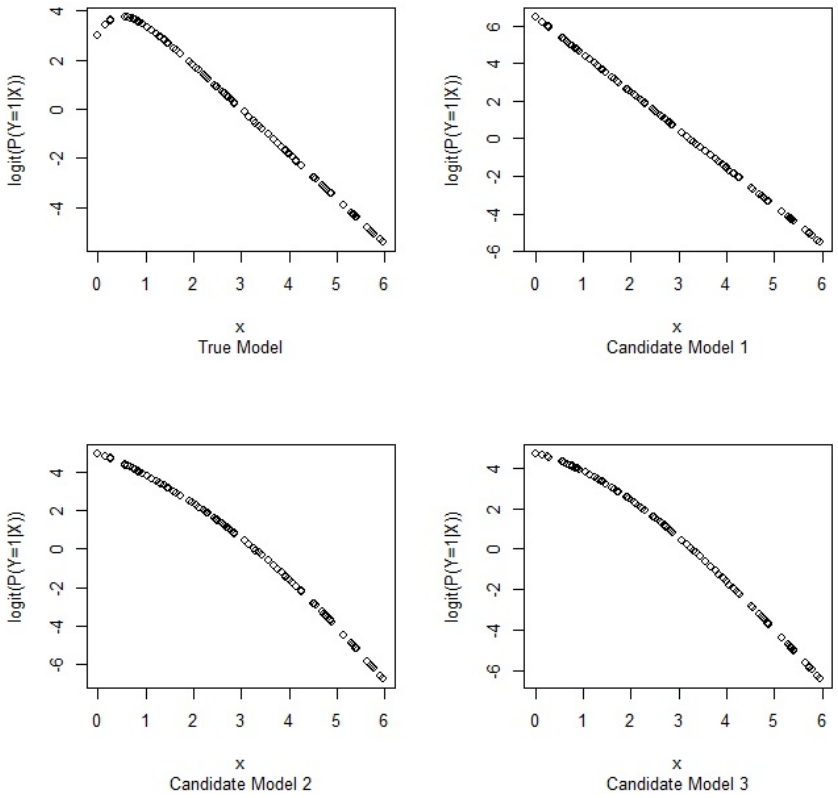

Figure 4.3: Illustration of true model and candidate models

Consistency

For this case, the consistency of the model selection criteria are associated with the Kullback-Liebler distance. The *Definition 3* for consistency will be used. In order to use the theorem given in 2.3.1, the columns of the covariate matrix should be linearly independent. Orthogonal polynomials are used to satisfy this condition (Rawlings et al., 1998). Candidate models are fitted with those orthogonal covariates. The number of times selecting the model with minimum KL distance are in Table 4.17.

According to Table 4.18, none of the model selection criteria perform satisfactorily in terms of consistency. For small sample sizes the cluster based criteria seem to outperform the common ones. For information based criteria, the problem seems to be related to convergence. Their performance fluctuate as the sample size increases. Cluster based criteria, on the other hand, worsen regularly as the sample size increases. This rises the need for a further investigation on penalty terms of cluster

based criteria.

Table 4.17: Frequency of selecting the model with minimum KL distance by each criterion out of 1000 replicates

Tool	n=100	n=250	n=500	n=750	n=1000	n=1500	n=5000	n=10000
AIC	414	416	359	301	293	258	307	481
CAIC	535	523	478	444	392	317	187	377
AIC_c	432	423	360	302	296	258	307	481
BIC	596	576	532	514	469	416	183	242
ICOMP	602	563	515	480	426	362	150	270
CC_{FM1}	684	640	607	591	562	516	347	282
CC_{FM2}	631	618	602	588	563	515	347	282
CC_{J1}	679	640	607	591	562	516	347	282
CC_{J2}	624	602	599	580	558	515	346	282

Efficiency

Table 4.18: Average observed efficiency rates

Tool	n=100	n=250	n=500	n=750	n=1000	n=1500	n=5000	n=10000
AIC	0.990	0.997	0.998	0.999	0.999	0.999	0.999	0.999
CAIC	0.984	0.995	0.997	0.998	0.999	0.999	0.999	0.999
AIC_c	0.990	0.996	0.998	0.999	0.999	0.999	0.999	0.999
BIC	0.979	0.993	0.996	0.998	0.998	0.999	0.999	0.999
ICOMP	0.976	0.993	0.997	0.998	0.998	0.999	0.999	0.999
CC_{FM1}	0.972	0.991	0.996	0.997	0.997	0.998	0.999	0.999
CC_{FM2}	0.977	0.992	0.996	0.997	0.997	0.998	0.999	0.999
CC_{J1}	0.972	0.991	0.996	0.997	0.997	0.998	0.999	0.999
CC_{J2}	0.977	0.992	0.996	0.997	0.997	0.998	0.999	0.999

Average observed efficiency rates are given in Table 4.18. In the context of efficiency, all criteria perform well. Among the information based criteria, AIC and AIC_c outperform CAIC, BIC, and ICOMP for smaller sample sizes. All get better and become comparable with each other as the sample size increases.

Cluster based criteria are a little less successful than common criteria for small sample sizes. They are more or less the same for larger sample sizes. When the sample size is 100 and 250, CC_{FM2} and CC_{J2} are better than CC_{FM1} and CC_{J1}. Unlike consistency, the efficiency rates of model selection criteria converges easily for small

sample sizes. All can be regarded as efficiency according to the efficiency definition given in section 2.3.1.2.

4.1.2.2 Nested Models

Nested models are conducted in this section in the same way as in section 4.1.4. Model selection criteria are evaluated for their performances in overfitting and underfitting problems.

i. Candidate set of overfitted and underfitted nested models

The first case is again the set of overfitted and underfitted models. Generating model is given by

$$logit(P(Y_i = 1|x_i)) = 2.5 + 0.5x_{i1} + 0.8x_{i2} + x_{i3} + 1.2x_{i4} - 4.33x_{i5}$$

where y is the binary response variable, and x_1, x_2, x_3, x_4, and x_5 follow U(0,6). Regression coefficients are chosen with the purpose of having an equal proportion of binary groups.

As seen in the following candidate models, now the generating model is not included.

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

Consistency

Evaluating the consistency of the criteria based on Table 4.19, *Definition 2* is used. 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 ∞.

Table 4.19: Frequency of selecting the model with minimum KL distance by each criterion out of 1000 replicates

Tool	n=500	n=1000	n=1500
AIC	773	781	797
CAIC	897	907	930
AIC_c	782	782	798
BIC	939	937	958
ICOMP	813	816	829
CC_{FM1}	948	924	815
CC_{FM2}	884	925	950
CC_{J1}	946	945	966
CC_{J2}	795	862	875

In this simulation study, model 5 had the minimum expected KL distance around 950 times out of 1000 replicates.

According to Table 4.19, among the conventional model selection criteria, BIC and CAIC show the highest performance in terms of agreeing with Kullback-Leibler distance. This is again a sign for handling the ovetfitting problem better than the others. As seen in the table, BIC slightly decreases with increasing sample size. This may be because BIC is based on the assumption that true model is in the candidate set. (Schwarz, 1978).

Cluster based criteria perform well. They guard against overfitting unlike AIC and AIC_c. Their performances are more satisfactory than ICOMP and comparable with CAIC and BIC. However, they do not show a significant increase in their performance with an increase in the sample size. This may be due to using KL distance as the reference. These cluster based criteria are not based on minimizing KL distance. Therefore, it may not be reasonable to evaluate them in reference to minimum KL distance.

Efficiency

Results given in Table 4.20 are very similar to those given in previous section. AIC, AIC_c, and ICOMP perform better than other criteria. Cluster based criteria with penalty term $ploglogn/100$ is slightly better than those with $plogn/100$ for this scenario. CC_{FM1}'s rate of efficiency decreased with an increase in the sample size.

Dependence on the sample size with a higher rate can be shown as the reason for that.

Table 4.20: Average observed efficiency rates

Tool	n=500	n=1000
AIC	0.995	0.998
CAIC	0.992	0.996
AIC_c	0.995	0.998
BIC	0.990	0.995
ICOMP	0.994	0.995
CC_{FM1}	0.988	0.975
CC_{FM2}	0.990	0.995
CC_{J1}	0.990	0.995
CC_{J2}	0.992	0.996

ii. Candidate set of underfitted nested models

Objective is to evaluate the performances of model selection criteria in dealing with underfitting problem. Generating model is the same as given in the following.

$$logit(P(Y_i = 1|x_i)) = 2.5 + 0.5x_{i1} + 0.8x_{i2} + x_{i3} + 1.2x_{i4} - 4.33x_{i5}$$

where y is the binary response variable, and x_1, x_2, x_3, x_4, and x_5 follow U(0,6). Regression coefficients are chosen with the purpose of having an equal proportion of binary groups.

Candidate models are given below

1. $logit(P(Y_i = 1|x_i)) = \beta_0 + \beta_{i1}x_1$
2. $logit(P(Y_i = 1|x_i)) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}$
3. $logit(P(Y_i = 1|x_i)) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}$
4. $logit(P(Y_i = 1|x_i)) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4}$

Consistency

Consistency definition based on the Kullback-Leibler distance is again used here, which is *Definition 2*. A model selection tool is consistent if the probability of selecting the model with the smallest Kullback-Leibler distance goes to 1 as the sample

size goes to infinity.

