UNIVERSITI PUTRA MALAYSIA

DYNAMIC ROBUST BOOTSTRAP ALGORITHM FOR LINEAR MODEL SELECTION USING LEAST TRIMMED SQUARES

HASSAN SAMI URAIBI

DYNAMIC ROBUST BOOTSTRAP ALGORITHM FOR LINEAR MODEL SELECTION USING LEAST TRIMMED SQUARES

By

HASSAN SAMI URAIBI

Thesis Submitted to the School of Graduate Studies, Universiti Putra Malaysia, in Fulfilment of the Requirements for the Degree of Master Science

September 2009

Dedicated to

My wife

My daughters

Sura , Shahed, Iman , Fatemah, Zainab, & Adyan.

The memory of my father

My Dear mother

Abstract of thesis presented to the Senate of Universiti Putra Malaysia in fulfilment of the requirement for the degree of Master of Science

DYNAMIC ROBUST BOOTSTRAP ALGORITHM FOR LINEAR MODEL SELECTION USING LEAST TRIMMED SQUARES

By

HASSAN S. URAIBI

September 2009

Chairman: Associate Professor Habshah Binti Midi, PhD

Faculty: Institute for Mathematical Research (INSPEM)

The Ordinary Least Squares (OLS) method is often used to estimate the parameters of a linear model. Under certain assumptions, the OLS estimates are the best linear unbiased estimates. One of the important assumptions of the linear model is that the error terms are normally distributed. Unfortunately, many researchers are not aware that the performance of the OLS can be very poor when the data set that one often makes a normal assumption, has a heavy-tailed distribution which may arise as a result of the presence of outliers. One way to deal with this problem is to use robust statistics which is less affected by the presence of outliers. Another possibility is to apply a bootstrap technique which does not rely on the normality assumption. In this thesis the usage of bootstrap technique is emphasized. It was a computer intensive method that can replace theoretical formulation with extensive use of computer. Unfortunately, many statistics practitioners are not aware of the fact that most of the classical bootstrap techniques are based on the OLS estimates which is sensitive to outliers. The problems are further

complicated when the percentage of outliers in the bootstrap samples are greater than the percentage of outliers in the original sample. To rectify this problem, we propose a Dynamic Robust Bootstrap-LTS based (DRBLTS) algorithm where the percentage of outliers in each bootstrap sample is detected. We modified the classical bootstrapping algorithm by developing a mechanism based on the robust LTS method to detect the correct number of outliers in the each bootstrap sample.

Kallel et al. (2002) proposed utilizing the bootstrap technique for model selection. They used the classical bootstrap method to estimate the bootstrap location and the scale parameters based on calculating the Mean of Squared Residual (MSR). It is now evident that the classical mean and classical standard deviation are easily affected by the presence of outliers. In this respect, we propose to incorporate our proposed DRBLTS in the bootstrap model selection technique. We also proposed to use an alternative robust location and scale estimates which are less affected by outliers instead of using the classical mean and classical standard deviation.

The performances of the newly proposed methods are investigated extensively by real data sets and simulations study. The effect of outliers is investigated at various percentage, i.e , 0%, 5%, 10%, 15% and 20%. The results show that the DRBLTS is more efficient than other estimators discussed in this thesis. The results on the model selection again signify that our proposed robust bootstrap model selection method is more robust than the classical bootstrap model selection.

Abstrak tesis yang dikemukakan kepada Senat Universiti Putra Malaysia sebagai memenuhi keperluan ijazah Master Sains

ALGORITMA BOOTSTRAP TEGUH DINAMIK UNTUK PEMILIHAN MODEL LINEAR MENGGUNAKAN KUASA DUA TERCANTAS TERKECIL

Oleh

HASSAN S. URAIBI

September 2009

Pengerusi: Associate Professor Habshah Binti Midi, Ph.D.

