Razi University

Faculty of Science

Department of Statistics

M.Sc. Thesis

Model Selection Based on Akaike Information Criterion

Supervisor :

Dr. Abdolreza Sayyareh

Advisor:

Dr. Davood Ghazvininejad

By:

Raouf Obeidi

May 2009


بسم الله الرحمن الرحيم

I am indebted to a lot of people, and I am sure that there are some that I will forget. If you are one of those, I am so sorry. Anyway, thank you for whatever you did for me. Those that I do have remembered to thank are;

First of all my dear supervisor, *Dr. Abdolreza Sayyareh* for his supervision, his encouragement, constructive guidance, excellent suggestion and critical evaluation at various key stages of the task.

I also would like to thank honorable members of the thesis committee, *Dr. Khosro Fazli*, the external examiner, *Dr. Reza Hashemi*, the internal examiner, and finally *Dr. Davood Ghazvininejad*, my advisor for his excellent suggestion.

Deep acknowledgement is offered to *Mrs. Leila Korani* the secretary of the Statistics department.

My deep acknowledgement is offered to *Asadollah Faghihi*, for all her kind friendship and honestly help during my work.

My special thank to *Romina Yachkaschi* for helping me to edit my chapter 5.

I would like to express my thanks to my best friends, *Abbas Tavasoli*, *Molood Abdollahi*, *Fereshteh Kahrari* and *Nasrin Hami Golzar* for all their kind friendship and accompaniment in various stages of this study.

I am also to indebted to my very kind friends who have been with me always and everywhere to all M. Sc. students in Biology, Chemistry, Mathematics and Statistics departments in Razi University for all their kind friendship.

I also would like to thank to *Dr. Hasan Doolah* and *Razieh Obeidi* for hearting me in all stages of my studies.

And finally, I would like to express my thanks to my family specially my mother and father for encouraging me to study and research in my field.

To

My Dear Parents

Abstract

Statistical modeling is an important part in statistics. Cox test is a solution of this problem which is modified likelihood ratio test for two non-nested models. It is an absolute hypothesis testing. Another method is Vuong test which considers equivalence of two competing models close to the data generating model. It is a relative hypothesis testing. Cox and Vuong tests are two model selection tests. On the other hand, AIC is a model selection criterion which is free of type-I error. These three methods are based on Kullback-Leibler divergence.

When the model is selected as the better model than the other ones with Vuong test or as the best model with AIC, it is not specified that how it is close to the data generating model. It is proposed to use the result of an absolute hypothesis testing, like Cox test, with Vuong test and AIC to answer to this question.

In this work, these three approaches which are essentially based on the likelihood function are studied. The asymptotic properties of them are verified. Finally, they are compared with each other.

Contents

List of Figures

List of Tables

Chapter 1

Some Definition

1.1 Introduction

An important problem in statistic concerning a sample of n independent and identically distributed observations is to test whether these observations come from a specified distribution. In such a uncertainty situation the statistical process based on data will construct statistical models for decision making. The analysis of models has followed two approaches in the literature; the hypothesis testing and the model selection criteria. Sometime a model is chosen which is at least not falsified. Clearly this approach is different of classical hypothesis testing. Two models may be nested or non-nested, and in the latter case they may be overlap. The nested models are frequently studied in both theoretical and applied statistics. But the non-nested models are less discussed. Historically a serious studies on non-nested models can be found in a period from Cox (1961,1962) to Vuong (1989). In search of similarities and differences between Cox test and Vuong test one may say that the Vuong test is a development of the Cox test. Both tests are a generalization of the likelihood ratio test under different sense. In Cox test the difference between the log-likelihood ratio and its expected value under the null hypothesis is considered. The Cox test says that a true model must be able to predict the performance of the specific alternatives, i.e. a true null should not distort the actual performance of the alternative model. The idea is to compare the true performance of the alternative model with the expected performance of the alternative model under the null hypothesis. The important points is that when a hypothesis is rejected, there is no means that it is rejected in favor of the specific alternative. For example the rejection of both models implies that neither model could predict the results of the other model. Then one concludes that both models are mis-specified. May be a solution to this difficulty is to use a model selection approach which chooses the model which is closest to the true model. Also, the other difficulty with Cox test is calculating the expected value of the log-likelihood ratio under the null hypothesis. Another candidate in a similar situation is Vuong test. In Vuong viewpoint, the best model is the model which maximizes the relevant part of Kullback-Leibler (Kullback