Table 4.21: Frequency of selecting the model with minimum KL distance by each criterion out of 1000 replicates

Tool	n=500	n=1000
AIC	982	1000
CAIC	912	1000
AIC_c	982	1000
BIC	770	996
ICOMP	971	1000
CC_{FM2}	160	140
CC_{J2}	129	109

In this simulation study, model 4 had the minimum expected KL distance for all the replicates.

In this scenario overfitted models are not in the candidate set. That is probably the reason of better performances of AIC, CAIC, and AIC_c. These criteria fail to handle overfitting problems due to an inadequate penalty term. BIC also shows a good result. Bozdogan (1998) suggests that when the true model is not included in the set of candidate models, both AIC and BIC are consistent when the true model is compared to smaller models. The fact that the other information based criteria works better than BIC in this situation may be because BIC is based on the assumption of having the true model in the candidate set (Schwarz, 1978).

Cluster based criteria do not seem to be useful for this scenario as seen from Table 4.21. Corresponding frequencies are strikingly low. CC_{FM1} and CC_{J1} never agreed with minimum KL reference (not shown in the tables). In order to understand the reason behind, this simulation is examined in more detail. Accordingly, model 4 has the minimum KL distance in all 1000 replicates. On the other hand, CC_{FM1} and CC_{J1} select model 1 900 times out of 1000 replicates. We found out that when the fitted model does not include x_5, the most significant factor in the true model, these clustering based criteria do not regard the model as a good fit, and judge them only for the number of parameters. In order to investigate this further, we carefully designed two additional simulation experiments: 1. One of the candidate models include x_5, 2.

All the models in the candidate set include x_5.

1. Only one candidate model includes x_5 Candidate models are given below.

$$
\begin{aligned}
1. \quad & logit(P(Y_i = 1|x_i)) &=& \beta_0 + \beta_{i1}x_{i2} \\
2. \quad & logit(P(Y_i = 1|x_i)) &=& \beta_0 + \beta_1 x_{i2} + \beta_2 x_{i3} \\
3. \quad & logit(P(Y_i = 1|x_i)) &=& \beta_0 + \beta_1 x_{i2} + \beta_2 x_{i3} + \beta_3 x_{i4} \\
4. \quad & logit(P(Y_i = 1|x_i)) &=& \beta_0 + \beta_1 x_{i2} + \beta_2 x_{i3} + \beta_3 x_{i4} + \beta_4 x_{i5}
\end{aligned}
$$

Table 4.22 indicates that all criteria picked the model with minimum KL distance. Here, model 4 has the minimum KL distance in all the 1000 replicates. Therefore, when the most effective factor is in the fitted model, our cluster based criteria are useful.

Table 4.22: Frequency of selecting the model with minimum KL distance by each criterion out of 1000 replicates

Tool	n=500	n=1000
AIC	1000	1000
CAIC	1000	1000
AIC_c	1000	1000
BIC	1000	996
ICOMP	1000	1000
CC_{FM1}	1000	1000
CC_{FM2}	1000	1000
CC_{J1}	1000	1000
CC_{J2}	1000	1000

2. All the models in the candidate set include x_5

Candidate models for this situation are given below.

$$
\begin{aligned}
1. \quad & logit(P(Y_i = 1|x_i)) &=& \beta_0 + \beta_{i1}x_{i5} \\
2. \quad & logit(P(Y_i = 1|x_i)) &=& \beta_0 + \beta_1 x_{i4} + \beta_2 x_5 \\
3. \quad & logit(P(Y_i = 1|x_i)) &=& \beta_0 + \beta_1 x_{i3} + \beta_2 x_{i4} + \beta_3 x_{i5} \\
4. \quad & logit(P(Y_i = 1|x_i)) &=& \beta_0 + \beta_1 x_{i2} + \beta_2 x_{i3} + \beta_3 x_{i4} + \beta_4 x_{i5}
\end{aligned}
$$

Table 4.23: Frequency of selecting the model with minimum KL distance by each criterion out of 1000 replicates

Tool	n=500	n=1000
AIC	1000	1000
CAIC	1000	1000
AIC_c	1000	1000
BIC	1000	1000
ICOMP	1000	1000
CC_{FM2}	605	617
CC_{J2}	777	829

Common criteria show a high success as expected. They are again able to select the model with minimum KL distance, which is model 4. Cluster based criteria with penalty term 1 failed for this scenario, too (results not shown in the table). They seem to be ignoring the significance of other factors, and penalize them more than necessary. They always pick the smaller models. Penalty term 2 is more reasonable. Even if it does not work as well as the common criteria, it is able to choose the largest model for most of the times.

Efficiency

Efficiency rates are obtained for the scenario in which x_5 is not included in any of the candidate models. Table 4.24 presents the results.

Table 4.24: Average observed efficiency rates

Tool	n=500	n=1000
AIC	0.999	1
CAIC	0.999	1
AIC_c	0.999	1
BIC	0.996	0.999
ICOMP	0.999	1
CC_{FM1}	0.940	0.942
CC_{FM2}	0.959	0.960
CC_{J1}	0.940	0.942
CC_{J2}	0.957	0.957

The common tools are all good at selecting the best approximation in terms of effi-

ciency. They can be called as efficient for this study, too. We know for this model neither CC_{FM1} or CC_{J1} were able to pick the true model. However, in Table 4.24, average observed efficiency rates are around 95%. When the interest is loss, which is defined as average squared distance between observed and predicted values, models are not very distinct from each other. That is why these efficiency rates are too high. On the other hand, it is obviously seen that they are very much lower than the efficiency rates of the common criteria. Moreover, higher dependence on the sample size again create a drawback for criteria with penalty term 1.

iii. Candidate set of overfitted nested models

As the last scenario of the nested models, overfitting problem is addressed. Data generation model is given by

$$logit(P(Y_i = 1|x_i)) = 2.5 + 0.5x_{i1} - 1.33x_{i2}$$

where y is the binary response variable, and x_1 and x_2 follow U(0,6). Regression coefficients are chosen with the purpose of having an equal proportion of binary groups.

Candidate model set contains the following models.

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

Consistency

Candidate set does not include the true model and includes the models that are larger than the (unknown) true model. Here the model with minimum Kullback-Leibler distance can be regarded as the best approximating model. Therefore, each method is examined in terms of its agreement with Kullback-Leibler distance. The consistency of criteria will be based on *Definition 2*. Frequency of each model selection criterion selecting the model with minimum KL distance out of 1000 replicates is given in Table 4.25.

Table 4.25: Frequency of selecting the model with minimum KL distance by each criterion out of 1000 replicates

Tool	n=500	n=1000
AIC	715	728
CAIC	914	939
AIC_c	729	734
BIC	957	974
ICOMP	772	783
CC_{FM1}	968	984
CC_{FM2}	840	916
CC_{J1}	968	984
CC_{J2}	802	885

As seen in in Table 4.25, BIC and CAIC again handles overfitting problem for the cases in which candidate set does not include the true model. AIC and AIC_c's tendency to overfit is also obvious.

Results in Table 4.25 show similar results with Table 4.19. CC_{FM1} and CC_{J1} outperform all other methods. Although cluster based criteria with penalty type 2 tends to overfit, their performances are better than AIC, AIC_c, and ICOMP. Frequency of selecting the model with minimum KL distance increases as the sample size increases. Hence, it can be said that these cluster based criteria are consistent in terms of *Definition 2*.

Efficiency

Table 4.26: Average observed efficiency rates

Tool	n=500	n=1000
AIC	0.992	0.996
CAIC	0.989	0.994
AIC_c	0.992	0.996
BIC	0.988	0.994
ICOMP	0.991	0.996
CC_{FM1}	0.987	0.994
CC_{FM2}	0.988	0.994
CC_{J1}	0.987	0.994
CC_{J2}	0.988	0.994

Average observed efficiency rates are given in Table 4.26.

When modeling is bound to overfitted models, all criterion showed a good performance. As in the previous sections, AIC, AIC_c, and $ICOMP$ outperform all other methods. Their efficiency mentioned in the literature is observed for this case, too. For the cluster based criteria, penalty terms do not differ too much for this scenario. Their efficiency rates are almost the same as BIC and ICOMP.

4.2 Modeling Purpose is Classification

In this section we focus on the scenarios in section 4.1 in which information based and classification based model selection criteria were comparable in terms of assessing the fit of the models. Each criterion is evaluated in terms of TCR, specificity and sensitivity. In each iteration of Monte Carlo simulations, each model selection criterion selects one of the candidate models. TCR, specificity and sensitivity are calculated for those models. Their Monte Carlo averages are given in the preceding tables.

4.2.1 True Model is in the set of Candidate Models

Firstly, we examine scenarios in which the true model is in the set of candidate models. We conduct the same simulation settings in which cluster based criteria show promising results.

4.2.1.1 Detecting Missing Interaction Terms

Firstly, we evaluated each criterion by using the simulation settings from section 4.1.2. Table 4.27 presents the results for Monte Carlo averages of TCR.