Institut Penyelidikan Matematik

Kaedah Kuasadua Terkecil Biasa (OLS) selalu digunakan untuk menganggar parameter model linear. Dalam andaian tertentu, penganggar OLS adalah penganggar saksama linear terbaik Salah satu daripada andaian yang penting tentang model linear adalah ralat bey taburan normal. Malangnya, kebanyakan penyelidik tidak sedar bahawa prestasi OLS boleh menjadi sangat lemah apabila set data yang biasa dianggap bey taburan normal mempunyai taburan yang berhujung tebal yang disebabkan kehadiran titik terpencil. Salah satu cara untuk mengatasi masalah ini adalah dengan menggunakan statistik teguh kurang yang dipengaruhi oleh titik terpencil. Antara kemungkinan lain adalah dengan menggunakan teknik '*bootstrap*'yang tidak bergantung kepada andaian normal. Dalam tesis ini, kegunaan teknik '*bootstrap*' adalah dititikberatkan. Ia merupakan kaedah intensif komputer yang boleh menggantikan perumusan teon dengan menggunakan komputer secara meluas. Malangnya,

kebanyakan pengamal statistik tidak sedar kenyataannya bahawa kebanyakan daripada teknik '*bootstrap*' klasik adalah berdasarkan kepada penganggar OLS di mana ianya sensitif terhadap titik terpencil. Masalah akan menjadi lebih sukar apabila peratus titik terpencil dalam sampel bootstrap adalah lebih besar berbanding dengan peratus titik terpencil dalam sampel asal. Untuk menyelesaikan masalah ini kita mencadangkan algoritma berdasarkan '*bootstrap*' Teguh Dinamik-LTS (DRBLTS)' dimana peratus titik terpencil di dalam setiap '*bootstrap*' sampel dikesan. Kita mengubahsuai algoritma '*bootstrap*' klasik dengan membina mekanisme berdasarkan kepada kaedah teguh LTS untuk mengesan bilangan outliers yang betul dalam setiap '*bootstrap*' sampel..

Kallel et al. (2002) mencadangkan teknik '*bootstrap*' digunakan untuk pemilihon model. Mereka menggunakan kaedah '*bootstrap*' klasik untuk mengganggarkan lokasi '*bootstrap*' dan skala parameter berdasarkan kepada min kuasa dua reja (MRS). Sekarang terbukti bahawa min klasik dan sisihan piawai klasik mudah dipengaruhi oleh kehadiran titik terpencil. Oleh itu, kami mencadangkan untuk menggabungkan DRBLTS dalam teknik pemilihan model '*bootstrap*'. Kami juga mencadangkan untuk menggunakan pengangar lokasi teguh dan penganggar skala teguh yang kurang dipengaruhi oleh titik terpencil selain menggunakan min klasik dan sisihan piawai klasik.

Prestasi kaedah baru yang dicadangkan dikaji secara meluas menggunakan dengan set data yang sebenar dan kajian simulasi. Keputusan- kajian menunjukkan bahawa penganrggar OLS lebih berjaya daripada kaedah yang dicadangkan dalam situasi di mana tiada titik terpencil dalam data. Kesan titik terpencil keatas kaedah yang dicadangkan telah diselidiki dalam pelbagai peratus iaitu 0%, 5%, 10%, 15% dan 20%. Keputusan juga menunjukkan bahawa DRBLTS adalah lebih efisien berbanding dengan penganggar-penganggar yang lain yang telah dibincangkan di dalam tesis ini apabila titik terpencil hadir di dalam data. Keputusan bagi pemilihan model sekali lagi menunjukkan bahawa kaedah pemilihan model '*bootstrap*' teguh adalah lebih teguh berbanding dengan pemilihan model teguh'*bootstrap*' klasik.

