Sumerianz Journal of Scientific Research, 2023, Vol. 6, No. 1, pp. 9-13
ISSN(e): 2617-6955, ISSN(p): 2617-765X
Website: https://www.sumerianz.com
DOI: https://doi.org/10.47752/sjsr.61.9.13

Original Article Open Access

Extreme Bound Analysis Based on Correlation Coefficient for Optimal Regression Model

Loc Nguyen
Loc Nguyen's Academic Network, Vietnam
Email: ng_phloc@yahoo.com

Article History
Received: 9 December 2022
Revised: 19 February 2023
Accepted: 24 February 2023
Published: 26 February 2023

How to Cite
Loc, Nguyen., 2023. "Extreme Bound Analysis Based on Correlation Coefficient for Optimal Regression Model." *Sumerianz Journal of Scientific Research*, vol. 6, pp. 9–13.

Abstract
Regression analysis is an important tool in statistical analysis, in which there is a demand of discovering essential independent variables among many other ones, especially in case that there is a huge number of random variables. Extreme bound analysis is a powerful approach to extract such important variables called robust regressors. In this research, I propose a so-called Regressive Expectation Maximization with RObust regressors (REMRO) algorithm as an alternative method beside other probabilistic methods for analyzing robust variables. By the different ideology from other probabilistic methods, REMRO searches for robust regressors forming optimal regression model and sorts them according to descending ordering given their fitness values determined by two proposed concepts of local correlation and global correlation. Local correlation represents sufficient explanatories to possible regressive models and global correlation reflects independence level and stand-alone capacity of regressors. Moreover, REMRO can resist incomplete data because it applies Regressive Expectation Maximization (REM) algorithm into filling missing values by estimated values based on ideology of expectation maximization (EM) algorithm. From experimental results, REMRO is more accurate for modeling numeric regressors than traditional probabilistic methods like Sala-I-Martin method but REMRO cannot be applied in case of nonnumeric regression model yet in this research.
Keywords: Extreme bound analysis; Regression analysis; Correlation coefficient; Expectation maximization (EM) algorithm.

1. Introduction
Given an dependent random variable Z and a set of independent random variables $X = (1, X_1, X_2,\ldots, X_n)^T$, regression analysis aims to build up a regression function $Z = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \ldots + \alpha_n X_n$ called regression model from sample data $(\boldsymbol{X}, \mathbf{z})$ of size N. As a convention, X_j (s) are called regressors and Z is called responsor whereas $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \alpha_2,\ldots, \alpha_n)^T$ are called regressive coefficients. The sample $(\boldsymbol{X}, \mathbf{z})$ is in form of data matrix as follows:

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{x}_1^T \\ \boldsymbol{x}_2^T \\ \vdots \\ \boldsymbol{x}_N^T \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1n} \\ 1 & x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N1} & x_{N2} & \cdots & x_{Nn} \end{pmatrix}$$

$$\boldsymbol{x}_i = \begin{pmatrix} 1 \\ x_{i1} \\ x_{i2} \\ \vdots \\ x_{in} \end{pmatrix}, \boldsymbol{x}_j = \begin{pmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{Nj} \end{pmatrix}$$

$$\boldsymbol{z} = \begin{pmatrix} z_1 \\ z_2 \\ \vdots \\ z_N \end{pmatrix}, \boldsymbol{Z} = \begin{pmatrix} 1 & z_1 \\ 1 & z_2 \\ \vdots & \vdots \\ 1 & z_N \end{pmatrix}$$

