MIDDLE EAST TECHNICAL UNIVERSITY
DEPARTMENT OF STATISTICS

Bootstrapping Conic Multivariate Adaptive Regression Splines (Bcmars)

İnci Batmaz, Ceyda Yazıcı, Fatma Yerlikaya-Özkurt

METU-STAT-Technical Report-2012- 002

March, 2012

DEPARTMENT OF STATISTICS
MIDDLE EAST TECHNICAL UNIVERSITY
ANKARA 06531 – TURKEY

TECHNICAL REPORT

A COMPUTATIONAL APPROACH TO NONPARAMETRIC REGRESSION: BOOTSTRAPPING CMARS METHOD

ABSTRACT

Bootstrapping is a computer-intensive statistical method which treats the data set as a population and draws samples from it with replacement. This resampling method has wide application areas especially in mathematically intractable problems. In this study, it is used to obtain the empirical distributions of the parameters to determine whether they are statistically significant or not in a special case of nonparametric regression, Conic Multivariate Adaptive Regression Splines (CMARS). CMARS is the modified version of the well-known nonparametric regression model, Multivariate Adaptive Regression Splines (MARS), which uses conic quadratic optimization (CQP). CMARS is at least as complex as MARS even though performs better with respect to several criteria. To achieve a better performance of CMARS with a less complex model, three different bootstrapping regression methods, namely, Random-X, Fixed-X and Wild Bootstrap are applied on four data sets with different size and scale. Then, the performances of the models are compared using various criteria including accuracy, precision, complexity, stability, robustness and efficiency. The results imply that Random-X method produces more precise, accurate and less complex models for medium size and medium scale data.

Keywords: Bootstrapping Regression, Conic Multivariate Adaptive Regression Splines, Fixed-X Resampling, Random-X Resampling, Wild Bootstrap

1. INTRODUCTION

Computational Statistics, which is a newer branch of statistics, is the method that uses algorithms dependent on computers in order to introduce a new methodology (Wegman, 1988). Computer-intensive statistical methods and visualization are the basic examples of this approach. The developments in the computer science make these methods feasible and popular especially after 1980s. Storing huge and high-dimensional data became easier with these improvements.

Bootstrap methods, classification and regression trees, generalized additive models and nonparametric regression are the basic methods of computer-intensive statistical methods which is another name for computational statistics (Efron and Tibshirani, 1991). Computational methods are preferred when numerically tractable and computationally intensive questions of interest exist. Thus, these approaches give a way of solving the problem.

Multivariate Adaptive Regression Splines (MARS) is a nonparametric regression, which is published by Friedman in 1991. MARS is widely used in modeling including biology, finance and engineering. This model is advantageous for handling the nonlinearity in the data. The model construction includes two parts: forward and backward. In the forward part, a large model is constructed. Then, some of the terms in the model are removed in the backward. Yerlikaya (2008) proposed a modification on the backward part of the model and call the new model as CMARS (Conic MARS). Later, Weber et al (2011) improved the model. However, the number of terms in the CMARS model can be more than the terms in the MARS model which leads a model at least as complex as MARS model.

The aim of this study is to decrease the complexity of the CMARS models. However, the mathematical intractability appears here as the lack of distribution fitting. If the distributions

of the parameters are known, the statistical significance can be tested with hypothesis testing or by constructing confidence intervals. Unfortunately, in the nonparametric regression the distributions of the parameters are not known in advance.

In this study, an empirical distribution is tried to be fitted to each parameter by a computational statistics method called bootstrap resampling. Bootstrap is a computer-intensive method that is heavily dependent on computers (Hjorth, 1994). In this approach, samples are drawn from the originals sample with replacement. For each bootstrap sample, the parameter of interest is calculated. The statistically significant model parameters are determined with the help of this method and insignificant terms are removed. For this purpose, three different bootstrapping regression methods, namely Fixed-X, Random-X and Wild Bootstrapping are run on four data sets chosen with respect to several criteria. Then, the performances of the models are compared according to complexity, stability, accuracy, precision, robustness and efficiency.

In section 2, a brief literature review is given. Then, the methods, including MARS, CMARS and bootstrap resampling are expressed in section 3. Applications and findings are discussed in the next section. In section 5, results and discussions are presented. Then, conclusions and further studies are given in section 6.

2. LITERATURE REVIEW

Ramanathan (2002) defines the models as the underlying, logical structure of the all analysis related with the social, economic or physical systems. A model represents the characteristics of the system and basic framework of the analysis. According to Hjorth (1994), models are simple form of the research phenomenon and the goal of models is to represent the ideas and conclusions. In statistics, formulating a model for the scientific question is the first step that should be conducted in an empirical study. After obtaining the data, the model should be estimated. If there are any assumptions in the model, these should be validated through hypothesis testing or with the help of visualization techniques. If the assumptions are satisfied, then the results can be interpreted statistically; otherwise necessary attempts have to be made to validate them.

Modeling has a wide range of applications, including engineering, biology, economics, and finance. In statistics, parametric and nonparametric models are the two major approaches. If the assumptions of the models are hold, parametric models give reliable results. On the other hand, in certain situations, it may not be possible to validate the assumptions of a parametric model. In this kind of situations, nonparametric modeling is recommended.

Multivariate Adaptive Regression Splines (MARS) is a nonparametric regression model, which is introduced by Jerome Friedman in 1991. This model is advantageous to handle high-dimensional data and approximate nonlinearity if exists. Moreover, MARS has a wide application from biology to finance.

In recent years, MARS has been conducted in a lot of studies. For instance, Kriner used this model for survival analysis in 2007. Zakeri et al. (2010) predict the energy expenditure for the first time in this research area by using MARS. Lin et al. (2011) applied MARS to time series data. In 2006, York et al. compared the power of the least squares fitting with polynomials with MARS. Deconinck et al. (2008) used MARS and Boosted Regression Trees (BRT) for the comparison of performances and show that it is better than BRT for fitting nonlinearities, being robust to small changes in the data and easier interpretation. Denison et al. (1998) provide a Bayesian algorithm for MARS.

In 2008, Yerlikaya proposed a modification to the model and call it as Conic MARS (CMARS). In this approach, the backward step of MARS is replaced with conic quadratic programming (CQP). Then, Batmaz et al. (2010) improved the model to fit nonlinearities better. The results indicate that MARS and CMARS perform better than Multiple Linear Regression (MLR). Last, in 2011, Weber et al. improved the model and compared it with MARS with respect to several criteria. The results show that CMARS is superior to MARS in terms of accuracy, robustness and stability under different data features.

Ozmen et al. (2011) propose a robustification to the CMARS. Alp et al. (2011) compare Generalized Additive Models (GAM), CMARS, MARS and Logistic Regression (LR) to detect a financial crisis before it occurs. In another study, Taylan et al. (2010), compare MARS and CMARS for classification and apply the method to a diabetes data set.

In the usual parametric modeling, the statistical significance of the model parameters can be investigated with hypothesis testing or constructing confidence intervals. However, if there is no information on the distributions of the parameters or normality assumption is not possible, and then methods in the computational statistics are suggested.

There are applications of computational methods for estimating the significance of parameters in a model. Efron (1988) applies bootstrap to Least Absolute Deviation (LAD) method. Fox (2002) uses Random-X and Fixed-X Resampling methods for robust regression which uses *M-estimator* with the Huber weight function. Also, Salibian-Barrera and Zamar (2002) apply bootstrapping to robust regression. Austin (2008) replaces bootstrap with backward elimination which results a better coverage in percentile CIs. Yetere-Kursun and Batmaz (2010) compare regression methods by employing different bootstrapping methods.

Flachaire (2003) compares the pairs bootstrap with wild bootstrap for heteroscedastic models. Efron and Tibshirani (1993) apply resampling residuals to a model based on Least Median of Squares (LMS). Chernick (2008) uses vector resampling for a kind of nonlinear model that is used in aerospace engineering. Montgomery et al. (2001) conduct bootstrapping residuals method to Michaelis-Menten model, which is a nonlinear regression.

3. METHODS

3.1.Multivariate Adaptive Regression Splines (MARS)

MARS is a nonparametric regression model in which there is no assumption between dependent and independent variables. It is developed by Friedman in 1991. In terms of approximating the nonlinearity in the data and handling the high dimensionality, MARS model is one of the best models. In addition to additive models, it is also possible to obtain the models with interaction terms.

MARS constructs models with two parts: forward and backward. In the forward part, a large model is obtained. However, this large model leads to overfitting. Thus, a backward part is conducted in order to remove terms that do not contribute to the model.

In nonparametric models, the relationship between response and predictor variables is not known. In general, the nonparametric regression model is defined as

$$y_i = f(\beta, x_i) + \varepsilon_i, \; i = 1,2,\ldots,n \qquad (1)$$

where β represents the parameters, n stands for the sample size and x_i shows the independent variables. In the model, f is the unknown form of the function.