ACKNOWLEDGEMENTS

Praise be to Allah for every thing and thanks a lot for him for helping me in good and bad times. Praise be to Allah who made Dr.Habshah Midi my supervisor in the master program. I would like to express my deep gratitude and warmest thanks to her invaluable guidance, her encouraging and her supporting in every stage of my thesis research. I also greatly value her judgment, her friendship, her kindness and her encouragement to study the robust regression and bootstrap technique. She provided me many opportunities for growth: from reading papers, writing a survey, turning ideas to implementation, many workshops, getting through the inevitable research setbacks and finishing thesis. She motivated me to enjoy a research work and I learned from her that it can provide me with lifetime benefits. Really, I consider myself lucky to access her supervision.

My thanks goes to the members of my supervisory committee members, Senior Lecturer Dr. Bashar Abdul Aziz Majeed Al-Talib, and Senior Lecturer Dr. Jabar Hassan Yousif for their invaluable discussions, comments, and help.

Special thanks to professor Dr. A. M. H. Rahmatullah Imon, statistics professor from Bangladesh, and professor Dr. Ricardo A. Maronna from Argentina for their useful remarks.

I gratefully acknowledge the financial support from the Universiti Putra Malaysia (UPM) as my sponsor during my studies. I would also like to extend my thanks to all members of Institute of Mathematical Research (INSPEM), UPM, for their kind assistance during my study.

I certify that a thesis Examination Committee has met on 16th September 2009 to conduct the final examination of Hassan S. Uraibi on his thesis entitled " Dynamic Robust Bootstrap Algorithm for Linear Regression Model Selection using Least Trimmed Squares (LTS)" in accordance with Universities and University Colleges Act 1971 and the Constitution of the Universiti Putra Malaysia [P.U.(A) 106] 15 March 1998. The Committee recommends that the student be awarded the M.Sc of Statistics.

Members of the Examination Committee are as follows:

Noor Akma Ibrahim, Ph.D
Associate Professor
Institute for Mathematical Research
Universiti Putra Malaysia
(Chairman)

Kassim Haron, PhD
Associate Professor
Faculty of Science
Universiti Putra Malaysia
(Internal Examiner)

Mohd Rizam Abu Bakar, PhD
Associate Professor
Faculty of Science
Universiti Putra Malaysia
(Internal Examiner)

Anton Abdulbasah Kamil, PhD
Associate Professor
School of Distance Education
Universiti Sains Malaysia
(External Examiner)

BUJANG KIM HUAT, PhD
Professor and Deputy Dean
School of Graduate Studies
Universiti Putra Malaysia

Date: 24 November 2009

This thesis was submitted to the Senate of Univirsiti Putra Malaysia and has been accepted as fulfilment of the requirement for the degree of Master of science. The members of the Supervisory Committee were as follows:

Habshah Binti Midi, PhD
Associate Professor
Laboratory of Applied and Computational Statistics,
Institute for Mathematical Research
Universiti Putra Malaysia
(Chairman)

Bashar Abdul Aziz Majeed Al-Talib, PhD
Senior Lecturer
Department of Mathematic
Faculty of science
Universiti Putra Malaysia
(Member)

Jabar Hassan Yousif, PhD
Senior Lecturer
Faculty of Computing and Information Technology
University of Sohar
(Member)

HASANAH MOHD GHAZALI, PhD
Professor and Dean
School of Graduate Studies
Universiti Putra Malaysia

Date: 10. December 2009

DECLARATION

I declare that the thesis is my original work except for quotations and citations which have been duly acknowledged. I also declare that it has not been previously, and is not concurrently, submitted for any other degree at Universiti Putra Malaysia or at any other institution.