and Leibler, 1951), divergence. The null hypothesis of Vuong test is the expectation under the true model of the log-likelihood ratio of the two candidate models which are equal to zero. It means that two candidate models are equivalent. This expectation however is unknown, but Vuong test works, because the decision making procedure by Vuong test does not depend on this unknown quantity.

On the other hand, some recent methods in model selection criteria are Akaike information Criterion (AIC), (Akaike, 1973), as a kind of penalized likelihood which its small value (notice that its small value has no intrinsic meaning) indicates the better model. The Bayesian information criterion (BIC), (Schwarz, 1978), Cross Validation technique, which is asymptotically equivalent to the AIC in the parametric case, bootstrap information criterion (EIC), (Ishiguro et al., 1997), ICOMP (Bozdogan, 2000), which is asymptotically equivalent to the BIC, are another model selection criteria. Notice that the conclusion of the criteria like AIC are never about the truth or falsity of a hypothesis, but about its closeness to the data generating model.

On the other hand, it seems that the rational behind the classical hypothesis testing is minimization of the type-I error and the type-II error which are incompatible. But the actual practice is a trade off between these two errors. There is another objection to the rationale of classical hypothesis testing. It may be difficult to find a correct specified model. It may still be relevant to choose the best model among a set of (misspecified) models. The idea of model selection is begin with a set of competing models to choice the best one. The decision making on this set is a important question in statistical inference. The Cox test, Vuong test and the AIC are designed to answer to this question that which of the competing models is the best one, or at least, which of them are equivalent to select as the bests. The literature on non-nested hypothesis testing in statistics was pioneered by Cox (1961, 1962) and Atkinson (1970), this subject applied by Pesaran (1974) and Pesaran and Deaton (1978). The analysis of non-nested regression models considered by Davidson and MacKinnon (1981), Fisher and McAleer (1981) and Dastoor (1983). Vuong (1989) considered the hypothesis testing when two competing models are nested, overlap and non-nested. His approach is based on the asymptotic distribution of difference of log-likelihood functions for two competing models. Recently the asymptotic distribution of AIC in linear regression models and the bias correction of this statistics are discussed by Yanagihara and Ohomoto (2005). Recently Commenges et al. (2008) has considered the normalized difference of

AIC as an estimate of a difference of Kullback-Leibler risks between two models. The comparison between the three essential approaches, Cox, Vuong and AIC, is of our interest. Genius and Strazzera (2002) have considered the behavior of three methods for regression models with discrete response. In this direction a question arises, What is the interpretation of equivalence of two rival models in Vuong approach? In fact we want to pay to this question, when two rival models are equivalent in Vuong test, they are two equivalent models which are close to unknown true model or far from it?. What is the result of Cox test and AIC in this case? In this work we want to answer to some kind of these questions by simulation.

The structure of this work is as follow. Chapter 1 gives some definitions which are required. Chapter 2 contains likelihood function. Nested and non-nested hypothesis testing are presented in chapter 3. Chapter 4 describes model selection based on Akaike information criterion. Finally, comparison between criteria is presented in chapter 5.

1.2 Statistical Model

Statistical models are needed when the random phenomena under study is not completely predictable. More precisely, in almost all empirical studies, there are uncontrollable elements of variability. For example if one observes the average hourly wind velocity at a given meteorological station, the variable under consideration is a random variable. It can not be generally predictable with certainty what value will be assumed in the next hour of observation. A complete characterization of the random variable being observed is given only if it can specified exactly with a statistical model.