Table 4.27: Monte Carlo Average of TCR for Each Criterion

Tool	Model 1		Model 2		Model 3		Model 4	
	n=100	n=500	n=100	n=500	n=100	n=500	n=100	n=500
TCR of true model	0.593	0.584	0.611	0.620	0.643	0.650	0.680	0.676
AIC	0.556	0.556	0.586	0.619	0.634	0.650	0.680	0.676
AIC_c	0.554	0.555	0.584	0.619	0.632	0.650	0.680	0.676
CAIC	0.545	0.539	0.569	0.612	0.617	0.650	0.680	0.676
BIC	0.537	0.529	0.557	0.600	0.604	0.650	0.679	0.676
ICOMP	0.569	0.573	0.597	0.620	0.642	0.650	0.680	0.676
CC_{FM1}	0.583	0.553	0.591	0.584	0.618	0.604	0.664	0.637
CC_{FM2}	0.597	0.588	0.615	0.619	0.643	0.648	0.678	0.676
CC_{J1}	0.588	0.564	0.597	0.592	0.622	0.618	0.670	0.648
CC_{J2}	0.596	0.586	0.614	0.621	0.643	0.648	0.680	0.676

Table 4.27 also show average TCR for the true model. True model itself does not have a high classification rate. The highest TCR is around 68 %, which occurs when the true model is Model 4. Monte Carlo averages of TCR for each criterion are very close to true model's TCR. Overall CC_{FM2} and CC_{J2} outperform all the others.Among the information based criteria, ICOMP outperform the others. AIC, AIC_c, CAIC, and BIC show similar results with ICOMP for Model 3 (when the sample size is 500) and for Model 4. Among cluster based criteria CC_{FM2} and CC_{J2} perform better than CC_{FM1} and CC_{J1}.

Table 4.28: Monte Carlo average of sensitivity for each criterion

Tool	Model 1		Model 2		Model 3		Model 4	
	n=100	n=500	n=100	n=500	n=100	n=500	n=100	n=500
sensitivity of true model	0.420	0.416	0.441	0.421	0.487	0.475	0.568	0.558
AIC	0.471	0.453	0.467	0.421	0.490	0.475	0.569	0.558
AIC_c	0.472	0.454	0.468	0.421	0.491	0.475	0.571	0.558
CAIC	0.474	0.475	0.478	0.429	0.496	0.475	0.569	0.558
BIC	0.478	0.482	0.480	0.440	0.506	0.475	0.573	0.558
ICOMP	0.456	0.434	0.453	0.421	0.488	0.475	0.568	0.558
CC_{FM1}	0.429	0.451	0.438	0.448	0.471	0.476	0.549	0.529
CC_{FM2}	0.420	0.411	0.430	0.418	0.476	0.468	0.564	0.558
CC_{J1}	0.426	0.434	0.435	0.439	0.470	0.473	0.554	0.534
CC_{J2}	0.419	0.411	0.429	0.418	0.476	0.468	0.565	0.558

From Table 4.28, it is seen that information based criteria are better than cluster based criteria in terms of sensitivity. Only for Model 3 and Model 4, they are comparable with the common criteria. All criteria are also comparable with the sensitivity for true model. In some cases, sensitivity rates for information based criteria are greater than sensitivity of true model.

Results for Monte Carlo averages of specificity rates are given in Table 4.29. Cluster based criteria perform better than information based criteria in terms of specificity. CC_{FM2} and CC_{J2} outperform CC_{FM1} and CC_{J1}, as well as the common criteria especially for sample size of 100. They are also greater than specificity of true model for some cases.

Table 4.29: Monte Carlo average of specificity for each criterion

Tool	Model 1		Model 2		Model 3		Model 4	
	n=100	n=500	n=100	n=500	n=100	n=500	n=100	n=500
specificity of true model	0.647	0.638	0.673	0.690	0.705	0.712	0.725	0.724
AIC	0.579	0.591	0.628	0.689	0.692	0.712	0.724	0.724
AIC_c	0.578	0.589	0.626	0.689	0.671	0.712	0.723	0.724
CAIC	0.566	0.589	0.604	0.676	0.689	0.712	0.722	0.724
BIC	0.556	0.548	0.588	0.658	0.647	0.712	0.719	0.724
ICOMP	0.602	0.618	0.651	0.689	0.703	0.712	0.724	0.724
CC_{FM1}	0.627	0.586	0.652	0.629	0.678	0.649	0.712	0.683
CC_{FM2}	0.648	0.640	0.682	0.690	0.710	0.712	0.724	0.724
CC_{J1}	0.634	0.607	0.660	0.644	0.685	0.668	0.718	0.697
CC_{FM2}	0.648	0.640	0.683	0.691	0.710	0.713	0.726	0.724

4.2.1.2 Nested Models

As another scenario, settings from section 4.1.4 are used. Monte Carlo averages of TCR, sensitivity and specificity for each criterion are given in three parts: i. when the candidate set includes both overfitted and underfitted models, ii. when the candidate set includes underfitted models, iii. when the candidate set includes overfitted models.

As seen in Table 4.30, all criteria have high TCR for nested models. Overall performances of CC_{FM2} and CC_{J2} are again better than the others. Monte Carlo average of TCR for each criterion are more or less the same with TCR for true model. For

the set of overfitted candidate models, cluster based criteria give higher TCR than the true model. Overall CC_{FM2} and CC_{J2} have higher rates than the true model outperforming the rest in terms of TCR.

Table 4.30: Monte Carlo Average of TCR for Each Criterion

Tool	overfitted and underfitted		underfitted		overfitted	
	n=500	n=1000	n=500	n=1000	n=500	n=1000
TCR of true model	0.914	0.894	0.914	0.894	0.721	0.710
AIC	0.912	0.895	0.914	0.894	0.724	0.710
AIC_c	0.913	0.895	0.914	0.894	0.723	0.710
CAIC	0.913	0.893	0.914	0.894	0.725	0.710
BIC	0.914	0.894	0.914	0.894	0.724	0.712
ICOMP	0.913	0.895	0.914	0.894	0.723	0.711
CC_{FM1}	0.914	0.894	0.914	0.894	0.724	0.712
CC_{FM2}	0.916	0.894	0.914	0.894	0.728	0.713
CC_{J1}	0.914	0.894	0.914	0.894	0.724	0.712
CC_{J2}	0.916	0.895	0.914	0.894	0.728	0.714

Sensitivity rates are given in Table 4.31 for this scenario. Rates for criteria are comparable with true model. They are all able to select the model with sensitivity rate close to the true model. For the set of overfitted candidate models, CC_{FM2} and CC_{J2} have higher rates than the true model outperforming the rest in terms of sensitivity.

Table 4.31: Monte Carlo average of sensitivity for each criterion

Tool	overfitted and underfitted		underfitted		overfitted	
	n=500	n=1000	n=500	n=1000	n=500	n=1000
sensitivity of true model	0.913	0.893	0.913	0.893	0.713	0.703
AIC	0.913	0.895	0.913	0.893	0.716	0.702
AIC_c	0.914	0.895	0.913	0.893	0.715	0.700
CAIC	0.913	0.892	0.913	0.893	0.716	0.701
BIC	0.914	0.893	0.913	0.893	0.713	0.702
ICOMP	0.913	0.895	0.913	0.893	0.715	0.701
CC_{FM1}	0.913	0.893	0.913	0.893	0.719	0.702
CC_{FM2}	0.915	0.894	0.913	0.893	0.719	0.704
CC_{J1}	0.914	0.893	0.913	0.893	0.715	0.702
CC_{J2}	0.916	0.895	0.913	0.893	0.722	0.706

As seen from Table 4.32, again CC_{FM2} and CC_{J2} outperform the others overall.

Table 4.32: Monte Carlo average of specificity for each criterion

Tool	overfitted and underfitted		underfitted		overfitted	
	n=500	n=1000	n=500	n=1000	n=500	n=1000
specificity of true model	0.914	0.895	0.914	0.895	0.728	0.716
AIC	0.912	0.896	0.914	0.895	0.731	0.719
AIC_c	0.913	0.896	0.914	0.895	0.730	0.720
CAIC	0.913	0.894	0.914	0.895	0.734	0.719
BIC	0.914	0.895	0.914	0.895	0.733	0.721
ICOMP	0.913	0.896	0.914	0.895	0.731	0.721
CC_{FM1}	0.914	0.895	0.914	0.895	0.732	0.720
CC_{FM2}	0.916	0.895	0.914	0.895	0.734	0.723
CC_{J1}	0.914	0.895	0.914	0.895	0.733	0.720
CC_{J2}	0.917	0.895	0.914	0.895	0.734	0.722

4.2.2 True Model is in not the set of Candidate Models

As another part, we now use the settings in which the true model is not in the candidate set. We evaluate the performances of model selection criteria in terms of TCR, sensitivity and specificity.

4.2.2.1 Nested Models

Simulation scenarios in section 4.2.2 are used to examine the performances of model selection criteria over the set of the nested models when the true model is not in the set of candidate models. Table 4.33 gives TCR for each criterion again in three parts.