HASSAN S. URAIBI

Date:

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

The general purpose of linear regression is to predict the behavior of response variable from some explanatory variable(s). In another word it assesses the degree of relationship between one response variable and one variable (simple regression) or more than one variable (Multiple regression). For verifying this task, a commonly used procedure is the ordinary least squares method (OLS). Historically it’s well known; easy of computation is the main reason OLS method had been initially used until today. Gauss in 1875 and Legendre in 1805 independently discovered the method of least squares for regression model. Lengendre in 1805 was the first to publish his results related to method of least squares, although Gauss is generally recognized as the “father “of least squares (Saccucci, 1985). As there is no computer when it was discovered, the OLS was extremely useful because it could be computed explicitly from the data through the use of matrix algebra (Anderson, *et al.*, 2001).

Multiple linear regression is the central model in this thesis. The general linear regression model can be written in a matrix form as follows:

$$y = X\beta + \varepsilon \tag{1.1}$$

where *y* is a $n \times 1$ vector, representing the observed response variable, *X* is the $n \times p$ matrix of predictor variables, β is unknown $p \times 1$ vector of regression parameters and ε is an $n \times 1$ vector of random errors assumed to be independent normally distributed with mean 0 and variance matrix σ^2. The Ordinary Least Squares method is often used to estimate the parameters of the model. It is a very popular method because of tradition and ease of computation. The OLS estimates are obtained by minimizing the error sum of squares. In order to use the regression correctly, the assumptions of OLS need to be met. These assumptions are as follows: (1) the errors are normally distributed, (2) the errors have the same variance at all levels of the independent variables (homoskedastic),(3) the explanatory variables are independent, also no correlation between explanatory variables and residuals, (4) the variables are measured without error (Anderson, 2001). When the OLS estimates satisfy all the above assumptions, the OLS is the Best Linear Unbiased Estimator (BLUE) which implies that among all the unbiased estimators, the OLS produces the minimum variance. However, in real situation, usually these assumptions are not met. When the assumptions are not met, the OLS can be highly inefficient, resulting in low power (Wilcox, 1997). In addition to that, the confidence bands become wider with increased alpha levels (Wilcox, 1997). The OLS approach may also produce unstable estimates when the assumption of normality of errors is not met (Ryan, 1997).

Unfortunately, many statistics practitioners are not aware of the fact that the violation of the normality assumption of the error terms may be due to one or more outliers in the data. Maronna *et al.* (2006) define outliers as observations that are well separated from the majority of the data or in some way deviate from the general pattern of the data. Fox (2003) considers the outliers in a linear model, a value of the response variable that is conditionally unusual given the values of the explanatory variables. Rouseeuw and Leroy (2003) describe regression outliers are cases for which $(x_{i1}, x_{i2}, \ldots, x_{ip}, y_i)$ deviate from the linear relation followed by the majority of the data, taking into account both the explanatory variables and response variable simultaneously. Outliers can occur for a variety of reasons including data entry errors, non-homogeneity:

Skyler J. Granmer (2005) stated that "Sometimes the data are not a homogeneous set to which a single model will apply, but rather a heterogeneous set of two or more types of cases".

model weaknesses, when the statistical model has no ability to represent a particular phenomenon thereby, is considered weak model, because most the statistical models are approximations to physical processes. The reasons of weak models may be due to randomness of human behaviors, left out variable, incomplete model, aggregation error and measurement error that are known error in equations and faulty distributional assumptions,

Incorrect assumptions about the distribution of the data can also lead to the presence of suspected outliers [e.g., Iglewicz & Hoaglin, 1993].

Outliers can occur in three directions. Rosseeuw and Zomeren (1990) described outliers in the X-direction as leverage points and if they are influential then they are generally known as high leverage points. The second types of outliers occur in the Y-direction. This type of outlier has a data point with a large squared residual from the fit. The third types of outliers occur in both X and Y directions, simultaneously.