In MARS model, instead of original predictor variables, a special form of them is used to construct models. These are called as Basis Function (BF) and represented with the following equations:

$$(x-t)_+ = \begin{cases} x-t, \text{ if } x>t, \\ 0, \text{ otherwise} \end{cases} \qquad (-x)_+ = \begin{cases} t-x, \text{ if } x<t, \\ 0, \text{ otherwise} \end{cases} \tag{2}$$

where $t \in \{x_{1,j}, x_{2,j}, \ldots, x_{n,j}\}$ and called as knot value. Here, these two BFs are the reflected pairs of each other. In the notation, p represents the number of independent variables.

Here, the purpose is to construct reflected pairs for each predictor at the knot value of x_{ij}. The collection of BFs are represented by

$$C = \{(x_j - t)_+, (-x_j)_+\} \tag{3}$$

The multivariate spline BFs take the following form to employ the BF that is tensor products of univariate spline functions:

$$B_m(x) = \prod_{k=1}^{K_m} [s_{km}(x_{(m)} - t_{km})_+], \tag{4}$$

where K_m represents the number of truncated functions in the m^{th} BF, x_{km} shows the input variable corresponding to the k^{th} truncated linear function in the m^{th} BF and t_{km} is the corresponding knot value and s_{km} takes the value of 1 or -1.

The MARS model is defined as

$$f(x) = \beta_0 + \sum_{m=1}^{M} \beta_m h_m(x) + \varepsilon, \tag{5}$$

where each h_m belongs to the set C and M represents the number of BFs in the current model. Given a choice for the h_m, the coefficients for the parameters (β_m) are estimated by minimizing the Residual Sum of Squares (RSS) with the same method similar to the one used in the usual Multiple Linear Regression (MLR). The important point is to determine the $h_m(x)$. The constant function $h_0(x) = 1$ is the first function that is used, and all functions in C are considered as candidate functions.

The decision on adding a new BF to the current model is explained with the following algorithm. Let M represent the current model set. The BFs in the current model are multiplied by the BFs in the candidate set C (with their reflected pairs) as shown below:

$$\hat{\beta}_{M+1} h_l(x)(x_j - t)_+ + \hat{\beta}_{M+2} h_l(x)(-x_j)_+, h_l \in M. \tag{6}$$

The BF which causes the most amount of reduction in the residual error is added to the model first. The parameters are determined by the LS approach. When the maximum number of terms (determined by the user) is obtained, the forward part finishes. After obtaining the large

model, backward part starts due to overfitting. In this step, a term in the model whose deletion causes the least amount of residual squared error is deleted first. This procedure estimates the best model, $\hat{f}_M$, of each size (number of terms) M. Cross validation (CV) is a possible solution for finding the optimal value of M. However, *generalized cross validation (GCV)* is used due to computational purposes. The GCV is defined as

$$GCV = \frac{\sum_{i=1}^{n} \left(y_i - \hat{f}_M(x_i) \right)^2}{\left(1 - C(M)/n \right)^2}, \tag{7}$$

where *n* represents the number of data samples. The numerator of the GCV is the usual RSS.

In general, $C(M)$ is calculated by using the following formula:

$$C(M) = trace\,(B(B^T B)^{-1} B^T) + 1. \tag{8}$$

$C(M)$ represents the cost penalty measure of a model in which there are M BFs. When the minimum value of the GCV is obtained, the MARS model is constructed.

3.2. Conic MARS (CMARS)

CMARS is an improved version of MARS. Yerlikaya (2008) proposed the model and later, it is improved by and Weber et al. (2011). It uses the BFs coming from forward step of the MARS and applies conic quadratic programming for the backward step to prevent overfitting. Instead of backward step, Penalized Residual Sum of Squares (PRSS) is constructed as follows:

$$PRSS = \sum_{i=1}^{n} \left(y_i - f(\bar{\boldsymbol{x}}_i) \right)^2 + \sum_{m=1}^{M_{\max}} \lambda_m \sum_{\substack{|\boldsymbol{\alpha}|=1 \\ \boldsymbol{\alpha}=(\alpha_1,\alpha_2)^T}}^{2} \sum_{\substack{r<s \\ r,s \in V(m)}} \int \theta_m^2 \left[D_{r,s}^{\alpha} \psi_m(\boldsymbol{t}^m) \right]^2 d\boldsymbol{t}^m, \tag{9}$$

where $M_{\max}$ is the number of BFs reached at the end of the forward algorithm; $V(m) = \{ \kappa_j^m \mid j = 1,2,...,K_m \}$ is the variable set associated with m^{th} BF, ψ_m. $\boldsymbol{t}^m = (t_{m_1}, t_{m_2}, \ldots, t_{m_{K_m}})^T$ represents the variables which contribute to the m^{th} BF, ψ_m.

The λ_m values are always nonnegative and used as the *penalty parameters* (m=1,2,…, M_{max}). Moreover, the α values in the following term (13) is taken as $D_{r,s}^{\boldsymbol{\alpha}} \psi_m(\boldsymbol{t}^m) = \frac{\partial^{|\boldsymbol{\alpha}|} \psi_m}{\partial^{\alpha_1} t_r^m \partial^{\alpha_2} t_s^m}(\boldsymbol{t}^m)$, $\boldsymbol{\alpha} = (\alpha_1, \alpha_2), |\boldsymbol{\alpha}| = \alpha_1 + \alpha_2,$ where $\alpha_1, \alpha_2 \in \{0,1\}$.

If $\alpha_i = 2,$ the derivative $D_{r,s}^{\alpha} \psi_m(t^m)$ disappears, and by addressing indices $r < s,$ the Schwarz's Theorem can be applied.

The optimization approach to the problem takes both the accuracy and lower complexity into account. The term *accuracy* refers to the small sum of squares of errors. The tradeoff between these two terms are expressed by penalty parameters and solved by CQP.

After making some arrangements, the PRSS takes the following form;

$$PRSS = \sum_{i=1}^{n}\left(\bar{y}_i - \theta_0 - \sum_{m=1}^{M}\theta_m \psi_m\left(\bar{\boldsymbol{x}}_i^m\right) - \sum_{m=M+1}^{M_{\max}}\theta_m \psi_m\left(\bar{\boldsymbol{x}}_i^m\right)\right)^2$$
$$+ \sum_{m=1}^{M_{\max}} \lambda_m \sum_{\substack{|\alpha|=1 \\ \alpha=(\alpha_1,\alpha_2)}}^{2} \sum_{\substack{r<s \\ r,s\in V(m)}} \int \theta_m^2 \left[D_{r,s}^{\alpha}\psi_m(\boldsymbol{t}^m)\right]^2 d\boldsymbol{t}^m, \tag{10}$$

where $\bar{\boldsymbol{x}}_i = (\bar{x}_{i,1}, \bar{x}_{i,2}, ..., \bar{x}_{i,p})^T$ represents any of the independent variables and $\bar{\boldsymbol{x}}_i^m = (\bar{x}_{i,\kappa_1}, \bar{x}_{i,\kappa_2}, ..., \bar{x}_{i,\kappa_{K_m}})^T$ stands for the corresponding projection vectors of $\bar{\boldsymbol{x}}_i$ onto those coordinates which contribute to the m^{th} BF, and they are related with the i^{th} output $\bar{y}_i$. Those coordinates are collected from the set V_m.

The input data which is represented by $\bar{\boldsymbol{x}}_l = (\bar{x}_{l,1}, \bar{x}_{l,2}, ..., \bar{x}_{l,p})^T$ generates a subdivision of any sufficiently large parallelpipe Q of IR^n .The parallelpipe, which is represented by Q, contains all the input data, and it is expressed as

$$Q = [a_1, b_1] \times [a_2, b_2] \times ... \times [a_p, b_p] = \prod_{j=1}^{p} Q_j \tag{11}$$

where $Q_j = [a_j, b_j]$, $a_j \le \bar{x}_{l,j} \le b_j$ $(j = 1,2,...,p;\ l = 1,2,...,N)$.