Results in Table 4.33 are similar to those in Table 4.30, 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 4.33: Monte Carlo Average of TCR for Each Criterion

Tool	overfitted and underfitted		underfitted		overfitted	
	n=500	n=1000	n=500	n=1000	n=500	n=1000
AIC	0.913	0.894	0.522	0.521	0.721	0.707
AIC_c	0.913	0.894	0.522	0.521	0.719	0.706
CAIC	0.913	0.892	0.522	0.521	0.720	0.710
BIC	0.913	0.892	0.521	0.521	0.720	0.711
ICOMP	0.913	0.894	0.522	0.521	0.721	0.706
CC_{FM1}	0.912	0.882	0.504	0.503	0.721	0.710
CC_{FM2}	0.914	0.891	0.504	0.507	0.726	0.713
CC_{J1}	0.913	0.890	0.504	0.503	0.721	0.710
CC_{J2}	0.916	0.892	0.504	0.507	0.729	0.713

Table 4.34: Monte Carlo average of sensitivity for each criterion

Tool	overfitted and underfitted		underfitted		overfitted	
	n=500	n=1000	n=500	n=1000	n=500	n=1000
AIC	0.911	0.893	0.459	0.452	0.714	0.699
AIC_c	0.911	0.893	0.459	0.452	0.713	0.699
CAIC	0.911	0.891	0.459	0.452	0.711	0.702
BIC	0.911	0.890	0.460	0.452	0.712	0.704
ICOMP	0.911	0.893	0.460	0.452	0.715	0.698
CC_{FM1}	0.910	0.880	0.474	0.474	0.712	0.703
CC_{FM2}	0.913	0.890	0.406	0.403	0.721	0.706
CC_{J1}	0.911	0.889	0.477	0.476	0.713	0.703
CC_{J2}	0.915	0.892	0.411	0.409	0.724	0.706

Sensitivity rates for this scenario are given in Table 4.34. When the true model is not in the candidate set, sensitivity rates for each criterion are less than those when the true model is involved in the candidate set. For the scenario in which the candidate set includes underfitted models, CC_{FM1} and CC_{J1} outperform CC_{FM2} and CC_{J2}, as well as the information based criteria. On the other hand, CC_{FM2} and CC_{J2} are better than all other criteria when only overfitted models are fitted as candidate models.

Table 4.35 presents the specificity rates for each criterion. CC_{FM2} and CC_{J2} are comparable with or better than all other criteria for any set of candidates. There is

again a sharp decrease in specificity rates when true model is excluded from the candidates and when only underfitted models are fitted as candidate models.

Table 4.35: Monte Carlo average of specificity for each criterion

Tool	overfitted and underfitted		underfitted		overfitted	
	n=500	n=1000	n=500	n=1000	n=500	n=1000
AIC	0.914	0.894	0.585	0.589	0.726	0.714
AIC_c	0.914	0.894	0.585	0.589	0.725	0.714
CAIC	0.914	0.892	0.586	0.589	0.728	0.717
BIC	0.914	0.891	0.583	0.589	0.728	0.717
ICOMP	0.914	0.894	0.585	0.589	0.726	0.715
CC_{FM1}	0.913	0.885	0.533	0.533	0.728	0.716
CC_{FM2}	0.915	0.892	0.602	0.610	0.731	0.719
CC_{J1}	0.914	0.891	0.532	0.530	0.728	0.716
CC_{J2}	0.917	0.892	0.597	0.605	0.733	0.719

CHAPTER 5

APPLICATION

We apply the model selection criteria to analyze the breast cancer data set obtained in a study conducted in Ankara Oncology Research and Education Hospital will be used to fit a logistic regression model. The data set includes tumor characteristics and the risk factors of 249 women with breast cancer and the covariates (breast cancer risk factors) for 251 women without breast cancer. This case-control study data set was firstly used by Dogan et al. (2011) and Erdem (2011) to investigate the etiologic heterogeneity of breast cancer, i.e. the association between the epidemiological risk factors and breast cancer by the disease characteristics. In the current chapter, a portion of the dataset is used to illustrate the model selection techniques considered in this thesis.

Data are described in detail and univariate analyses are given in section 5.1. In section 5.2, logistic regression models are fitted with the significantly effective covariates. According to the literature and univariate models in section 5.1, the set of candidate models are constructed. Model selection process is held by both using the common and newly proposed criteria.

5.1 Data Description

Data set includes a binary response for 249 cases (women with breast cancer), and 251 controls (healthy women). Covariates consist of potential risk factors. These risk factors to be investigated are age, height, weight, body mass index (BMI), region, education level, menstrual regularity, menopause status, age at menopause, age at first

menstruation, birth status, number of births, age at first birth, age at last birth, breast-feeding duration, smoking status, smoking duration, hormone replacement therapy (HRT) status, family history, mammography history, cyst history, biopsy history. Variable types for these factors will briefly be explained. Age variable consists of the ages of women with breast cancer and ages of healthy women. It is a continuous variable including positive natural numbers. BMI is known as body mass index, and it is calculated by $weight/height^2$ (kg/m^2). Region is a categorical variable with seven levels. It s based on geographical regions in Turkey. In this variable, reference level is indicated by 1, and it refers to Mediterranean Region . 2 shows Eastern Anatolia Region, 3 is Aegean Region, 4 is Southeastern Anatolia Region, 5 is Central Anatolia Region, 6 is Black Sea Region, and 7 is Marmara Region. Education level of cases and controls is presented in another categorical variable. It has 5 level, from no education as the reference level to a bachelor degree. Menstrual regularity can also be regarded as a risk factor for breast cancer. It is a categorical variable with 4 levels. The reference level is coded as 1, and it shows regularity in pre-menopausal period. 2 is for irregularity in pre-menopausal period, 3 is for perimenopausal period, and 4 stands for women in their post-menopausal period. Another categorical variable show the menopause status of cases and controls. 0 is for women who are not in their menopause period, 1 is for women who went through menopause. For women in their menopause period, their entering age exists as another continuous variable. Having had a birth or not is a binary variable coded as 1 and 0. For women who had a birth, their number of births, age at first and last births, and breast-feeding duration are other potential risk factors in the data set. Smoking status is also a binary variable coded as 1 and 0. For whom smoking, their smoking duration is given as a continuous variable. Hormone replacement therapy (HRT) status of women is given in another categorical variable. 0 is for reference level of having no HRT. 1 is for HRT with estrogen receptor (ER), 2 is for HRT with progesterone receptor (PR), and 3 is for both. Women having no family history of breast cancer is coded as 0 in another categorical variable. This variable has three levels. First order relative history is coded as 1, and second order relative history is coded as 2. Regular mammography check can also be related to the risk of breast cancer. This is given in the next binary variable. Women who never had a mammography is coded as 0, and women who go through mammography twice as year is coded as 1. Breast cyst history also exists in the data

set, and given by 1 and 0. For women who had a cyst in their breasts, their having a biopsy or not is another binary variable coded as 1 and 0.

These variables are all potential risk factors for breast cancer. For the analysis of such a data, the first thing to do is to examine their univariate relations with the response variable. For categorical variables, chi-square tests for independence are conducted. In Table 5.1 p-values are given for each. Significance level for univariate analyses can be taken as 0.15. Accordingly, region, education level, menstrual regularity, menopausal status, smoking, HRT, mammography, cyst history, and pathology variables are related to breast cancer. They should be included in the overall analysis. Comparison of the averages of continuous variables for cases and controls is held by t-test. Their p-values are given in Table 5.2. For 0.15 significance level, risk for breast cancer depends on age, BMI, number of births, age at first birth, age at last birth, breast-feeding duration, and smoking duration (marked by stars in the table). They will be included in the overall model, too.

Table 5.1: Chi-square test for independence

Factor	Factor Levels	Case	Control		p-value
				Total	
Region	Mediterranean Region	12	17		
	Eastern Anatolia Region	16	19		
	Aegean Region	9	5		
	Southeastern Anatolia Region	18	7		
	Central Anatolia Region	146	165		
	Black Sea Region	43	35		
	Marmara Region	5	3		
	Total				0.144*
Education	No education	39	15		
	Primary school	112	105		
	Secondary school	31	27		
	High school	29	53		
	University	38	46		
	Total				0.000*
Menstrual regularity	Premenaposal regural	36	115		
	Premenaposal irregular	21	37		
	Perimenaposal	43	10		
	Postmenaposal	149	89		
	Total				<0.000*
Menaposal status	0	94	162		
	1	155	89		
	Total				0.000*
Birth status	0	25	24		
	1	224	227		
	Total				0.976
Smoking status	0	189	185		
	1	60	96		
	Total				0.000*
HRT	Never	210	197		
	ER	10	14		
	PR	16	35		
	ER and PR	13	5		
	Total				0.008*
Family history	Absent	192	195		
	1^st order relative	42	46		
	2^st order relative	15	10		
	Total				0.550
Mammography	Never	159	104		
	1	26	38		
	2	64	109		
	Total				0.000*
Cyst history	0	209	157		
	1	40	64		
	Total				0.000*
Pathology	0	14	56		
	1	26	38		
	Total				0.016*

Table 5.2: t-test for the difference of means

Factor	Case	Control	
	Average (sd)	Average (sd)	p-value
Age	51.301 (10.516)	45.967 (9.553)	0.000*
BMI	29.068 (4.942)	27.183 (5.316)	0.000*
Age at menapouse	46.839 (5.424)	45.955 (5.502)	0.226
Age at first menstruation	13.486 (1.374)	13.478 (1.386)	0.949
Number of births	2.781 (1.356)	2.507 (1.311)	0.029*
Age at first birth	22.112 (4.992)	21.471 (4.349)	0.147*
Age at last birth	29.116 (5.446)	27.471 (5.220)	0.001*
Breast-feeding duration	29.237 (28.320)	24.771 (21.916)	0.062*
Smoking duration	15.183 (12.526)	12.167 (9.872)	0.117*

5.2 Analysis

Data analysis of this data set is conducted in three part. Firstly the univariate analyses are held in order to pick the significant factors. Then multivariate analyses are done to obtain the overall model. Several candidate models are fitted. In model selection part, best approximation is selected by using the information based and cluster based model selection criteria.