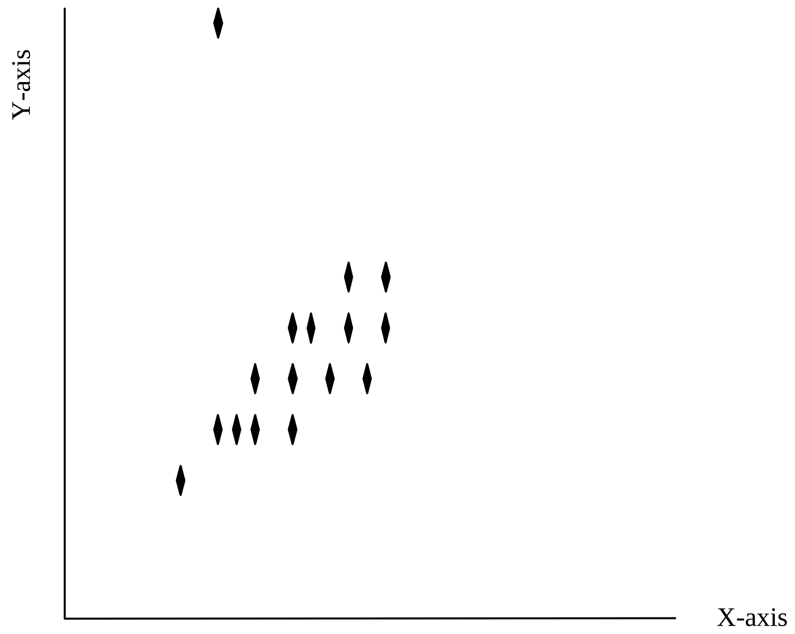

Figure 1.1 The Y- axis outlier.

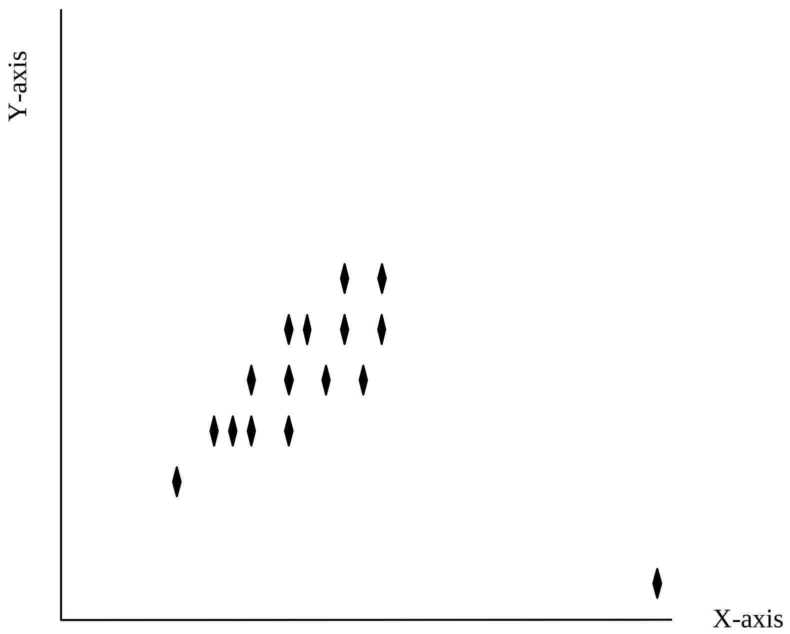

Figure 1.2 Tthe X-axis outlier.