The parallelpipe is expressed as

$$Q = \bigcup_{\sigma^j=0}^{n} \prod_{j=1}^{p} \left[\bar{x}_{l_{\sigma^j},j}, \bar{x}_{l_{\sigma^{j+1}},j}\right] \tag{12}$$

So the PRSS takes the following form;

$$PRSS \approx \sum_{i=1}^{n} \left(y_i - \theta^T \psi(\bar{\boldsymbol{d}}_i)\right)^2$$
$$+ \sum_{m=1}^{M_{\max}} \lambda_m \sum_{\substack{|\alpha|=1 \\ \alpha=(\alpha_1,\alpha_2)^T}}^{2} \sum_{\substack{r<s \\ r,s\in V(m)}} \sum_{(\sigma^{\kappa_j})} \theta_m^2 \left[D_{r,s}^{\alpha}\psi_m(\bar{x}_{l_{\sigma^{\kappa_j}_j,\kappa_j^m}}, \bar{x}_{l_{\sigma^{\kappa_j}_j,\kappa_j^m}}, ..., \bar{x}_{l_{\sigma^{\kappa_j}_j,\kappa_j^m}}) \right]^2 \prod_{j=1}^{K_m} \left(\bar{x}_{l_{\sigma^{\kappa_j}_{j+1},\kappa_j^m}} - \bar{x}_{l_{\sigma^{\kappa_j}_j,\kappa_j^m}} \right),$$

where $\left(\sigma^{\kappa_j}\right)_{j=1,2,...,K_m} \in \{1,2,\ldots,n+1\}^{K_m}$. (13)

The following notations related with (σ^{κ_i}) are defined in order to use in the forward steps:

$$\hat{\boldsymbol{x}}_i^m = \left(\bar{x}_{l_{\sigma^{\kappa_j}_j,\kappa_j^m}}, \bar{x}_{l_{\sigma^{\kappa_j}_j,\kappa_j^m}}, ..., \bar{x}_{l_{\sigma^{\kappa_j}_j,\kappa_j^m}} \right), \tag{14}$$

$$\Delta\hat{\boldsymbol{x}}_i^m = \prod_{j=1}^{K_m} \left(\bar{x}_{l_{\sigma^{\kappa_j}_{j+1},\kappa_j^m}} - \bar{x}_{l_{\sigma^{\kappa_j}_j,\kappa_j^m}} \right). \tag{15}$$

The approximation to the PRSS is defined as

$$PRSS \approx \sum_{i=1}^{n} y_i - \boldsymbol{\theta}^T \boldsymbol{\psi}(\bar{\boldsymbol{d}}_i)^2 + \sum_{m=1}^{M_{\max}} \lambda_m \theta_m^2 \sum_{i=1}^{(N+1)^{K_m}} \left(\sum_{\substack{|\boldsymbol{\alpha}|=1 \\ \boldsymbol{\alpha}=(\alpha_1,\alpha_2)^T}}^{2} \sum_{\substack{r<s \\ r,s\in V(m)}} \left[D_{r,s}^{\alpha} \psi_m(\hat{\boldsymbol{x}}_i^m) \right]^2 \right) \Delta \hat{\boldsymbol{x}}_i^m. \tag{16}$$

$$PRSS \approx \left\| y - \psi(\bar{\theta}) \theta \right\|_2^2 + \sum_{m=1}^{M_{\max}} \lambda_m \sum_{i=1}^{(n+1)^{K_m}} L_m^2 \theta_m^2,$$

where $\boldsymbol{\psi}(\bar{\boldsymbol{d}}) = \boldsymbol{\psi}(\bar{\boldsymbol{d}}_1),...,\boldsymbol{\psi}(\bar{\boldsymbol{d}}_N)^T$ is a matrix with dimensions of $(N \times (M_{\max}+1))$, and $\|\cdot\|_2$ denotes the Euclidean norm and the numbers L_{im} are defined as

$$L_{im} = \left[\left(\sum_{\substack{|\boldsymbol{\alpha}|=1 \\ \boldsymbol{\alpha}=(\alpha_1,\alpha_2)^T}}^{2} \sum_{\substack{r<s \\ r,s\in V(m)}} \left[D_{r,s}^{\alpha} \psi_m(\hat{\boldsymbol{x}}_i^m) \right]^2 \right) \Delta \hat{\boldsymbol{x}}_i^m \right]^{1/2}. \tag{17}$$

The PRSS can be taken from the view point of CQP, a technique used for continuous optimization. Thus, the Tikhonov regularization problem can be formulated again by using the CQP. The optimization problem below is considered by putting an appropriate bound, *M*.

$$\begin{aligned} &\min \quad \left\| \boldsymbol{\psi}(\bar{\boldsymbol{d}})\boldsymbol{\theta} - \boldsymbol{y} \right\|_2^2 \\ &\text{subject to} \quad \left\| \boldsymbol{L}\boldsymbol{\theta} \right\|_2^2 \le M. \end{aligned} \tag{18}$$

Thus, the problem can be expressed as a CQP problem with the following way

$$\min_{\boldsymbol{x}} \ \boldsymbol{c}^T \boldsymbol{x}, \quad \text{subject to} \ \left\| \boldsymbol{D}_i \boldsymbol{x} - \boldsymbol{d}_i \right\|_2^2 \le \boldsymbol{p}_i^T \boldsymbol{x} - q_i \ \ (i = 1, 2, ..., k). \tag{19}$$

where

$\boldsymbol{c} = (1, 0_{M_{\max}+1}^T)^T$, $\boldsymbol{x} = (t, \boldsymbol{\theta}^T)^T$, $\boldsymbol{D}_1 = (0_n, \boldsymbol{\psi}(\bar{\boldsymbol{d}}))$, $\boldsymbol{d}_1 = \boldsymbol{y}$, $\boldsymbol{p}_1 = (1, 0, ..., 0)^T$, $q_1 = 0$,

$\boldsymbol{D}_2 = (0_{M_{\max}+1}, \boldsymbol{L})$, $\boldsymbol{d}_2 = \boldsymbol{0}_{M_{\max}+1}$, $\boldsymbol{p}_2 = 0_{M_{\max}+2}$ and $q_2 = -\sqrt{M}$.

The problem (Equation 18) should be reformulated to obtain the optimality condition as the following

$$\min_{t,\theta} \quad t,$$

$$\text{such that} \quad \chi = \begin{pmatrix} \mathbf{0}_n & \psi(\bar{\boldsymbol{d}}) \\ 1 & \mathbf{0}^T_{M_{\max}+1} \end{pmatrix} \begin{pmatrix} t \\ \theta \end{pmatrix} + \begin{pmatrix} -\boldsymbol{y} \\ 0 \end{pmatrix},$$

$$\eta = \begin{pmatrix} \mathbf{0}_{M_{\max}+1} & \boldsymbol{L} \\ 0 & \mathbf{0}^T_{M_{\max}+1} \end{pmatrix} \begin{pmatrix} t \\ \theta \end{pmatrix} + \begin{pmatrix} \mathbf{0}_{M_{\max}+1} \\ \sqrt{M} \end{pmatrix}, \tag{31}$$

$$\chi \in L^{n+1},\ \eta \in L^{M_{\max}+2},$$

where L^{n+1}, $L^{M_{\max}+2}$ are the $(n+1)$ and $(M_{\max}+2)$ dimensional *ice-cream* (or *second-order*, or *Lorentz*) *cones*, defined by

$$L^{n+1} = \quad \boldsymbol{x} = (x_1, x_2, ..., x_{n+1})^T \in \mathbb{R}^{n+1} \mid x_{n+1} \geq \sqrt{x_1^2 + x_2^2 + ... + x_n^2} \qquad (n \geq 1). \tag{32}$$

Moreover, $(t, \theta, \chi, \eta, \omega_1, \omega_2)$ is a *primal dual optimal solution* if and only if

$$\chi = \begin{pmatrix} \boldsymbol{0}_n & \psi(\bar{\boldsymbol{d}}) \\ 1 & \boldsymbol{0}^T_{M_{\max}+1} \end{pmatrix} \begin{pmatrix} t \\ \theta \end{pmatrix} + \begin{pmatrix} -y \\ 0 \end{pmatrix},$$

$$\eta = \begin{pmatrix} \boldsymbol{0}_{M_{\max}+1} & 1 \\ 0 & \boldsymbol{0}^T_{M_{\max}+1} \end{pmatrix} \begin{pmatrix} t \\ \theta \end{pmatrix} + \begin{pmatrix} \boldsymbol{0}_{M_{\max}+1} \\ \sqrt{M} \end{pmatrix},$$

$$\begin{pmatrix} \boldsymbol{0}^T_n & 1 \\ \psi(\bar{\boldsymbol{d}})^T & \boldsymbol{0}_{M_{\max}+1} \end{pmatrix} \omega_1 + \begin{pmatrix} \boldsymbol{0}^T_{M_{\max}+1} & 0 \\ L^T & \boldsymbol{0}_{M_{\max}+1} \end{pmatrix} \omega_2 = \begin{pmatrix} 1 \\ \boldsymbol{0}_{M_{\max+1}} \end{pmatrix}, \tag{33}$$

$$\omega_1^T \chi = 0,\ \omega_2^T = 0,$$

$$\omega_1 \in L^{n+1},\ \omega_2 \in L^{M_{\max}+2},\ \chi \in L^{n+1},\ \eta \in L^{M_{\max}+2}.$$

3.3 BOOTSTRAP

The bootstrap is a resampling technique which takes samples from the original data set with replacement. It is a data-based simulation method useful for producing inferences. The application of this method is not difficult, but depends heavily on computers. Thus, they are called computer-intensive methods (Chernick, 2008). The application of bootstrap includes estimation of standard errors and bias, constructing confidence intervals, hypothesis testing, classification, etc. The bootstrap procedure can be explained with the following steps.