5.2.1 Univariate Analysis

Before conducting an overall logistic regression, it would be beneficial to see the results for univariate logistic models. All factors are fitted univariately to see their individual effects. Table 5.3 presents outputs for each model. Factors that are significant at α=0.15 nominal level are marked by a star sign. Age, BMI, region, education level, menstrual regulation, menopausal status, number of births, age at first birth, age at last birth, breast-feeding duration, smoking status, smoking duration, HRT, mammography, cyst history, and pathology variables are seem to be related to breast cancer risk. Based on univariate analyses, odds of having breast for older women is higher (OR=1.055). Women having higher BMI have higher odds of having breast

cancer, too (OR=1.075). Women living in region 3 (Southeastern Anatolia Region), have significantly higher odds than women living in region 1 (Mediterranean Region) (OR=3.643). For education level, having no education is taken as the reference. Univariate analysis of this factor reveals that higher education levels decreases the breast cancer risk. Primary school graduation makes women less likely to have a breast cancer. (OR=0.392). Women graduated from secondary school have less odds of having breast cancer than women with no education. (OR=0.441). In the same way, women graduated from high school have less odds of having breast cancer than women with no education (OR=210). A bachelor degree makes women have less odds of having breast cancer than women with no education, too (OR=0.318). Irregularities in menstrual periods also enhances the risk of breast cancer. Women having irregularity in their premenopausal period have higher odds of breast cancer than women having regular menstruations (OR=1.813). Women in their perimenopausal period have also higher odds of having breast cancer (OR=13.736). Being in postmenopausal period, women have also higher odds of having breast cancer (OR=5.348). Another significantly effective factor is the menopausal status based on the univariate analyses. Women who entered their menopause have a higher risk of breast cancer (OR=3.001). Number of births also increases the odds for having breast cancer (OR=1.171). Giving their first birth at a higher age, women have higher odds for having breast cancer (OR=1.030). In the same direction, having the last birth at a higher age makes women have higher odds of having breast cancer (OR=1.059). Breast-feeding duration is a significant factor, too. Based on this data set, women with a longer breast-feeding duration have higher odds for having breast cancer (OR=1.007). Smoking seems to decrease the odds of having breast cancer (0.512). However, for women who smoke, smoking for a longer time increases the odds of having breast cancer (1.025). HRT status show significance for two levels. Women who had HRT with PR have a less risk of having breast cancer (OR=0.429). On the other hand, women who had both ER and PR as a HRT have higher odds of having breast cancer (OR=2.439). Women who had mammography have less odds than women who never had mammography. (0.400). Having a cyst history decreases the odds of having breast cancer (OR=0.320). On the other hand, if women with a cyst history had a biopsy, their odds of having breast cancer is higher than women who never had a biopsy (2.737). The results are consistent with chi-square test for independence and t-test for the mean differences.

Table 5.3: Univariate Models

Factor	OR	95% CI	p value
Age	1.055	(1.036,1.076)	0.000*
BMI	1.075	(1.038,1.115)	0.000*
Region1	1.192	(0.442,3.261)	0.728
Region2	2.550	(0.700,10.171)	0.164
Region3	3.643	(1.196,12.013)	0.027*
Region4	1.253	(0.583,2.773)	0.566
Region5	1.740	(0.739,4.203)	0.208
Region6	2.361	(0.485,13.372)	0.296
Education1	0.392	(0.199,0.737)	0.005*
Education2	0.441	(0.197,0.961)	0.042*
Education3	0.210	(0.097,0.437)	0.000*
Education4	0.318	(0.149,0.652)	0.002*
Menstrual reg1	1.813	(0.936,3.476)	0.074*
Menstrual reg2	13.736	(6.501,31.522)	0.000*
Menstrual reg3	5.348	(3.414,8.533)	0.000*
Menopause	3.001	(2.091,4.333)	0.000*
Menopause age	1.030	(0.982,1.082)	0.224
Menstruation age	1.004	(0.884,1.141)	0.949
Birth	0.947	(0.523,1.712)	0.857
Number of births	1.171	(1.017,1.356)	0.031*
Age at the first birth	1.030	(0.990,1.072)	0.148*
Age at the last birth	1.059	(1.023,1.098)	0.001*
Breast-feeding duration	1.007	(0.999,1.016)	0.068*
Smoking	0.512	(0.347,0.752)	0.000*
Smoking duration	1.025	(0.995,1.056)	0.100*
HRT1	0.670	(0.283,1.295)	0.347
HRT2	0.429	(0.225,0.787)	0.008*
HRT3	2.439	(0.902,7.716)	0.096*
Family history1	0.927	(0.582,1.474)	0.750
Family history2	1.523	(0.675,3.583)	0.317
Mammography	0.400	(0.278,0.573)	0.000*
Cyst history	0.320	(0.207,0.485)	0.000*
Pathology	2.737	(1.284,6.031)	0.010*

5.2.2 Multivariate Analysis

Significantly effective covariates are chosen to fit an overall model given in Table 5.4. The probability of having a breast cancer is modelled by these covariates adjusted for each other.

From Table 5.3, it is seen that region variable show significance for only one level. Therefore, it is changed to a binary variable. If a woman is from region 3 (Southeastern Anatolia Region), it is coded as 1, otherwise it is 0. This variable is put into the overall model as a two-level categorical variable.

Table 5.4: Overall Model

Factor	OR	95 % CI	p value
Age	1.033	(1.002,1.066)	0.038*
BMI	1.027	(0.982,1.076)	0.239
Region	2.789	(0.989,8.503)	0.059
Education1	0.596	(0.262,1.306)	0.205
Education2	0.924	(0.334,2.507)	0.877
Education3	0.326	(0.117,0.869)	0.028*
Education4	0.793	(0.283,2.168)	0.655
Menstrual reg1	2.594	(1.197,5.609)	0.015*
Menstrual reg2	13.172	(5.629,33.465)	0.000*
Menstrual reg3	6.103	(3.822,9.141)	0.000*
Menopause	1.021	(0.541,3.252)	0.977
Number of births	0.711	(0.543,0.926)	0.012*
Age at the first birth	0.971	(0.911,1.033)	0.355
Age at the last birth	1.043	(0.983,1.108)	0.162
Breast-feeding duration	1.006	(0.995,1.017)	0.271
Smoking	0.723	(0.362,1.430)	0.354
Smoking duration	1.016	(0.980,1.055)	0.403
HRT1	0.462	(0.172,1.195)	0.115
HRT2	0.628	(0.280,1.373)	0.250
HRT3	4.839	(1.486,18.003)	0.012*
Mammography	0.287	(0.169,0.480)	0.000*
Cyst	0.278	(0.125,0.581)	0.001*
Pathology	3.610	(1.448,9.409)	0.007*

In Table 5.4, significantly effective factors are marked by a star sign for the overall model. The significance level is taken as 0.05.

5.2.3 Model Selection

In order to obtain the first candidate model, insignificant covariates are removed from the overall model, and the model is fitted again. Age, BMI, and smoking status variables are kept in the model. The other variables are adjusted according to these factors. By eliminating the insignificant factors, the final significant model adjusted for age, BMI, and smoking status. As a result, the first candidate model is given by

$$\begin{aligned} M1 : logit(P(Y=1))_i = \beta_0 &+ \beta_1 AGE_i + \beta_2 BMI_i + \beta_3 MenstrualReg_i \\ &+ \beta_4 Smoking_i + \beta_5 HRT_i + \beta_6 Mammography_i \\ &+ \beta_7 Cyst_i + \beta_8 Pathology_i \end{aligned}$$

Table 5.5 presents outputs for M1.

Table 5.5: Candidate Model 1

Factor	OR	95 % CI	p value
Age	1.026	(0.998,1.055)	0.076
BMI	1.030	(0.987,1.074)	0.176
Menstrual reg1	2.647	(1.275,5.497)	0.009*
Menstrual reg2	14.153	(6.168,35.299)	0.000*
Menstrual reg3	5.229	(2.777,10.079)	0.000*
Smoking	0.820	(0.518,1.298)	0.395
HRT1	0.439	(0.169,1.099)	0.081
HRT2	0.515	(0.237,1.085)	0.086
HRT3	4.564	(1.474,16.264)	0.012*
Mammography	0.316	(0.191,0.516)	0.000*
Cyst	0.331	(0.156,0.667)	0.003*
Pathology	3.146	(1.337,7.789)	0.010*

Nonlinearity of the relation between age and breast cancer probability is suspected.

Therefore, along with the model shown in Table 5.5, another candidate model including a quadratic age term is conducted as given by

$$M2 : logit(P(Y=1))_i = \beta_0 + \beta_1 AGE_i + \beta_2 AGE_i^2 + \beta_3 BMI_i + \beta_4 MenstrualReg_i + \beta_5 Smoking_i + \beta_6 HRT_i + \beta_7 Mammography_i + \beta_8 Cyst_i + \beta_9 Pathology_i$$

Outputs for M2 is given in Table 5.6.

Table 5.6: Candidate Model 2

Factor	OR	95 % CI	p value
Age	1.025	(0.888,1.184)	0.730
Age^2	1.000	(0.999,1.001)	0.998
BMI	1.030	(0.986,1.075)	0.183
Menstrual reg1	2.648	(1.274,5.501)	0.009*
Menstrual reg2	14.154	(6.148,35.420)	0.000*
Menstrual reg3	5.229	(2.771,10.108)	0.000*
Smoking	0.820	(0.518,1.299)	0.396
HRT1	0.439	(0.169,1.099)	0.082
HRT2	0.515	(0.237,1.085)	0.086
HRT3	4.563	(1.470,16.334)	0.012*
Mammography	0.316	(0.190,0.519)	0.000*
Cyst	0.331	(0.155,0.668)	0.003*
Pathology	3.146	(1.337,7.680)	0.010*

In another candidate model, nonlinearity of BMI is checked by adding a quadratic BMI term. Now, the age variable is kept in a linear relation. The fitted model is given by

$$M3 : logit(P(Y=1))_i = \beta_0 + \beta_1 AGE_i + \beta_2 BMI_i + \beta_3 BMI_i^2 + \beta_4 MenstrualReg_i + \beta_5 Smoking_i + \beta_6 HRT_i + \beta_7 Mammography_i + \beta_8 Cyst_i + \beta_9 Pathology_i$$

Table 5.7 indicates outputs for M3.