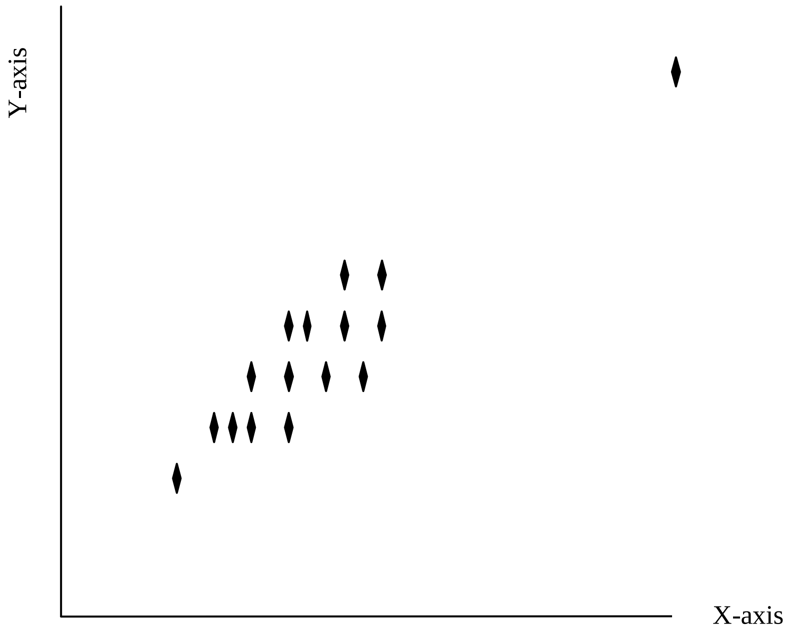

Figure 1.3 Both X-axis and Y-axis outlier.

The classical Ordinary Least Squares (OLS) method has long been subjugated the literature and applications of linear models. According to Gauss-Markov theorems, the OLS is the optimal procedure under the assumption that the distribution of the errors is normal. Many researchers are not aware that the performance of the OLS can be very poor when the data set that one often makes a normal assumption, has a heavy-tailed distribution which may arise as a result of outliers. Outliers which arise from 'bad' data points may have large influence on the OLS estimates. According to Hampel *et al.* (1986), the existence of 1-10% outliers in a routine data is rather rule than exceptions. Midi *et al.* (2009) pointed out that the detection of outliers is crucial due to their responsibility for misleading conclusion about the fitting of multiple linear regression model, causing multicollinearity problems, masking and swamping of outliers.

Chatterjee , Hadi, and Price;(2000) stated that "Masking occurs when the data contain outliers but we fail to detect them. This can happen because some of the outliers may be hidden by other outliers in the data. Swamping occurs when we wrongly declare some of the non-outlying points as outliers. This can occur because outliers tend to pull the regression equation toward them, hence make other points lie far from the fitted equation. Thus, masking is a false negative decision whereas swamping is a false positive."

Hampel (1971) pointed out that even one single outlier can have an arbitrary large effect on the OLS estimates. One of central concepts to understand robust regression technique is the breakdown points (BP). Hampel (1971) introduced a BP as the proportion of outliers that it would take to render the estimator useless. The robustness of each estimator is measured by the BP. An estimator becomes more robust as the value of BP increases. The BP of the OLS estimator is 0% which implies that it can be easily affected by a single outlier. A better approach is to consider a robust procedure. This procedure fits a regression by using estimators that dampen the impact of unusual observations or outliers; those points lying far away from the pattern formed by the good points and has large residuals from the robust fit. According to Giloni *et al.* (2006), robust methods are those methods that can fit the bulk of the data well. It is worth mentioning that the results obtained from robust methods are expected to be fairly close to the classical methods in the situation where there is no outlier(s) in the data sets. Several works on robust estimation have been proposed in the literature. Among them are Edgeworth proposed the Least Absolute Values (LAV) estimator in 1887, and also Huber (1973) who introduced M-estimators. However, none of these estimators achieved high breakdown point. Rousseeuw & Leroy in 1987 introduced the most robust estimator having the highest possible breakdown point of n/2 or 50% which is known as Least Median Squares (LMS) and Least Trimmed of Squares (LTS). Yohai (1987) improved further the efficiency of the high breakdown estimators by introducing the MM-estimators. If a robust estimation technique has 50% BP then 50% of the data could contain outliers and the coefficients would remain usable (Hampel *et al.*, 1986). In the literature several methods proposed to detect the outlying observations problem, according to their impact and location. (see: Huber P.J ; 1973, Cook; 1977, Belsley Kuh and Welsch; 1980, Hawkins; 1980, Velleman and Welsch; 1981, Atkinson; 1982 , Cook and Weisberge; 1982 , Rousseeuw; 1984, Rousseeuw and Yohai; 1984, Rousseeuw; 1985, Rousseeuw and Leroy; 2003, Chatterjee and Hadi; 1988, Rousseeuw and Zomeren; 1990, Fox; 1991, Barrett and Lewis; 1994, Huber M. and Rousseeuw; 1996, Habshah Midi; 1999, Chatterjee , Hadi, and Price; 2000, Hampel F;2000, Montgomery, Peck, and Vining;2001,Imon; 2002; 2005a; 2005b; 2007, Habshah Midi;2002, Imon and Ali;2005, Midi *et al.*, 2009.