The Oxford dictionary defines a model as a simplified or idealized description of a particular system, situation or process, often in mathematical terms that is put forth as a basis for a theoretical or empirical understanding, or for calculations, predictions, etc. Thus, a good definition of the *statistical model* is, a simplified or idealized description of a random phenomenon, generally in probabilistic terms, that is put forth as a basis for a theoretical or empirical understanding, or for conclusions, inferences, predictions, etc. All of the statistical models (the probability distribution functions) depend on one or several parameters. The parameters vary over a specified range, called the *parameter space*.

A fundamental difficulty in statistical analysis is the choice of an appropriate model,

estimating and determining the order of dimension of a model. This is a common problem when the statistical model contains many parameters. The main purpose of model selection is to understand structure of the observed data. Thus, one needs to select a model. In the following section, definitions which needs to model selection are considered.

1.3 Definitions

The objective of model simplicity is noted by *parsimony*. Occam's Razor is a principle credited to the medieval English philosopher and Franciscan monk William of Ockham $(1285-1349)$. "Plurality should not posited without necessity", is a quote of William of Ockham. Translated to statistical modeling, Occam's Razor is sometimes referred to as the law of parsimony which notes no more causes should be assumed than those will account for the effect. In other words, everything should be made as simple as possible, but not simpler (Albert Einstein). Choose the simplest model which adequately fits to the data is a goal of model selection.

The *true* or *generating* model is the model that presumably gave rise to the data, and it will be showed by probability density function, p.d.f. (or probability mass function, p.m.f.) $h(.)$.

The *candidate* (*postulate*, *approximating*, *competing* or *proposed*) model is the model that could potentially be used to describe the data. It is chosen such that is close, in some sense, to the true model. It is noted by $f(.,\theta)$ or $g(.,\gamma)$ where θ and γ are vector of the parameters.

A collection of candidate models is a *candidate family*. In a model selection framework, this family is formulated to represent all interest candidate models . It is shown with $\mathcal{F}_\theta = \{f(.,\theta); \theta \in \Theta \subset \Re^p\}$ and $\mathcal{G}_\gamma = \{g(.,\gamma); \gamma \in \Gamma \subset \Re^q\}$.

A candidate model which has the same structure as the true model is called *correctly* (or *well*) *specified* model. If it is not correctly specified, is called *mis-specified* model.

A *candidate* model which has been fitted to the data is called a *fitted* model, say $f(.,\hat{\theta}_n)$.

A *fitted* model that has a more complex structure than the *true* model is called an *overspecified* model. It includes more parameters, explanatory variables, etc.

A *fitted* model which has a more simplistic structure than the true model is known as

underfitted model.

The two models, $f(.;\theta)$ and $g(.;\gamma)$, are *nested* if one model can be obtained from the other one by imposing restriction(s). If two models can not be obtained from each others, are *non-nested* models. In other words, $f(.;\theta)$ and $g(.;\gamma)$ are non-nested models if $f(x;\theta) \cap g(x;\gamma) = \emptyset$ and $f(.;\theta)$ is nested in $g(.;\gamma)$ if $f(x;\theta) \subset g(x;\gamma)$, (Commenges et al, 2008).

Mis-specified, nested and non-nested models are considered in more details, latter.

To determine that which of the fitted models in the candidate collection models, best resembles the true model, one requires a measure which provides a suitable reflection of the disparity between the true model and a fitted candidate model. The Kullback-Leibler (Kullback and Leibler, 1951) divergence fulfills this objective. This measure is introduced in the section (2.6.2).