Table 5.7: Candidate Model 3

Factor	OR	95 % CI	p value
Age	1.019	(0.991,1.049)	0.197
BMI	1.599	(1.152,2.245)	0.005*
BMI^2	0.993	(0.987,0.998)	0.008*
Menstrual reg1	2.547	(1.215,5.336)	0.013*
Menstrual reg2	13.945	(6.050,34.899)	0.000*
Menstrual reg3	5.554	(2.929,10.796)	0.000*
Smoking	0.814	(0.513,1.292)	0.382
HRT1	0.430	(0.165,1.086)	0.076
HRT2	0.492	(0.225,1.043)	0.069
HRT3	4.877	(1.543,17.767)	0.010*
Mammography	0.307	(0.185,0.502)	0.000*
Cyst	0.325	(0.152,0.660)	0.003*
Pathology	3.155	(1.334,7.748)	0.010*

The last candidate model is given in Table 5.8. Both quadratic age term, and quadratic BMI term is now included in the fitted model.

Table 5.8: Candidate Model 4

Factor	OR	95 % CI	p value
Age	1.000	(0.863,1.159)	0.997
Age^2	1.000	(0.999,1.002)	0.796
BMI	1.606	(1.155,2.259)	0.005*
BMI^2	0.993	(0.987,0.998)	0.008*
Menstrual reg1	2.559	(1.220,5.364)	0.012*
Menstrual reg2	14.096	(6.089,35.442)	0.000*
Menstrual reg3	5.603	(2.944,10.940)	0.000*
Smoking	0.816	(0.514,1.296)	0.389
HRT1	0.429	(0.165,1.084)	0.076
HRT2	0.492	(0.225,1.043)	0.069
HRT3	4.814	(1.521,17.497)	0.010*
Mammography	0.309	(0.185,0.509)	0.000*
Cyst	0.327	(0.152,0.665)	0.003*
Pathology	3.148	(1.330,7.731)	0.010*

The formula of the model is given by

$$\begin{aligned} M4 : logit(P(Y=1))_i = \beta_0 &+ \beta_1 AGE_i + \beta_2 AGE_i^2 + \beta_3 BMI_i + \beta_4 BMI_i^2 \\ &+ \beta_5 MenstrualReg_i + \beta_6 Smoking_i + \beta_7 HRT_i \\ &+ \beta_8 Mammography_i + \beta_9 Cyst_i + \beta_{10} Pathology_i \end{aligned}$$

Overall significance of the candidate models are tested by using Hosmer-Lemeshow test (Hosmer and Lemeshow, 2000) and the p-values are given in Table 5.9. Nominal significance level is set at α=0.05. As a result, the conducted models are all significant.

Candidate models M1-M4 are compared using the model selection methods considered in this thesis. Results are presented in Table 5.9.

Table 5.9: Comparison

Criterion	Overall	M1	M2	M3	M4
AIC	560.424	562.091	564.091	**557.012**	558.945
AICc	562.951	562.840	564.957	**557.878**	559.937
CAIC	610.999	589.486	593.594	**586.515**	590.555
BIC	661.575	616.881	623.096	**616.017**	622.165
ICOMP	829.211	**555.563**	577.855	574.955	591.178
CC_{FM1}	1.437	**0.959**	1.021	1.009	1.081
CC_{FM2}	0.691	**0.564**	0.582	0.570	0.599
CC_{J1}	1.608	**1.130**	1.192	1.180	1.253
CC_{J2}	0.862	**0.735**	0.753	0.741	0.771
Hosmer-Lemeshow p-values	0.380	0.824	0.824	0.477	0.595
TCR	0.73	0.73	0.73	0.73	0.23
Sensitivity	0.59	0.72	0.72	0.66	0.14
Specificity	0.87	0.73	0.73	0.81	0.45

Overall model is also presented in Table 5.9. According to Hosmer-Lemeshow test, this model is significant. However, it cannot be a candidate model, since it includes insignificant factors. Model selection criteria also regard this model as the worst model. Proposed criteria are able to discriminate between poor and better approximations.

In Table 5.9, the minimum value of each criterion is given in bold for each candidate model. AIC, AIC_c, CAIC and BIC select M3, where M3 is the model including linear relation with age, and nonlinear relation with BMI index. On the other hand, ICOMP and the clustering based criteria choose M1, where M1 is linear in age and BMI.

M1 versus M3 ?

When model selection criteria used for a particular data analysis lead to different models, care should be taken and more investigation should be conducted before eventually deciding on the final model. These include i. taking the *statistical power* statistical power of the model selection criteria used into account, ii. existent literature on the subject matter.

As seen in Table 4.1, when compared with other common criteria, ICOMP is more successful in determining the need for a quadratic term. That is, if there was a curvilinear relationship between the logit(P(Y=1)) and age or BMI, ICOMP would have selected the logistic regression model with quadtratic terms in age and BMI. For this data analysis, the proposed criteria choose the same model with ICOMP. Hosmer-Lemeshow significance test p values also favor the candidate Model 1 over candidate Model 3, since it has a greater p-value.

In Table 5.9, TCR, sensitivity and specificity for each candidate model are also given. There is no significant difference between M1 and M3 in terms of TCR. Sensitivity rate of M3 is remarkably low, whereas specificity rate for the same model is higher than M1.

Linear relation between age and breast cancer is supported by an earlier study (Kessler, 1992). Figure 5.1 shows that a sample taken in 1987 shows that breast cancer incidence increases until age 75, and then starts to decrease. This is actually a sign for a nonlinear relation between age and breast cancer. However, most of the women included in our data set are under age 75. There are only 5 women older than 75 out of 500 women. Therefore, for our sample we expect to observe a linear relation between age and breast cancer based on Figure 5.1.

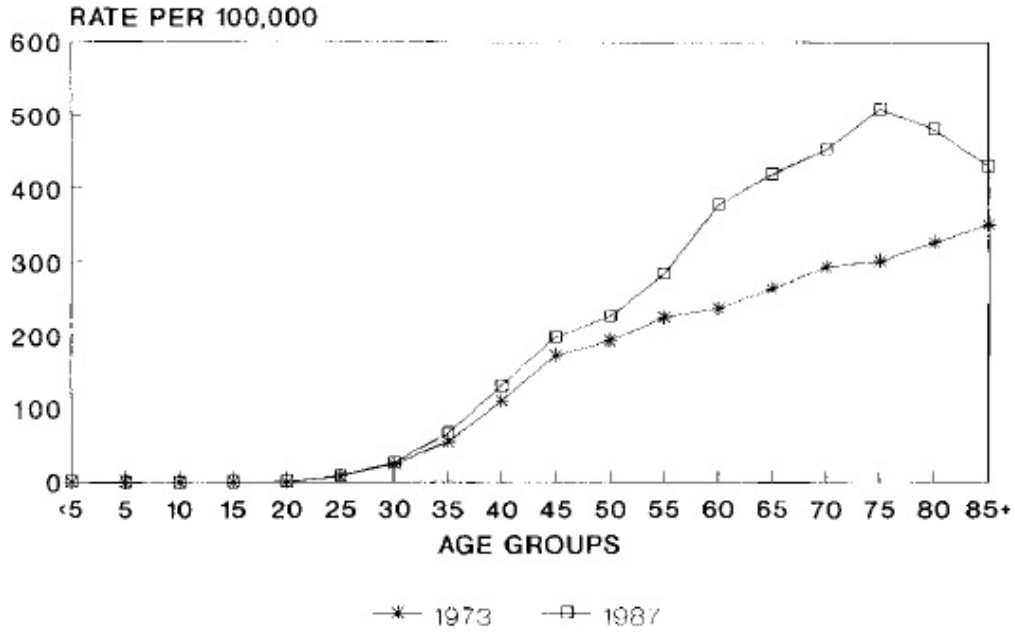

Figure 5.1: Breast Cancer Incidence by Age

Figure 5.2 presents the relation between age and proportion of cases in the data, and Figure 5.3 shows the relation between BMI and proportion of cases in the data. These are also an indicator of linear relations.