1. Generate a random sample (x^{*b}) of size *n* (the same sample size with the original data) from the empirical distribution with replacement.
2. Compute the value of the statistic of interest for this sample.
3. Repeat steps 1-2 *B* times (i.e. b = 1,…, B).

3.3.1. Bootstrapping Regression

Let $Y_i = x_i^T \beta + \varepsilon_i, \quad for i = 1, \ldots, n$ be the usual MLR model. In the model, x_i represents the independent variables and β shows the parameters. The error terms, ε_i, are normally distributed with zero mean and constant variance. The parameters, β, are distributed normally.

If all assumptions of the model are satisfied, then the model is appropriate for the data and the results will be reliable. However, in the following cases there are some problems (Hjorth, 1994). If

- the model is non-linear,
- the statistical analysis of estimation has no direct classical solution,
- errors are not normally distributed,
- there are parameters dependent on another function.

Efron and Tibshirani (1993) indicate that bootstrap is applicable to general models including non-linearity of parameters; fitting methods different from LS approach by giving reasonable outputs. According to them, bootstrapping regression is applicable to models that have a mathematical form in addition to models that have no mathematical solution.

3.3.2.1 Fixed-X Resampling (Residual Resampling)

Random-X Resampling (Pairs Bootstrap)

It is recommended to be used when there is heteroscedasticity in the residual variance or correlation structure in the residuals, or it is suspected that some important parameters are missing in the model (Chernick, 2008).

Step 1: Select *B* bootstrap samples of $z_i^{'} = (y_i, x_{i1}, x_{i2}, \ldots, x_{ik}),\ i = 1, \ldots, n$

Step 2: Fit a model to the vector $z_i^{'}$ and obtain the estimates of parameters *(β)* and save them.

Step 3: Repeat this procedure B times and obtain bootstrap estimates of parameters.

The design X is assumed to be deterministic as in the usual regression approach. But this approach makes the X matrix random so the estimates will lead to the variability.

This method can be more advantages to be used in the following cases:

- If the distribution of the error terms is different for the independent variables (i.e. heteroscedasticity, skewness),
- If the non-linearity part is not well-defined,
- For large sample sizes, if the data consists influential observations, in case of heteorscedasticity or skewness.

3.3.2.2 Fixed-X Resampling (Residual Resampling)

In this model, the response values are taken as random due to the error components. Its use is recommended in case of identically distributed errors (Fox, 2002).

Step 1: Fit a model to the data and obtain the fitted values, $\hat{y}_i$ and the residuals, $\hat{\varepsilon}_i$.

Step 2: Select a bootstrap sample of residuals and add them to the fitted values. These new fitted values are now new response variables, $y_{new} = \hat{y}_i + \hat{\varepsilon}_b$.

Step 3: Fit a model to the original independent variables and new response variables. Obtain the new parameters, $y_{new} = X\beta + \varepsilon$.

Step 4: Repeat this procedure B times and collect the parameters.

This method can be more advantages to be used in the following cases:

- If there is no doubt about the adequacy of the model,
- If the predictors are considered as fixed,
- For small data sets or data with influential observations.

3.3.2.3 Wild Bootstrap

The wild bootstrap is a new approach for heteroscedastic models. According to Liu (1988), the errors of the model have two-point distribution which is called *Rademacher* distribution, and defined as follows:

$$f(x)=\begin{cases}0.5, & x=1\\ 0.5, & x=-1\\ 0, & otherwise\end{cases}. \tag{34}$$

Step 1: Fit a model to the data and obtain the fitted values, $\hat{y}_i$, and the residuals, $\hat{\varepsilon}_i$.

Step 2: These new fitted values are now new response variables, $y_{new}=\hat{y}_i+\hat{\varepsilon}_b$, where the error distribution is the f(x) given in (34).

Step 3: Repeat this procedure B times and collect the parameters.

In the wild bootstrap, the errors are randomly assigned as 1 or -1 and attached to the fitted values.

Flachaire (2005) suggests the use of wild bootstrap instead of pairs bootstrap in case of heteroscedasticity since the simulation studies give better results.

The choice of the bootstrapping regression model depends on how well the assumptions of the model are satisfied. For instance, if the model is MLR, then the errors must be independent from the covariates and must be i.i.d. Then, the Fixed-X resampling is reliable. However, Random-X resampling is not as conservative as the Fixed-X resampling. It performs better even when the assumptions are not satisfied.

Percentile Interval

Percentile interval uses the Empirical Cumulative Distribution Function (ECDF) of the bootstrap sample to find the upper and lower endpoints. It is defined as

$$\left(\hat{\theta}_B^{*(\alpha/2)},\hat{\theta}_B^{*(1-\alpha/2)}\right).$$

3.3.4. Bootstrap Estimate of Bias

Bias is used to investigate the performance of a measure (Martinez and Martinez, 2002). Actually, it measures the statistical accuracy of a measure. It is defined as

$$bias(T)=E[T]-\theta, \tag{35}$$

In general, it is the difference between the expected value of a statistic and the parameter value. For bootstrap estimate of bias, the empirical distribution of the parameter is used. It is defined as the following formula.

$$\hat{bias}_B=\bar{\hat{\theta}}^{*}-\hat{\theta}, \tag{36}$$

where $\bar{\hat{\theta}}^{*} = \frac{1}{B}\sum_{b=1}^{B}(\hat{\theta}^{*b})$. (37)

where $\bar{\hat{\theta}}^{*}$ is the mean of the values obtained by bootstrapping. The bias corrected estimate is explained by the following formula.

$$\tilde{\theta} = 2\hat{\theta} - \bar{\hat{\theta}}^{*}. \quad (38)$$

3.3.4. Cross Validation (CV) Technique and the Performance Criteria

In the comparison of models, 3-fold CV technique is used (Martinez and Martinez, 2002; Gentle, 2009). In this technique, data sets are randomly divided into three parts (folds). At each attempt, two folds (66.6% of observations) are used to develop models while the other fold (33.3% of observations) is kept to test them.

The performances of the models developed are evaluated with respect to different criteria. These include accuracy, precision, complexity, stability, robustness and efficiency. The accuracy criterion is used to measure the predictive ability of the models while precision criterion is used to determine how variable the parameter estimates are; the less variability indicates more precision. The MAE, R^2, PWI and PRESS measures are used to evaluate the models according to accuracy. On the other hand, the precision of parameter estimates are determined by their empirical CIs. Other criterion used in the comparisons is the complexity; it is measured by the MSE. Besides, the stabilities of the accuracy and complexity measures obtained from the training and test data sets are also evaluated. The definitions of these measures are placed in Appendix. Furthermore, robustness of the measures with respect to different data sets are evaluated by considering the standard deviations of the measures. Moreover, to assess the efficiency of the models build, computational run times are utilized.

4. APPLICATION AND FINDINGS

Table 1. Data Sets Used in the Comparisons

		Scale (p)	
	(n, p)	**Small**	**Medium**
Sample Size (n)	**Small**	Data Set 1: Concrete Slump (CS) (Yeh, 2007). (103,7)	Data Set 2: Uniform Sampling (US) (Kartal, 2007) (160,10)
	Medium	Data Set 3: PM10(Aldrin, 2006). (500,7)	Data Set 4: Forest Fires (FF) (Cortez and Morais, 2007). (517,11)

In order to evaluate and compare the performances of the models developed by using the MARS, CMARS and Bootstrapping CMARS (BCMARS) methods, they are run on four different data sets to observe the effects of certain data characteristics such as size and complexity on the methods' performances. In this study, two-third of the observations is used

for training and the rest is used for testing the model. Each fold is taken as testing once, thus 3 models are obtained for a data set.

To fit a MARS model to a data set, the R package *"Earth"* (Milborrow, 2009) is preferred due to the lack of MARS code in MATLAB. Then, the code written in MATLAB (2009a, The MathWorks, U.S.A.) by Yerlikaya (2008) and developed further by Batmaz et al. (2010) is used to obtain CMARS models. For optimization process, the MOSEK optimization software (6, MOSEK ApS, Denmark) is utilized. Then, all computations, including nonparametric bootstrap, are run using the code written in MATLAB.

The following steps belong to the algorithm followed for obtaining three different BCMARS models, labeled as BCMARS-1 (uses Fixed-X Resampling), BCMARS-2 (uses Random-X Resampling) and BCMARS-3 (uses Wild Bootstrap).

Step 1: The set of BFs (from the first part of the MARS algorithm) are obtained. The BFs are considered fixed and they will be used for bootstrapping.

Step 2: A CMARS Model is constructed and the optimal value of $\sqrt{M}$ is found. To achieve this, the curve of $\sqrt{RSS}$ versus norm of $L\theta$ in the log-log scale is obtained (see Figure 4). The optimal value of this curve is the corner point which is demonstrated by a red point. The selected value gives the best solution for both accuracy and complexity.