1.4 Model Selection

The evaluation of competing statistical models is central to the process of scientific inquiry. When the competing models are stated in the form of predictors from quantitative models, their adequacy with respect to observed data can be rigorously assessed. Given K plausible candidate models of the underlying process that has generated the observed data, we should like to know which model approximates the true process better than the other ones. More generally, we should like to know how much statistical evidence the data provide for each of K models, preferably in terms of likelihood (Royall, 1997) or the probability of each of models being correct (or the most correct, because the generating model may never to be known for certain). The process of evaluating candidate models is called *model selection*.

A straightforward solution to the problem of evaluating several candidate models is to select the model that gives the most accurate description of the data. However, the process of model selection is complicated by the fact that a model with many free parameters is more flexible than a model with only a few parameters. It is clearly not desirable to always deem the most complex model, is the best, and it is generally accepted that the best model is the one that provides an adequate account of the data while using a minimum number of parameters. Thus, any criterion for model selection needs to address this trade off between descriptive accuracy and minimizing the num-

ber of parameters.

On other hand, the likelihood theory is an important concept of inference from data, given a model. It assumes that the model is correctly specified and only the parameters in the structural model are to be estimated. In the next chapter, likelihood theory is considered.

Chapter 2

Likelihood Function and Related Topics

2.1 Introduction

A method of maximum likelihood is one of the most important tools for estimation and inference. A fundamental assumption underlying classical results on the properties of the maximum likelihood estimator (MLE) is that stochastic law which determines the behavior of the phenomena investigated (the true structure) is know to lie within a specified model. In other words, the probability model is assumed to be correctly specified. In many (if not most) circumstances, one may not have complete confidence that is so. In this case, properties of MLE should be considered, again.
In this chapter, first the likelihood function in classical situation is presented in section (2.2). Section (2.3) contains asymptotic theory for MLE. Section (2.4) describes lemma and theorems which are used in this work. Finally, section (2.5) gives mis-specification which includes quasi-maximum likelihood estimator, its properties, information matrix test and Kullback-Leibler information criterion.

2.2 The Likelihood Function in Classical Situation

The basic concepts of likelihood estimation are defined in this section.

Definition 2.2.1 (Likelihood Function) Suppose $X_1, \ldots, X_n$ are random variables with joint p.d.f. (or p.m.f.) $f(x;\theta)$ where $\theta \in \Theta$. Given observations, the *likelihood function* is defined as $L_f(\theta) = f(x_1, \ldots, x_n; \theta)$ which is a function of θ.

For each independently identical distributed (i.i.d.) random variables $X_1, \ldots, X_n$, the likelihood function $L_f(\theta)$ is real-valued function defined on the parameter space Θ.

Definition 2.2.2 (Maximum Likelihood Estimation) Suppose for a random sample $X_1, \ldots, X_n$, $L_f(\theta)$ is maximized over Θ at $\hat{\theta}_n$ such that:

$$\sup_{\theta \in \Theta} L_f(\theta) = L_f(\hat{\theta}_n),$$

where $\hat{\theta}_n \in \Theta$. Then the statistic $\hat{\theta}_n$ is called the *maximum likelihood estimator* (MLE) of θ.

The subscript n of $\hat{\theta}_n$ denotes the dependence of estimation on the number of observations, which assumed n can increase infinitely. This allows to consider the asymptotic behavior of the estimators which are obtained of n random sample of population.
The MLE has favorite properties. As the important one is the invariance under transformation which means that if $\varphi = g(\theta)$ where g is an arbitrary function and $\hat{\theta}_n$ is the MLE of θ, then $g(\hat{\theta}_n)$ is the MLE of φ.
There are essentially two methods for finding MLE:

(i) **Direct maximization**: Examine $L_f(\theta)$ directly to determine which value of θ maximizes $L_f(\theta)$ for a given observed values of the $X_1, \ldots, X_n$. This method is particularly useful when the range (or support) of the data depends on the parameters.