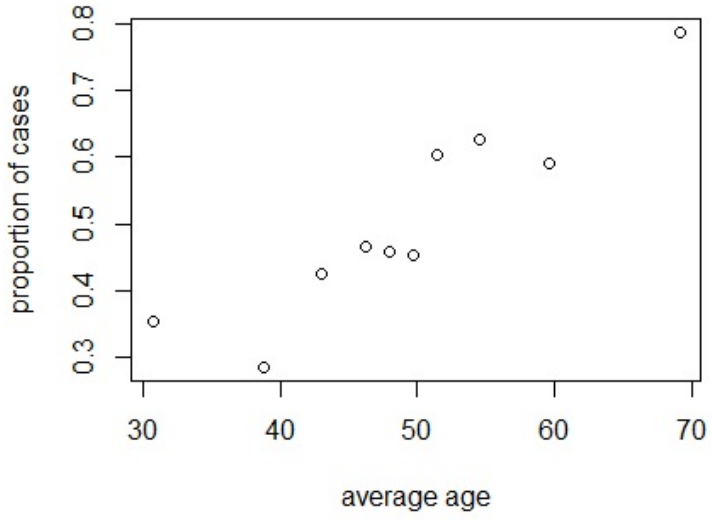

Figure 5.2: Age vs. P(Y=1)

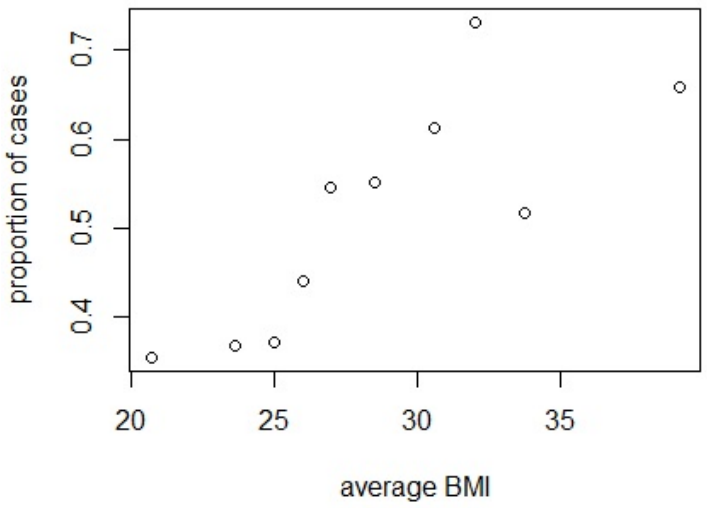

Figure 5.3: BMI vs. P(Y=1)

As a result of these additional findings, M1 is preferred over M3.

CHAPTER 6

CONCLUSION

In this study, we focus on the lack of importance given to modeling purposes in model selection process. Our objective in this thesis is to provide a new model selection approach for logistic regression based on classification. Logistic regression is commonly used for classification purposes. With this incentive, we use clustering tree distances to assess adequacy of logistic regression models.

In the first part of this thesis, likelihood based model selection criteria, namely AIC, AIC_c, CAIC, BIC and ICOMP, are reviewed. This part also provides information about different characteristics of model selection criteria for different purposes. Consistency and efficiency definitions are given for the cases in which the modeling purpose is model fitting. For the cases when classification is the modeling purpose, TCR, sensitivity and specificity measures are explained.

In the next part, new approach for model selection is given in detail. We here focus on the idea that logistic regression can be used as a classification tool. When observed and predicted values of a logistic fit are presented by two cluster trees, the similarity of these trees is used as a goodness of the model as a classification tool. Existing clustering similarity measures and the reasons for choosing FM and Jaccard among them are explained. Their behaviour as a model selection tool are assessed by conducting small simulation studies. The need for a penalty term is shown. Theory behind the existing penalty terms are investigated. The behaviour of likelihood based model selection criteria for the same simulation scenario is also assessed. In order to obtain such behaviour, 1-FM and 1-Jaccard are penalized for the number of parameters. The proposed penalties are also based on the sample size. These new penalized cluster

based criteria are denoted by CC_{FM} and CC_J.

The performances of CC_{FM} and CC_J with two different possible penalty terms are evaluated in simulation studies chapter. They are also compared with commonly used information based criteria, AIC, AIC_c, CAIC, BIC and ICOMP. Simulation scenarios are divided into two based on modeling purposes of model fitting and classification. Their performances are also assessed both when the true model is in the set of candidate models and when it is not. Results present that there are no outstanding differences in performances for different modeling purposes. Cluster based criteria are not better than information based criteria when the modeling purpose is classification. They show similar performances for both purposes. The most remarkable results of these simulations are given in two parts. For the scenarios in which the modeling purpose is model fitting, the results are as follows:

- When the candidate set includes rather parsimonious models,
  - In terms of consistency, ICOMP outperform all other criteria. Among information based criteria, AIC and AIC_c are better than CAIC and BIC. Among cluster based criteria, CC_{FM2} and CC_{J2} are better than CC_{FM1} and CC_{J1}. Information based criteria perform better than cluster based criteria.
  - In terms of efficiency, ICOMP, AIC and AIC_c are better than all other. Two groups of criteria are comparable to each other.
- When the candidate set includes rather saturated models,
  - In terms of consistency, among information based criteria CAIC and BIC perform better than AIC, AIC_c and ICOMP. Among cluster based criteria, CC_{FM1} and CC_{J1} are better than CC_{FM2} and CC_{J2}. Cluster based criteria perform better than cluster based criteria.
  - In terms of efficiency, ICOMP, AIC and AIC_c are again better than all other. Two groups of criteria are comparable to each other.

When the modeling purpose is classification, our simulations result in the following:

- All criteria show similar performances.

- CC_{FM2} and CC_{J2} are better than others especially for detecting a missing interaction term when sample size is moderate and for nested models when the candidate set involves overfitted models.

We also use cluster based criteria along with the information based criteria to model a breast cancer data set. Candidate model set is constructed after univariate analyses. Models are selected by comparing the values of CC_{FM}, CC_J, AIC, AIC_c, CAIC, BIC and ICOMP. Cluster based criteria selects the same model with ICOMP, whereas the others selects an another model as the best approximation. The difference between these two models is a quadratic term. In our simulations, ICOMP performs better than any other criteria in detecting a missing quadratic term. Therefore, our model choice is reasonable.

This thesis leads a different point of view for model selection of logistic regression. It also provides an extensive comparison of the existing model selection criteria. However, our studies imply that proposed penalty terms require a further research. As a future study, we consider to investigate those terms in more detail.

REFERENCES

[1] Aparicio, T. and Villanua, I. (2007). *Some Selection Criteria for Nested Binary Choice Model: A Comparative Study*. Computational Statistics, 22, 635-660.

[2] Behboodian, J. (1964). *Information for Estimating the Parameters in Mixtures of Exponential and Normal Distributions*. Unpublished doctoral dissertation, Department of Mathematics, University of Michigan, Ann Arbor.

[3] Baudry, J. P., Cardoso, M., Celeux, G., Amorim, M. J. and Ferreira, A. S. (2015). *Enhancing the selection of a model-based clustering with external categorical variables*. Advanced Data Analysis and Classification, 9, 177-196.

[4] Biernacki, C., Celeux, G. and Govaert, G. (2000). *Assessing a Mixture Model for Clustering with the Integrated Completed Likelihood*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(7).

[5] Bouchard, G. and Celeux, G. (2006). *Selection of Generative Models in Classification*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 28(4).

[6] Box, G. E. P (1976). *Science and Statistics*. Journal of American Statistical Association, 71(356), 791-799.

[7] Bozdogan, H. (1987). *Model Selection and Akaike's Information Criterion (AIC): The General Theory and its Analytical Extensions*. Psychometrika, 52(3), 345-370.

[8] Bozdogan, H. (1988). *ICOMP: A New Model Selection Criterion.* H.H. Bock (Ed.), Classification and Related Methods of Data Analysis, North-Holland, Amsterdam, 599-608.

[9] Bozdogan, H. (2000). *Akaike's Information Criterion and Recent Developments in Information Complexity*. Journal of Mathematical Psychology, 44, 62-91.

[10] Bozdogan, H. and Haughton, D. M. A. (1998). *Informational Complexity Criteria for Regression Models*. Computational Statistics & Data Analysis, 28, 51-76.

[11] Cavanaugh, J. E. (2004). *Criteria For Linear Model Selection Based on Kullbakc's Symmetric Divergence*. Australian New Zealand Journal of Statistics, 46(2), 257-274.

[12] Claeskens, G., Croux, C. and Kerckhoven, J. V. (2006). *Variable Selection for Logistic Regression Using a Prediction-Focused Information Criterion*. Biometrics, 62, 972-979.

[13] Claeskens, G. and Hjort, N. L. (2003). *The Focused Information Criterion (with Discussion)*. Journal of the American Statistical Association, 98, 900-916.

[14] Claeskens, G. and Hjort, N. L. (2008). *Model Selection and Model Averaging*. Cambridge: Cambridge University Press.

[15] Dogan, L., Kalaylioglu, Z., Karaman, N., Ozaslan, C., Atalay, C. and Altinok, M. (2011). *Relationships Between Epidemiological Features and Tumor Characteristics of Breast Cancer*. Asian Pacific Journal of Cancer Prevention, 12, 3375-3380.

[16] Downton, M. and Brennan, T. (1980). *Comparing Classifications: An Evaluation of Several Coefficients of Partition Agreement*. Paper presented at the meeting of the Classification Society, Boulder, Colorado.

[17] Erdem, M. T. (2011). *Modeling Diseases with Multiple Disease Characteristics: Comparison of Models and Estimation Methods*. Unpublished Master's thesis, Middle East Technical University, Ankara, Turkey.

[18] Fitzmaurice, G. M., Laird, N. M. and Ware, J. H. (2004). *Applied Longitudinal Analysis*. New York: John Wiley & Sons Inc.

[19] Fowlkes, E. B. and Mallows, C. L. (1983). *A Method for Comparing Two Hierarchical Clusterings*. Journal of the American Statistical Association, 78(383), 553-569.