Step 3: Since there is not a distributional assumption, nonparametric bootstrap is used for the analysis.

- *BCMARS-1:* the original data is used to obtain the residuals and fitted values. The bootstrap sample of residuals are selected with replacement and added to the fitted values, so the new dependent variable is obtained. A model is constructed by using the fixed independent variables and this new dependent variable to obtain the parameters of BFs.
- *BCMARS-2:* the bootstrap sample of the data (including independent and dependent variables) is selected. This bootstrap sample and the BFs coming from Step 1 are used to obtain the parameters of the model (including the intercept).
- *BCMARS-3:* a model is fitted to the original data and the fitted values are obtained. The bootstrap sample of errors is obtained with Rademacher distribution. The fitted values and the bootstrap sample of errors are added to obtain new response variable. A model is constructed by using the fixed independent variables and this new dependent variable to obtain the parameters.

Step 4: Step 3 is repeated 1000 times and the ECDF of each parameter is obtained.

Step 5: For the significance level taken as $\alpha = 0.1$, the percentile CI of each parameter is constructed. If this interval includes zero, the corresponding BF is removed from the model.

Step 6: The Steps 2-5 are reapplied with the remaining BFs until all the CIs of the parameters do not include zero.

The percentile method is used for conducting the CIs, since there is no know form of the distribution of parameters. Efron and Tibshirani (1993) suggest the number of bootstrap samples to be as at least 1000 to construct percentile intervals. Then, the performance measures of each model obtained in three different ways are calculated. Moreover, the computational run time of the methods are recorded to be compared.

5. RESULTS AND DISCUSSION

In this section, it is aimed to compare the performances of the methods studied, namely MARS, CMARS, BCMARS (Fixed-X Resampling, Random-X Resampling and Wild Bootstrapping) in general (Section 5.1), and also, according to different features of data sets such as size (Section 5.2) and scale (Section 5.3). In these comparisons, various criteria including accuracy, precision, stability, efficiency (Section 5.4) and robustness are considered.

5.1. Comparison with respect to Overall Performances

The mean and standard deviations of measures obtained from four data sets are given in Table 2. These values are calculated for training and testing data sets in addition to the stability of measures. Definitions of the measures are given in Appendix. In this table, lower means for MAE, MSE and PRESS and higher means for R^2 and PWI measures indicate better performances. On the other hand, smaller standard deviations imply robustness for the corresponding measure. The following conclusions can be drawn from this table:

For training data sets:

- Fixed-X Resampling provides best performance with respect to MAE and R^2 accuracy measures. This method is the most robust among the others with respect to the same measures. These findings are also valid with respect to the complexity measure, MSE, as well.
- MARS, however, performs best with respect to the other accuracy measures PWI and PRESS. This method is the most robust among the others with respect to the same measures.
- When the bootstrapping models are compared among themselves, the Fixed-X Resampling method overperforms with respect to the means and spreads of all measures except the spread of PWI. Random-X Resampling is the most robust one with respect to the PWI measure.

For testing data sets:

- Random-X performs best with respect to most of the measures, namely MSE, R^2 and PRESS. It also produces more robust models for the same measures. Moreover, it gives the least complex models as well by providing the smallest MSE mean value.
- MARS has the best performance with respect to the only one accuracy measure, MAE. It is also the most robust for the same measure.
- CMARS, on the other hand, is the best performing and also the most robust method in terms of PWI.
- When only the bootstrapping methods are considered, Fixed-X Resampling is the best one with respect to the performance measure MAE, and Wild bootstrapping is the most robust one for the same performance measure. Moreover, Random-X Resampling has the highest PWI coverage, and Wild bootstrapping is the most robust with respect to PWI.

Table 2. Overall Performances (Mean±Std. Dev.) of the Methods

Performance Measures	**Training**				
	MARS	**CMARS**	**BCMARS-1**	**BCMARS-2**	**BCMARS-3**
MAE	0.3453 ±0.2336	0.4040 ±0.3980	0.3204*±0.2260**	0.3356 ±0.2263	0.4251 ±0.2797
MSE	0.4015 ±0.3064	0.6070 ±0.9080	0.3117*±0.2700**	0.4230 ±0.3990	0.5770 ±0.4950
R^2	0.6005 ±0.2797	0.5911 ±0.3407	0.6827*±0.2492**	0.6120 ±0.3350	0.5127 ±0.3398
PWI	0.9944*±0.0082**	0.9942 ±0.0082**	0.9909 ±0.0153	0.9932 ±0.0140	0.9855 ±0.0158
PRESS	0.0097*±0.0230**	72.0000 ±248.80	0.2390 ±0.7570	1.2090 ±3.0150	$13.5x10^6$±$4.7x10^6$
Performance Measures	**Testing**				
	MARS	**CMARS**	**BCMARS-1**	**BCMARS-2**	**BCMARS-3**
MAE	0.4576*±0.2956**	0.5800 ±0.4580	0.4838±0.3076	0.6460±0.6110	0.4977±0.2998
MSE	3.0700±7.0900	1.5780 ±2.1350	1.2670±1.9970	0.5480*±0.3660**	1.0720 ±1.2710
R^2	0.4480±0.3820	0.3630±0.4030	0.4500±0.3800	0.4530*±0.3770**	0.3840±0.4010
PWI	0.9930*±0.0108	0.9930*±0.0106**	0.9884±0.0177	0.989±0.0169	0.9878±0.0120
PRESS	470±996	491±287	459±1037	107.700*±189.10**	$1.4x10^6$±$0.5x10^6$
Performance Measures	**Stability**				
	MARS	**CMARS**	**BCMARS-1**	**BCMARS-2**	**BCMARS-3**
MAE	0.7657±0.1848	0.7440±0.2383	0.7252±0.1939	0.7375±0.2870	0.8690*±0.1783**
MSE	0.5500 ±0.3710	0.5690±0.3400	0.5550±0.3450	0.6374±0.2174**	0.7616*±0.2852
R^2	0.6070±0.3680	0.4690±0.3940	0.5750±0.3640	0.6577*±0.3063**	0.6300±0.3650
PWI	0.9950*±0.0070	0.9940±0.0070	0.9940±0.0070	0.9950*±0.0060**	0.9940±0.0080
PRESS	0.0003±0.0005	0.0±0.0**	0.0100±0.0270	0.0020±0.0050	0.1000*±0.2733

*indicates better performance with respect to means; **indicates better performance with respect to spread

For stability;

- Random-X Resampling and Wild bootstrapping methods are more stable when compared to the other methods.
- Random-X is more stable with respect to R^2 and PWI; it has the most robust stability with respect to the same measures, and also has the most robust stability with respect to the MSE.
- Besides, Wild bootstrapping is more stable in terms of MAE, MSE and PRESS; it has the most robust stability with respect to the MAE measure only.
- CMARS has the most robust stability with respect to PRESS.

5.2. Comparison with respect to Sample Sizes

Table 3 presents the performance measures of the studied methods with respect to two sample size categories: small and medium. Depending on the results given in the table, following conclusions can be reached:

- Small training and testing data sets produce better models for all measures except PRESS compared to the medium training and testing data sets.
- All methods are more stable in small data sets with respect to R^2, PWI and PRESS.
- Wild bootstrapping is more stable in small data sets with respect to all measures except PWI.
- Fixed-X method produces the lowest MAE for small size training data sets, while MARS has the best value for this measure in testing samples.
- Fixed-X produces the lowest MSE value in both small and medium sized training samples. However, MARS is the best method for the MAE in small data while Random-X is the best one in medium size testing data.
- Fixed-X method is superior to other methods in terms of R^2 for small and medium size training data sets, while MARS is the best one for testing small and medium size data sets.
- MARS and CMARS are the best methods with respect to PWI measure in both types of testing data. Both methods also perform similar with respect to the same measure in training samples.
- Random-X Resampling is the best method for the PRESS measure in small testing samples while MARS is the best model for PRESS is medium training samples. On the other hand, Fixed-X gives the best result in small training samples with respect to the same measure.
- In terms of the complexity measure, MSE as well as the accuracy measures MAE and R^2, Wild bootstrapping and the Random-X are the most stable methods in small and medium size data sets.
- MARS and Wild bootstrapping methods are the most stable methods in small and medium size data sets with respect to PWI, respectively.
- Wild bootstrapping is the most stable method in both size data in terms of the PRESS measure.