One important aspect in statistical inference is to acquire the standard errors of the parameter estimates and to construct the T-statistics and confidence intervals for the parameters of a model. The OLS technique is often used to estimate the parameters of a model. The construction of confidence intervals requires that the estimates can be treated as samples from a normal distribution. Nonetheless, many measurements are not normal and have a heavy tailed distribution which may be the result of outliers. In this situation, we may use an alternative method such as robust method or the bootstrapping method, which is a distribution free method. The Bootstrapping method, which was introduced by Efron in 1979 , has been increasingly popular because it has many interesting properties. The basic idea of bootstrapping method is to generate a large number of sub-samples by randomly drawing observations with replacement from the original dataset or full sample. These sub-samples are then being termed as bootstrap samples and are used to recalculate the estimates of the regression coefficients. In fact re-sampling methods do not need some resampling assumptions that have related to the form of the estimator distribution in the ordinary sampling techniques, because the sample is thought as population (Sahinler, 2007). Some re-sampling procedures such as jackknife (Quenouille, 1949), permutation methods that introduced by Fisher and Pitman in 1930, and use of computers to do simulation also goes back to the early days of computing in the late 1940. They were introduced before nonparametric bootstrap that was introduced by Efron in 1979, who was unified the ideas and connected the simple nonparametric for independent and identically distributed (iid) observations, which resamples the data with replacement (Chernick, 2008). Bootstrap method has been successful in attracting statistics practitioners as its usage does not rely on the normality assumption. An interesting feature of the bootstrap method is that it can provide the standard errors of any complicated estimator without requiring any theoretical calculations. These interesting properties of the bootstrap method have to be traded off with computational cost and time. There are considerable papers that deal with bootstrap methods, see Efron and

Tibshiriani (1986) and Efron and Tibshiriani (1993). Kallel *et al.* (2002) proposed using the bootstrap technique for model selection. They used the random –x Re-Sampling together with the OLS method in their bootstrapping algorithm. Furthermore, the computation of the bootstrap location and bootstrap scale estimates are based on the classical mean and classical standard deviation formulation. As already been mentioned, the OLS is very sensitive to the presence of outliers and will produce less efficient results. One possible approach to deal with this problem is to incorporate a robust method which is not sensitive to outliers in the bootstrapping algorithm. In addition of using the robust method, we shall propose using a robust location and robust scale formulation for the bootstrap estimates. Hence a new robust bootstrap method is proposed for model selection criteria.

However, the development of robust bootstrap methods in the presence of outliers has received little attention. There are not many papers that deal with robust bootstrapping methods in linear regression. Amado and Pires (2004) proposed a resampling plan which is not so much affected by the outlying observations. They applied re-sampling probabilities to ascribe more importance to some samples values than others, but not in the context of linear regression. Singh (1998) robustified the bootstrap method by applying winsorization for certain L and M estimators. But according to Amado and Pirez (2004) this winsorized bootstrap is difficult to apply to multivariate samples. Imon and Ali (2005) proposed a Diagnostics – Before-Bootstrap whereby the suspected outliers are identified and omitted from the analysis before performing bootstrap with the remaining set of