Therefore, x_{ij} and z_i is the i^{th} instances of regressor X_j and responsor Z at the i^{th} row of matrix $(\boldsymbol{X}, \mathbf{z})$. Because the sample $(\boldsymbol{X}, \mathbf{z})$ can be incomplete in this research, $\boldsymbol{X}$ and $\mathbf{z}$ can have missing values and so, let z_i^- and x_{ij}^- denote missing values of responsor Z and regressor X_j at the i^{th} row of matrix $(\boldsymbol{X}, \mathbf{z})$. When both responsor and regressors are random variables, the assumption of their normal distribution is specified by the probability density function (PDF) of Z as follows:

$$P(Z|X,\boldsymbol{\alpha}) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(Z-\boldsymbol{\alpha}^T X)^2}{2\sigma^2}\right) \quad (1.2)$$

Note, $\boldsymbol{\alpha}^T X$ and σ^2 are mean and variance of Z with regard to $P(Z \mid X, \boldsymbol{\alpha})$, respectively. The superscript "^{T}" denotes transposition operator in vector and matrix. The popular technique to build up regression model is least squares method which produces the same result to likelihood method based on the PDF of Z but the likelihood method can produce more results with estimation of the variance σ^2. The PDF $P(Z \mid X, \boldsymbol{\alpha})$ is essential to calculate likelihood function of given sample. Let $\hat{\boldsymbol{\alpha}} = (\hat{\alpha}_0, \hat{\alpha}_1, \hat{\alpha}_2, \ldots, \hat{\alpha}_n)^T$ be the estimates of regressive coefficients $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \alpha_2, \ldots, \alpha_n)^T$ resulted from least squares method or likelihood method, the estimate of responsor Z is easily calculated by regression function as follows:

$$\hat{Z} = \hat{\alpha}_0 + \sum_{j=1}^{n} \hat{\alpha}_j X_j = \hat{\boldsymbol{\alpha}}^T X \quad (1.3)$$

When there is a large number of random variables which consumes a lot of computing resources to produce regression model, there is a demand of discovering essential independent variables among many other ones. Extreme bound analysis (EBA) is a powerful approach to extract such important variables called robust regressors. Traditional EBA methods focus on taking advantages of probabilistic appropriateness of regressors. With concerning domain of EBA, let A, B, and C be free set, focus set, and doubtful set of regressors, respectively, the regression function k of regression model k is rewritten without loss of its meaning as follows:

$$Z(k) = \alpha_0 + \boldsymbol{\alpha}_A^T A + \alpha_k X_k + \boldsymbol{\alpha}_D^T D \quad (1.4)$$

Where D is a combination of regressors taken from doubtful set C without regressor X_k and consequently, $\boldsymbol{\alpha}_A$ and $\boldsymbol{\alpha}_D$ are regressive coefficients extracted from $\boldsymbol{\alpha}$ corresponding to free set A and combination set D, respectively. According to Levine, Renelt, and Leamer, suppose variance of each model k is σ_k^2, if 95% confidence interval of α_k as $[\alpha_k - 1.96\sigma_k^2, \alpha_k + 1.96\sigma_k^2]$ [1] is larger or smaller than 0 then, the regressor X_k is robust. Alternately, Sala-I-Martin estimated the mean $\hat{\alpha}_k$ of α_k weighted by K likelihood values over K models where K is the number of combinations taken from doubtful set C. Later on, Sala-I-Martin calculated every fitness value of every regressor X_k and such fitness value is represented by cumulative density function (cdf) at 0 denoted cdf(0) given mean $\hat{\alpha}_k$ and model variance σ_k^2. The larger the cdf(0) is, the more robust the regressor is. In general, these probabilistic methods are effective enough to apply into any data types of regressors and responsor although they may not evaluate exactly the regressors which are independent from any models because probabilistic analysis inside these methods is required concrete regression models which are already built. Therefore, in this research, I propose an alternative method based on correlation beside these probabilistic methods for analyzing robust variables, in which highly independent regressors are concerned more than ever. The proposed algorithm is described in the next section.