Table 3. Averages of Performance Measures with Respect to Different Sample Sizes

Sample Size	Performance Measures	Training				
		MARS	**CMARS**	**BCMARS-1**	**BCMARS-2**	**BCMARS-3**
Small	**MAE**	0.2340	0.3570	0.1899*	0.2092	0.3410
	MSE	0.1773	0.6020	0.1158*	0.1387	0.3000
	R^2	0.8208	0.7840	0.8824*	0.8596	0.7350
	PWI	1.0000*	0.9970	0.9910	0.9910	0.9870
	PRESS	0.0170	144	0.0150*	0.0140	0.0340
Medium	**MAE**	0.4563	0.4498*	0.4769	0.4874	0.5090
	MSE	0.6257	0.6125	0.5469*	0.7630	0.8540
	R^2	0.3802	0.3978	0.4431*	0.3140	0.2908
	PWI	0.9888	0.9900*	0.9890	0.9940*	0.9830
	PRESS	0.0020*	0.2440	0.5080	2.6400	$27x10^6$
Sample Size	**Performance Measures**	**Testing**				
		MARS	**CMARS**	**BCMARS-1**	**BCMARS-2**	**BCMARS-3**
Small	**MAE**	0.3300*	0.5560	0.3440	0.7280	0.3790
	MSE	0.3520*	1.0010	0.3980	0.3670	0.3570
	R^2	0.7110*	0.5760	0.6770	0.6800	0.6500
	PWI	1.0000*	1.0000*	0.9910	0.9910	0.9920
	PRESS	23.200	122.70	35.000	18.650*	22.700
Medium	**MAE**	0.5849	0.6052	0.6518	0.5468*	0.6160
	MSE	5.7800	2.1500	2.3100	0.7658*	1.7880
	R^2	0.1853*	0.1497	0.1765	0.1817	0.1178
	PWI	0.9860*	0.9860*	0.9850	0.9860*	0.9830
	PRESS	918.00	860.00	968.00	215.00*	$2.9x10^6$
Sample Size	**Performance Measures**	**Stability**				
		MARS	**CMARS**	**BCMARS-1**	**BCMARS-2**	**BCMARS-3**
Small	**MAE**	0.2250	0.7300	0.7265	0.6110	0.9359*
	MSE	0.4980	0.5770	0.5750	0.5530	0.8835*
	R^2	0.7700	0.5960	0.7350	0.7510	0.7710*
	PWI	1.0000*	0.9970	0.9990	0.9990	0.9940
	PRESS	0.0007	0.0469	0.0189	0.0040	0.1660*
Medium	**MAE**	0.4431	0.7578	0.7236	0.8888*	0.8022
	MSE	0.5760	0.5620	0.4410	0.7049*	0.6150
	R^2	0.4450	0.3410	0.3830	0.5460*	0.4890
	PWI	0.9915	0.9900	0.9898	0.9920	0.9948*
	PRESS	0.0000	0.0003	0.0000	0.0012	0.0457*

*indicates better performance with respect to the corresponding measure and sample

5.3. Comparisons with respect to Scales

In Table 4, the performance measures of the studied methods with respect to two scale types; small and medium are presented. Depending on the results given in the table, following conclusions can be drawn:

- Medium scale training data sets produce better models for all methods with respect to all measures except PWI.

Table 4. Averages of Performance Measures with Respect to Different Scale

Scale	Performance Measures	Training				
		MARS	**CMARS**	**BCMARS-1**	**BCMARS-2**	**BCMARS-3**
Small	**MAE**	0.5229	0.4140*	0.4720	0.4910	0.6561
	MSE	0.4572	0.8992	0.3830*	0.4040	0.7728
	R^2	0.5483	0.4985	0.6139*	0.5928	0.4078
	PWI	0.9970	0.9924	0.9980*	0.9980*	0.9934
	PRESS	0.0214*	143.66	0.4320	2.1910	2.7000
Medium	**MAE**	0.1677	0.1773	0.1384*	0.1492	0.1940
	MSE	0.3417	0.3500	0.2260*	0.4450	0.3810
	R^2	0.6591	0.6630	0.7650*	0.6340	0.6170
	PWI	0.9913	0.9920*	0.9820	0.9870	0.9770
	PRESS	0.0017	0.0010*	0.0080	0.0300	0.0060
Sample Size	**Performance Measures**	**Testing**				
		MARS	**CMARS**	**BCMARS-1**	**BCMARS-2**	**BCMARS-3**
Small	**MAE**	0.6696	0.5445*	0.6747	0.6776	0.7130
	MSE	0.7327*	2.3959	0.7717	0.7443	0.8469
	R^2	0.3297	0.3377*	0.3240	0.3293	0.2721
	PWI	0.9964*	0.9901	0.9960	0.9960	0.9932
	PRESS	213.00	800.69	176.94	141.72*	$2.9x10^6$
Medium	**MAE**	0.2703	0.2790	0.2550*	0.6070	0.2820
	MSE	5.4630	1.7800	1.8600	0.3130*	1.2980
	R^2	0.5107	0.5040	0.6000	0.6020*	0.4960
	PWI	0.9892	0.9900*	0.9790	0.9810	0.9820
	PRESS	1012.8	714.00	714.00	66.800*	419.00
Sample Size	**Performance Measures**	**Stability**				
		MARS	**CMARS**	**BCMARS-1**	**BCMARS-2**	**BCMARS-3**
Small	**MAE**	0.5008	0.7801	0.7041	0.7300	0.9200*
	MSE	0.6515	0.5771	0.5183	0.5539	0.8378*
	R^2	0.6521*	0.3714	0.5540	0.5539	0.6277
	PWI	0.9984*	0.9930	0.9977	0.9979	0.9980
	PRESS	0.0003	0.0828*	0.0005	0.0013	0.0471
Medium	**MAE**	0.7666	0.7900	0.7505	0.7470	0.8100*
	MSE	0.3474	0.5920	0.3480	0.8040*	0.6850
	R^2	0.5628	0.5310	0.6000	0.7600*	0.6330
	PWI	0.9931*	0.9930*	0.9910	0.9930	0.9920
	PRESS	0.0004	0.0460*	0.0220	0.0040	0.1650

* indicates better performance with respect to the corresponding measure and scale

- Medium scale data sets produce more stable models for all methods for MAE, R^2 and PRESS. On the other hand, small scale data yield more stable models for MSE and PWI.
- In small scale training samples, CMARS produces similar results with Fixed-X and Random-X Resampling for MAE measure. However, in small scale testing samples, MARS, Fixed-X and Random-X yield similar values for the same accuracy measure.
- Fixed-X Resampling is the best method with respect to the complexity measure, MSE, in small and medium scale training data sets. However, MARS and Random-X are the best methods for the same measure in small and medium scale testing samples, respectively.

- The best model in terms of R^2 is yielded by Fixed-X in both scale training samples, while CMARS and Random-X produce the best for the same measure in small and medium scale testing data, respectively.
- Fixed-X and Random-X models are superior to others in terms of PWI in small scale training samples. In medium scaled training samples, however, CMARS produces the best value for the same measure. On the other hand, in testing samples, MARS and CMARS give the best models with respect to PWI in both small and medium scale.
- Random-X Resampling is the best model for PRESS in all testing samples. But, MARS and CMARS result in better PRESS values all training data.
- Wild bootstrap is superior to other methods with respect to the stability of MAE in all type data sets.
- MARS seems more stable in terms of PWI in type data sets.
- Wild bootstrapping is superior to other methods with respect to the stability of MSE, the complexity measure, in small scaled while Random-X Resampling is the best method in medium scaled data with respect to the stability of the same measure.
- MARS and Random-X are the most stable in R^2 in small scale data and medium scale data, respectively.
- CMARS is the most stable method with respect to the PRESS measure for both scales of data.

5.4. Evaluation of the Efficiencies

The elapsed time of each method for each data set are recorded on Pentium (R) Dual-Core CPU 2.80 GHz processor and 32-bit operating system Windows ® computer during the runs (Table 14). Depending on the results, following conclusions can be stated:

- Run times increases as sample size and scale increases.
- As expected, it takes the bootstrap methods considerably longer times to run than MARS and CMARS.

Table 5. Runtimes (in seconds) of Methods with respect to Size and Scale of Data Sets

		Scale	
		Small	**Medium**
Sample Size	**Small**	MARS: < 0.0800 sec.*	MARS: < 0.0800 sec.*
		CMARS: < 4.4666 sec.	CMARS: < 19.5269 sec.
		BCMARS-1: < 1,595 sec.	BCMARS-1: < 13,262 sec.
		BCMARS-2: < 1,578 sec.	BCMARS-2: < 18,537 sec.
		BCMARS-3: < 1,599 sec.	BCMARS-3: < 15,617 sec.
	Medium	MARS: < 0.0840 sec.*	MARS: < 0.0900 sec.*
		CMARS: < 18.2008 sec.	CMARS: < 21.6737 sec.
		BCMARS-1: < 15,958 sec.	BCMARS-1: < 18,664 sec.
		BCMARS-2: < 7,076 sec.	BCMARS-2: < 31,590 sec.
		BCMARS-3: < 8,374 sec.	BCMARS-3: < 16,753 sec.