[20] Hosmer, D. W., Hosmer, T., Le Cessie, S. and Lemeshow, S. (1997). *A Comparison of Goodness-of-Fit Tests for the Logistic Regression Model*. Statistics in Medicine, 16, 965-980.

[21] Hosmer, D. W. and Lemeshow, S. (2000). *Applied Logistic Regression*.(2nd Edition). New York: John Wiley & Sons Inc.

[22] Hubert, L. and Arabie, P. (1985). *Comparing Partitions*. Journal of Classification, 2, 913-218.

[23] Hurvich, C. M. and Tsai C. L. (1989). *Regression and Time Series Model Selection in Small Samples*. Biometrika, 76(2), 297-307.

[24] Kalaylioglu, Z. I. and Ozturk, O. (2013). *Bayesian Semiparametric Models for Nonignorable Missing Mechanisms in Generalized Linear Models*. Journal of Applied Statistics, 40(8), 1746-1763.

[25] Kessler, L. G. (1992). *The Relationship Between Age and Incidence of Breast Cancer*. Cancer Supplement, 69(7).

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

[27] Lee, J. W., Lee, J. B., Park, M. and Song, S. H. (2005). *An Extensive Comparison of Recent Classification Tools Applied to Microarray Data*. Computational Statistics & Data Analysis, 48, 869-885.

[28] Milligan, G. W. and Schilling, D. A. (1985). *Asymptotic and Finite Sample Characteristics of Four External Criterion Measures*. Multivariate Behavioral Research, 20, 97-109.

[29] Morey, L. and Agresti, A. (1984). *The measurement of classification agreement: An adjustment to the Rand statistic for chance agreement.* Educational and Psychological Measurement, 44, 33-37.

[30] Muller, S. and Welsh, A. H. (2010). *On Model Selection Curves*. International Statistical Review, 78, 240-256.

[31] Nelder, J. A. and Wedderburn R. W. M. (1972). *Generalized Linear Models*. Journal of the Royal Statistics Society. Series A (General), 135(3), 370-384.

[32] Rand, W. M. (1971). *Objective criteria for the evaluation of clustering methods.* Journal of the American Statistical Association, 66, 846-850.

[33] Rao, C. R. and Wu, Y. (2001). *On Model Selection*. IMS Lecture Notes, Monograph Series, 38.

[34] Rawlings, J. O., Pantula, S. G. and Dickey, D. A. (1998). *Applied Regression Analysis: A Research Tool*. New York: Springer-Verlag.

[35] Qian, G. and Field, C. (2002). *Law of Iterated Logarithm and Consistent Model Selection Criterion in Logistic Regression*. Statistics & Probability Letters, 56, 101-112.

[36] Schwarz, G. (1978). *Estimating the Dimension of a Model*. The Annals of Statistics, 6(2), 461-464.

[37] Seghouane, A. and Amari, S. (2007). *The AIC Criterion and Symmetrizing the Kullback-Liebler Divergence*. IEEE Transactions on Neural Networks, 18(1).

[38] Seghouane, A. (2010). *Asymptotic Bootstrap Corrections of AIC for Linear Regression Models*. Signal Processing, 90, 217-224.

[39] Steinley, D. (2004). *Properties of the Hubert-Arabie Adjusted Rand Index*. Psychological Methods, 9(3), 386-396.

APPENDIX A

A.1 Axiom for "Simplest correct polynomial has the smallest KL divergence from the true nonlinear model"

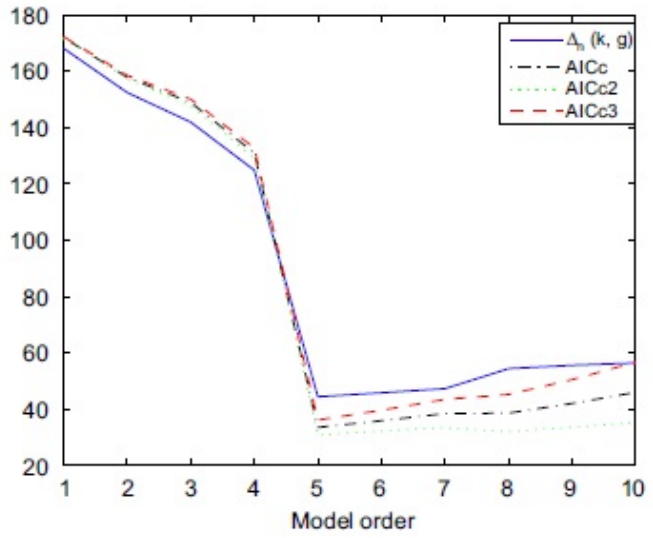

Figure A.1: Model selection criteria vs. model order

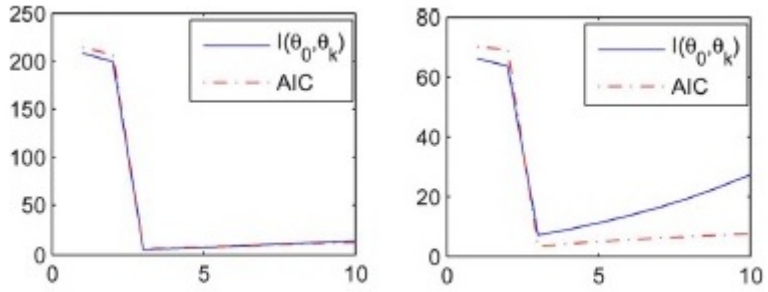

Figure A.2: Model selection criteria vs. model order

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 is called full model.

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 nonlinear 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 A.1 (Seghouane, 2010) and A.2 (Seghouane and Amari, 2007). Let k_{min} be the solution of $\frac{d}{dk}KL = 0$. 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.

A.2 Likelihood of Clustering Similarity Measures

Let $g(\beta) = \frac{e^{x^T\beta}}{1+e^{x^T\beta}}$. Also let $g(\hat{\beta}(\alpha)) = \frac{e^{x_\alpha^T\hat{\beta}(\alpha)}}{1+e^{x_\alpha^T\hat{\beta}(\alpha)}}$ denote the fitted model and, $g(\beta_0(\alpha)) = \frac{e^{x_\alpha^T\beta_0(\alpha)}}{1+e^{x_\alpha^T\beta_0(\alpha)}}$ denote the true model.

$$L(\pi_{00},\pi_{01},\pi_{10},\pi_{11}) \propto \prod_{i=1}^{n}\prod_{j=i+1}^{n}[\pi_{00,ij}^{(1-y_i)(1-y_j)}\pi_{01,ij}^{(1-y_i)(y_j)}\pi_{10,ij}^{(y_i)(1-y_j)}(1-\pi_{00,ij}-\pi_{01,ij}-\pi_{10,ij})^{y_iy_j}]$$

where

$$\begin{aligned}
\pi_{00,ij} = P_f(Y_i=0,Y_j=1)P_g(Y_i=0,Y_j=1) &+ P_f(Y_i=0,Y_j=1)P_g(Y_i=1,Y_j=0)+\\
P_f(Y_i=1,Y_j=0)P_g(Y_i=0,Y_j=1) &+ P_f(Y_i=1,Y_j=0)P_g(Y_i=0,Y_j=1)\\
\pi_{01,ij} = P_f(Y_i=0,Y_j=1)P_g(Y_i=0,Y_j=0) &+ P_f(Y_i=1,Y_j=0)P_g(Y_i=0,Y_j=0)+\\
P_f(Y_i=0,Y_j=1)P_g(Y_i=1,Y_j=1) &+ P_f(Y_i=1,Y_j=0)P_g(Y_i=1,Y_j=1)\\
\pi_{10,ij} = P_f(Y_i=0,Y_j=0)P_g(Y_i=0,Y_j=1) &+ P_f(Y_i=0,Y_j=0)P_g(Y_i=1,Y_j=0)+\\
P_f(Y_i=1,Y_j=1)P_g(Y_i=0,Y_j=1) &+ P_f(Y_i=1,Y_j=1)P_g(Y_i=1,Y_j=0)\\
\pi_{11,ij} = P_f(Y_i=0,Y_j=0)P_g(Y_i=0,Y_j=0) &+ P_f(Y_i=1,Y_j=1)P_g(Y_i=1,Y_j=1)+\\
(P_f(Y_i=0,Y_j=0)P_g(Y_i=1,Y_j=1) &+ P_f(Y_i=1,Y_j=1)P_g(Y_i=0,Y_j=0)
\end{aligned}$$

where g is the true density function and f is the predicted density function. (y_i, y_j) denotes the all possible pairs which are classified. $\pi_{00,ij}$ is the probability of having a pair in the same cluster by both of trees. $\pi_{01,ij}$ is the probability of having a pair put in different clusters by the first tree and put in the same cluster by the second tree. $\pi_{10,ij}$ is the probability of having a pair put in the same cluster by the first tree and put in different clusters by the second tree. $\pi_{11,ij}$ is the probability of having a pair in different clusters by both of trees.

$$\pi_{00,ij} = (1 - g_i(\hat{\beta}(\alpha)))(1 - g_j(\hat{\beta}(\alpha)))(1 - g_i(\beta_0))(1 - g_j(\beta_0)) + g_i(\hat{\beta}(\alpha))g_j(\hat{\beta}(\alpha))g_i(\beta_0)g_j(\beta_0)$$

where β_0 is the true coefficients of full (unknown) model $X^T\beta$; $\hat{\beta}(\alpha)$ is the estimated coefficients of the submodel $x_\alpha^T\beta$; $g(.) = \frac{e^\cdot}{1+e^\cdot}$. Others follow similarly.