(ii) **Likelihood equations**: If the range of data does not depend on the parameter, the parameter space Θ is an open set, and the likelihood function is differentiable with respect to θ over Θ, then the MLE of θ satisfies the equations $\nabla_\theta \log L_f(\hat{\theta}_n) = 0$.

Note that log(.) shows the natural logarithm. The equations in (ii) are called the likelihood equations and $\log L_f(\theta)$ is called the log-likelihood function. ∇_θ is the gradient operator with respect to θ. The log-likelihood function is used for convenience. Because if $\hat{\theta}_n$ maximize $L_f(\theta)$, it also maximizes $\log L_f(\theta)$. In addition in the independent case, $L_f(\theta)$ expressed as a product, so $\log L_f(\theta)$ becomes sum, which is easier to differentiate. The likelihood equations can have multiple solutions, so it is important to check that a given solutions indeed maximizes the likelihood function.

2.3 Asymptotic Theory for MLE

Under some conditions as *regularity conditions* MLE is a consistent and asymptotically normal estimator of parameter. These conditions mainly relate to differentiability of the density and the ability of interchanging differentiation and integration. They are as follow:

Assumption (A1): Let $X_1, \ldots, X_n$ be i.i.d. random variables with p.d.f. (or p.m.f.) $f(x; \theta)$ where $\theta \in \Theta$.

Assumption (A2): The parameters are identifiable. It means if $\theta_1 \neq \theta_2$ then $f(x;\theta_1) \neq f(x;\theta_2)$.

Assumption (A3): The range of random variable X, say R, does not depend on θ and $f(x;\theta)$ is differentiable in θ.

Assumption (A4): The parameter space Θ contains an open set which the true parameter value, θ_0, is an its interior point.

Assumption (A5): For every $x \in R$, the density $f(x;\theta)$ is three times differentiable with respect to θ, the third derivative is continuous in θ, and $\int_R f(x;\theta)dx$ can be differentiated three times under the integral sign.

Assumption (A6): For $\theta_0 \in \Theta$, there exists a positive number c and a function $M(x)$ (both of them may depend on θ_0) such that

$$|\nabla_\theta^{(3)} \log f(x;\theta)| \leq M(x) \quad \forall x \in R,\ |\theta_0 - \theta| < c \ \text{ with, } \ E_{\theta_0}\{M(X)\} < \infty,$$

where $\nabla_\theta^{(i)}$ for $i = 1, \ldots, n$ is gradient operator with respect to $f(.;\theta)$.
LeCam (1953) showed that under regularity conditions for all n there exists a MLE, $\hat{\theta}_n$. In the next Theorem, it is shown that MLE is a consistent estimator. For more details see Stuart et al. (1999).

Theorem 2.3.1 *Let $X_1, \ldots, X_n$ be i.i.d. random variables distributed $f(x;\theta)$ and $L_f(\theta) = \prod_{t=1}^n f(x_t;\theta)$ be the likelihood function, and $\hat{\theta}_n$ denote the MLE of θ. Under regularity conditions on $f(x;\theta)$ and $L_f(\theta)$,*

$$\lim_{n \longrightarrow \infty} P(|\hat{\theta}_n - \theta_0| \geq \epsilon) = 0 \quad \forall \epsilon > 0 \quad \textit{and} \quad \forall \theta \in \Theta.$$

In other words, $\hat{\theta}_n \xrightarrow{P} \theta_0$.

Since θ_0 is the limiting of $\hat{\theta}_n$, $(\hat{\theta}_n - \theta_0) = O_P(1)$ which means $(\hat{\theta}_n - \theta_0)$ is bounded in probability, its asymptotic distribution function can be obtained.

Theorem 2.3.2 *Let $X_1, \ldots, X_n$ be i.i.d. random variables distributed $f(x;\theta)$. Under regularity conditions on $f(x;\theta)$ and likelihood function;*

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{L} N_p(0, I^{-1}(\theta_0)),$$

where $I(\theta)$ is the Fisher information.