*indicates better performance with respect to run times

- Three bootstrap regression methods have almost the same efficiencies in small size and small scale data sets. Note that run times of these methods increases almost ten times as much as the scale increases from small to medium.

- Random-X and Wild bootstrapping have similar efficiencies in medium size small scale data sets; Fixed-X runs twice as much to those of Random-X and Wild bootstrapping, whose run times increase almost five times as much as the sample size increases.
- Fixed-X and Wild bootstrapping have similar run times for medium size medium scale data sets while Random-X runs almost twice as much to that of Fixed-X and Wild bootstrapping.

5.5. Evaluation of the Precisions of the Model Parameters

In addition to performance measures of the models, the CIs and standard deviations of the parameters are calculated after bootstrapping. These values are compared with those values obtained from bootstrapping CMARS. The smaller the lengths of the CIs and the standard deviations, the more precise the parameter estimates are.

According to the results, following conclusions can be drawn:

In medium size and medium scale data:

- The length of CIs is larger in Wild bootstrapping than the ones obtained by Fixed-X Resampling. Thus, Fixed-X gives more precise parameter estimates.
- The standard deviations obtained by bootstrapping (STD(BS)) are smaller for Wild bootstrapping method than for Fixed-X Resampling.
- In general, both types of standard deviations are smaller than the ones obtained from CMARS.

In small size and medium scale data set:

- In fold 2, standard deviations of Wild bootstrapping are smaller compared to those of CMARS, while the STD (BS) are not. However, the lengths of CIs become narrower after bootstrapping.

In medium size and small scale data set:

- In general, the length of CIs of Random-X is smaller than CMARS. Thus, Random-X produces more precise parameter estimates.
- Random-X Resampling produces narrower CIs than Fixed-X. So, parameter estimates of Random-X are more precise.
- The standard deviations of parameters obtained by Random-X and Fixed-X are similar.

In small size and small scale data set:

- The lengths of CIs become narrower and standard deviations of the parameters become smaller after bootstrapping, thus, resulting in more precise parameter estimates.
- STD(BS) values obtained for Fixed-X Resampling are smaller than ones obtained from Random-X.

6. CONCLUSION AND FURTHER RESEARCH

In this study, three different bootstrap methods are applied to a nonparametric regression, called CMARS, which is an improved version of the backward step of the widely used method MARS. MARS has two-step algorithm to build a model: forward and backward. CMARS uses inputs obtained from the forward step of MARS, and then, by utilizing the CQP technique, it constructs the large model. Although CMARS overperforms MARS with respect to several criteria, it constructs models which are at least as complex as MARS (Weber et al., 2011).

In this study, it is aimed to reduce the complexity of CMARS models. To achieve this aim, bootstrapping regression methods, namely Fixed-X and Random-X Resampling, and Wild bootstrapping, are utilized by adopting an iterative approach to determine whether the parameters statistically contribute to the developed CMARS model or not. If there are any which do not contribute, they are removed from the model, and a new CMARS model is fitted to the data by only retaining the statistically significant parameters until none of them is found to be insignificant. The reason of using a computational method here is the lack of prior knowledge regarding the distributions of the model parameters.

The performances of the methods are empirically evaluated and compared with respect to several criteria by using four data sets which are selected in such a way that they can represent the small and medium sample size and scale categories. The criteria include accuracy (with MAE, R^2, PWI and PRESS measures), complexity (with the MSE measure), stability (by comparing the performances in training and test samples), robustness (by comparing the performances in different data sets), efficiency (using run times) and precision (by evaluating the length of CIs of parameters). All performance criteria are explained in Appendix A. In order to validate all models developed; three-fold CV approach is used. For this purpose, these data sets are divided into three parts (folds) and two of them are used for building (training) and the remaining one is used for testing.

Depending on the comparisons presented in the previous section, Section 5, one may conclude the followings:

- In general, BCMARS methods perform better than MARS and CMARS with respect to most of the measures, and also lead to development of robust models with respect to the same measures.
- Either one of the BCMARS methods yields models which are less complex than that of MARS and CMARS.
- In overall, Random-X Resampling or Wild bootstrapping produce more stable models with respect to most of the measures considered.
- Fixed-X method performs the best in small size training data in terms of most measures.
- Fixed-X also performs the best in medium size training data sets with respect to MSE and R^2.
- MARS and Random-X Resampling overperform in small and medium size test data sets, respectively.
- Wild bootstrapping and Random-X methods are more stable in small and medium size test data sets, respectively.
- Fixed-X is performing equally well on both scale of training data sets.
- Random-X performs best in medium scale data while MARS and CMARS perform best in small scale data.
- Random-X Resampling is more stable in medium scale data set.

- It is apparent that by decreasing the number of terms in the model by bootstrapping, the CIs become narrower compared to those of CMARS. Moreover, the standard errors of the parameters which obtained empirically decreases after bootstrapping. Thus, bootstrapping results in more precise parameter estimates.
- The main drawback of bootstrapping is its computational effort. Since it is heavily dependent on computers, it takes significantly more time than the other methods, MARS and CMARS.

In short, depending on the above conclusions, it may be suggested that Random-X Resampling method leads to more accurate and more precise and less complex models particularly for medium size and medium scale data. Nevertheless, it is the least efficient method among the others for this type of data set in terms of run time.

Future studies are planned in several directions. First, BCMARS methods are going to be applied on different data sets with small to large size and scale. Then, Repeated Analysis of Variance (RANOVA) will be applied to test whether there is statistically significant difference between the performances of methods. Besides, replicated CV is going to be used while validating the models. Then, after well-documented, the written MATLAB code will be issued as on open source to make it available for interested researchers.

REFERENCES

Aldrin, M. (2006). Improved Predictions Penalizing both Slope and Curvature in Additive Models. *Computational Statistics and Data Analysis*, 50 (2), 267–284.

Alp, O. S., Büyükbebeci, E., Iscanoglu Cekic, A., Yerlikaya-Özkurt, F., Taylan, P. and Weber, G.-W. (2011). CMARS and GAM & CQP - Modern Optimization Methods Applied to International Credit Default Prediction. *Journal of Computational and Applied Mathematics (JCAM)*, 235, 4639-4651.

Austin, P. (2008). Using the Bootstrap to Improve Estimation and Confidence Intervals for Regression Coefficients Selected using Backwards Variable Elimination. *Statistics in Medicine*, 27 (17), 3286–3300.

Batmaz, İ., Yerlikaya-Özkurt, F., Kartal-Koç, E., Köksal, G. and Weber, G. W. (2010). Evaluating the CMARS Performance for Modeling Nonlinearities. *Proceedings of the 3rd Global Conference on Power Control and Optimization, Gold Coast (Australia)*, 1239, 351-357.

Chernick, M. (2008). *Bootstrap Methods: A Guide for Practitioners and Researchers.* New York: Wiley.

Cortez, P., and Morais., A. (2007). Data Mining Approach to Predict Forest Fires using Meteorological Data. In J. Neves, M. F. Santos and J. Machado (Ed.), New Trends in Artificial Intelligence, *Proceedings of the 13th EPIA 2007 - Portuguese Conference on Artificial Intelligence*, December, Guimarães, Portugal, 512-523.

Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap Methods and their Application.* Cambridge: Cambridge Series in Statistical and Probabilistic Mathematics.

Deconinck, E., Zhang, M. H., Petitet, F., Dubus, E., Ijjaali, I., Coomans, D., and Vander Heyden, Y. (2008). Boosted Regression Trees, Multivariate Adaptive Regression Splines and Their Two-Step Combinations with Multiple Linear Regression or Partial Least Squares to Predict Blood-Brain Barrier Passage: A case study. *Analytica Chimica Acta*, 609 (1), 13–23.

Denison, D. G. T., Mallick, B. K., and Smith, F. M. (1998). Bayesian MARS. *Statistics and Computing*, 8 (4), 337-346.

Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, 7 (1), 1-26.

Efron, B. (1979). Computers and the Theory of Statistics: Thinking the Unthinkable. *SIAM Review*, 21 (4), 460-479.

Efron, B. (1988). Computer-Intensive Methods in Statistical Regression. *Society for Industrial and Applied Mathematics*, 30 (3), 421-449.

Efron, B. (1992). Jackknife-After-Bootstrap Standard Errors and Influence Functions. Journal *of the Royal Statistical Society*. Series B (Methodological), 54 (1), 83-127.

Efron, B., and Tibshirani, R.J. (1986). Bootstrap methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy. *Statistical Science*, 1 (1), 75-77.

Efron, B., and Tibshirani, R.J. (1991). Statistical Data Analysis in the Computer Age. *Science*, 253, 390-395.

Efron, B., and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap.* New York: Chapman & Hall.

Flachaire, E. (2003). *Bootstrapping Heteroskedastic Regression Models: Wild Bootstrap vs. Pairs Bootstrap*. Working paper.

Fox, J. (2002). Bootstrapping Regression Models. *An R and S-PLUS Companion to Applied Regression:Web Appendix to the Book.* Sage, CA: Thousand Oaks.

Freedman, D.A. (1981). Bootstrapping Regression Models. *The Annals of Statistics*, 9 (6), 1218-1228.

Friedman J. (1991). Multivariate Adaptive Regression Splines. *Annals of Statistics*, 19 (1), 1-67.

Gentle, J. E. (2009). *Computational Statistics*. New York: Springer.

Givens, G. H., and Hoeting, J. A. (2005). *Computational Statistics*. New York: John Wiley & Sons.

Godfrey, L. (2009). *Bootstrap Tests for Regression Models*. Palgrave Macmillian.

Gonçalves, S., White, H., (2005). Bootstrap Standard Error Estimates for Linear Regression. *American Statistical Association*, 100 (471), 970-979.

Hastie, T., Tibshirani, and R., Friedman, J., (2001). *The Elements of Statistical Learning, Data Mining, Inference and Prediction.* New York: Springer.

Hjorth, J. S. U., (1994). *Computer Intensive Statistical Methods: Validation Model Selection and Bootstrap.* New York: Chapman & Hall.

Kartal, E. (2007). *Metamodeling Complex Systems Using Linear and Nonlinear Regression Methods.* Master Thesis, Graduate School of Natural and Applied Sciences, Department of Statistics, METU, Ankara, Turkey.

Kriner, M. (2007). *Survival Analysis with Multivariate Adaptive Regression Splines.* Dissertation, LMU Munchen: Faculty of Mathematics, Computer Science and Statistics, Munchen.

Lin, C. J., Chen, H. F., and Lee, T. S. (2011). Forecasting Tourism Demand Using Time Series, Artificial Neural Networks and Multivariate Adaptive Regression Splines: Evidence from Taiwan. *International Journal of Business Administration,* 2 (2), 14-24.

Liu, R. Y. (1988). Bootstrap Procedure under Some non-i.i.d. Models. *Annals of Statistics*, 16 (4), 1696-1708.

Loughin, T. M., and Koehler, K. J. (1997). Bootstrapping Regression Parameters in Multivariate Survival Analysis. *Lifetime Data Analysis*, 3(2), 157–177.

Martinez, W. L., and Martinez, A. R. (2002). *Computational Statistics Handbook with Matlab.* New York: Chapman & Hall.

MATLAB Version 7.8.0 (R2009a).

Matlab-R link Retrieved from http://www.mathworks.com/matlabcentral/fileexchange/5051 (accessed on February 24, 2011).

Milborrow, S. (2009). earth: Multivariate Adaptive Regression Spline Models. R Software Package. Retrieved from http://cran.r-project.org/web/packages/earth/ index.html (accessed on February 24, 2011).

Montgomery, D. C., Peck, E. A., and Vining, G. G. (2001). *Introduction to Linear Regression Analysis.* New York: John Wiley & Sons.

MOSEK, Version 6. A very powerful commercial software for CQP. Retrieved from http://www.mosek.com (accessed January 7, 2011).

R Development Core Team (2009). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from http://www.R-project.org. (accessed on February 24, 2011).

Ramanathan, R. (2002). *Introductory Econometrics with Applications*. South Western College Publishing.

Salibian-Barrera, M., and Zamar, R. Z. (2002). Bootstrapping Robust Estimates of Regression. *The Annals of Statistics*, 30 (2), 556-582.

Taylan, P., Weber, G.-W. and Yerlikaya, F. (2010). A new approach to multivariate adaptive regression spline by using Tikhonov regularization and continuous optimization, *TOP (the Operational Research journal of SEIO (Spanish Statistics and Operations Research Society)*, 18 (2), 377-395.

Weber, G. W., Batmaz, I., Koksal, G., Taylan, P., and Yerlikaya-Ozkurt, F. (2011). CMARS: A New Contribution to Nonparametric Regression with Multivariate Adaptive Regression Splines Supported by Continuous Optimisation, *Inverse Problems in Science and Engineering* (in print).

Wegman, E., (1988). Computational Statistics: A New Agenda for Statistical Theory and Practice. *Journal of the Washington Academy of Sciences*, 78, 310-322.

Wu, C. F. J. (1986). Jackknife, Bootstrap and Other Resampling Methods in Regression Analysis. *The Annals of Statistics*, 14 (4), 1261-1295.

Yeh, I-Cheng, (2007). Modeling Slump Flow of Concrete using Second-order Regressions and Artificial Neural Networks. *Cement and Concrete Composites*, 29 (6), 474-480.

Yerlikaya, F. (2008). *A New Contribution to Nonlinear Robust Regression and Classification with MARS and its Applications to Data Mining for Quality Control in Manufacturing*. Master Thesis, Graduate School of Applied Mathematics, Department of Scientific Computing, METU, Ankara, Turkey.

Yetere-Kurşun, A., Batmaz, İ. (2010). Comparison of Regression Methods by Employing Bootstrapping Methods. *COMPSTAT2010: 19th International Conference on Computational Statistics.* Paris, France. August 22-27. Book of Abstracts, 92.

York, T. P., Eaves, L. J., and Van Den Oord, E., J., C., G. (2006). Multivariate Adaptive Regression Splines: A Powerful Method for Detecting Disease-Risk Relationship Differences among Subgroups. *Statistics in Medicine*, 25 (8), 1355–1367.

Zakeri, I. F., Adolph, A. L., Puyau, M., R., Vohra, F. A., Butte, N. F. (2010). Multivariate Adaptive Regression Splines Models for the Prediction of Energy Expenditure in Children and Adolescents. *Journal of Applied Physchology*, 108, 128–136.

APPENDIX

Nomencleature:

y_i is the response value for the i^{th} observation,

$\hat{y}_i$ is the estimated response value for the i^{th} observation,

$\bar{y}$ is the value of the mean response,

n is the number of observations (sample size),

p is the number of terms (BFs) in the model,

$\bar{\hat{y}}$ is the value of the mean of the estimated responses,

$s(y)^2$ is the sample variance of the observed response values,

$s(\hat{y})^2$ is the sample variance of the estimated response values,

$e_i = y_i - \hat{y}_i$ is the residual for the i^{th} observation,

h_i is the leverage value of the i^{th} observation. It is obtained from the i^{th} diagonal element of the hat matrix; H. The hat matrix is defined with the following formula $H = X(X^T X)^{-1} X^T$. Here, X represents the design matrix and rank of it is p.

Accuracy Measures

Mean Absolute Error (MAE)

It is defined as follows:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|e_i| = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|. \quad (48)$$

Small values are the better.

The Coefficient of Determination (R^2)

This value shows how much variation in the response variable is explained by the model. It is defined by the following formula:

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}. \quad (49)$$

Higher values indicate better fit.

Proportion of Residuals within Some User-Specified Range (PWI)

PWI is the proportion of residuals within some user-specified range such as two or three sigma. In this study, three sigma coverage is considered. The greater the percentage is the better the performance.

Prediction Error Residual Sum of Squares (PRESS)

PRESS measures the predictive capability of the model. The formula used to calculate this measure is defined as:

$$PRESS = \sum_{i=1}^{n}\left(\frac{e_i}{1-h_i}\right)^2. \quad (50)$$

Small values of PRESS, indicates a higher ability of prediction.

Precision Measure

Bootstrap Estimate of Standard Deviation

The bootstrap estimate of standard error is calculated with the following formula (Martinez and Martinez, 2002).

$$\hat{se}_B = \left\{ \frac{1}{B-1} \sum_{i=1}^{B} \left(\hat{\theta}^{*b} - \bar{\hat{\theta}}^{*} \right)^2 \right\}^{1/2}, \tag{51}$$

where

$$\bar{\hat{\theta}}^{*} = \frac{1}{B} \sum_{b=1}^{B} \hat{\theta}^{*b} \tag{52}$$

and $\hat{\theta}^{*b}$ is the bootstrap replication of $\hat{\theta}$.

It measures the variation around the mean. The standard deviations of the parameters from ECDF are obtained.

Complexity Measure

Mean Square Error (MSE)

In this study, the MSE is used to measure the model complexity. Larger values of the MSE indicate more complex models. The formula for the MSE is given below:

$$MSE = \frac{1}{n-p} \sum_{i=1}^{n} \left(y_i - \hat{y}_i \right)^2 \tag{53}$$

Stability Measure

The model is said to be stable if it performs well on both training and testing data sets. It is measured by the following formula (Osei-Bryson, 2004):

$$\min \left\{ \frac{CR_{TR}}{CR_{TE}}, \frac{CR_{TE}}{CR_{TR}} \right\} \tag{54}$$

CR_{TR} and CR_{TE} represents the performance measures obtained from training and testing samples. If the stability measure is close to one, it indicates higher stability.