2. Methodology

In this section, I describe a proposed EBA method based on correlation coefficient for optimal regression model. Essentially, I propose two concepts of correlation such as local correlation and global correlation. Local correlation is also called model correlation, which implies fitness of a target regressive parameter with subject to a given regression model. Note, regressive parameter $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \alpha_2, \ldots, \alpha_n)^T$ is the set of regressive coefficients corresponding to regressors $X = (X_1, X_2, \ldots, X_n)$ and let Z and $\hat{Z}$ be the responsor and its estimate, respectively. Given regression model k, let $R_k(X_j, \hat{Z})$ and $R_k(\hat{Z}, Z)$ be the correlation between X_j and $\hat{Z}$ and the correlation between $\hat{Z}$ and Z within model k, respectively.

$$R_k(X_j,\hat{Z}) = \frac{\sum_{i=1}^{N}(x_{ij}-\bar{x}_j)(z_i-\bar{z})}{\sqrt{\sum_{i=1}^{N}(x_{ij}-\bar{x}_j)^2}\sqrt{\sum_{i=1}^{N}(z_i-\bar{z})^2}}$$
$$R_k(\hat{Z},Z) = \frac{\sum_{i=1}^{N}(\hat{z}_i-\bar{\hat{z}})(z_i-\bar{z})}{\sqrt{\sum_{i=1}^{N}(\hat{z}_i-\bar{\hat{z}})^2}\sqrt{\sum_{i=1}^{N}(z_i-\bar{z})^2}} \quad (2.1)$$

Where,

$$\bar{x}_j = \frac{1}{N}\sum_{i=1}^{N} x_{ij}\,, \bar{z} = \frac{1}{N}\sum_{i=1}^{N} z_i\,, \bar{\hat{z}} = \frac{1}{N}\sum_{i=1}^{N} \hat{z}_i$$
$$\hat{z}_i = \boldsymbol{\alpha}^T \boldsymbol{x}_i = \alpha_0 + \sum_{j=1}^{n} \alpha_j x_{ij}$$

Suppose the estimate of the j^{th} coefficient α_j with regard to regressor X_j is $\hat{\alpha}_j$, let $R_k(X_j, Z)$ be the local correlation of X_j and Z within model k. Obviously, $R_k(X_j, Z)$ reflects fitness or appropriateness of the regressive coefficient estimate $\hat{\alpha}_j$ regarding model k. The local correlation $R_k(X_j, Z)$ is defined as product of $R_k(X_j, \hat{Z})$ and $R_k(\hat{Z}, Z)$ as follows:

$$R_k(X_j, Z) = R_k(X_j, \hat{Z})R_k(\hat{Z}, Z) \tag{2.2}$$

Indeed, local correlation is a conditional correlation of a regressor along its estimated coefficient given the condition which is the estimated regression model and so, the intermediate variable representing such condition is the estimated response $\hat{Z}$. For K models which are estimated, averaged local correlation $\bar{R}(X_j, Z)$ is calculated as follows:

$$\bar{R}(X_j, Z) = \frac{1}{K}\sum_{k=1}^{K} R_k(X_j, Z) \tag{2.3}$$

Global correlation implies fitness of the target regressive parameter without concerning any regression models. Let $R(X_j, Z)$ denote the global correlation between regressor X_j and responsor Z, which is defined as usual correlation coefficient as follows:

$$R(X_j, Z) = \frac{\sum_{i=1}^{N}(x_{ij} - \bar{x}_j)(z_i - \bar{z})}{\sqrt{\sum_{i=1}^{N}(x_{ij} - \bar{x}_j)^2}\sqrt{\sum_{i=1}^{N}(z_i - \bar{z})^2}} \tag{2.4}$$

A regressor X_j along with its implicit regressive coefficient α_j are good if they can give sufficient explanatories to possible models and they can be more independent to reflect the responsor Z. In other words, the first condition of sufficient explanatories to possible models is represented by local correlation and the second condition of independent reflection is represented by global correlation. Therefore, the fitness of X_j and α_j are defined as product of the averaged local correlation $\bar{R}(X_j, Z)$ and the global correlation $R(X_j, Z)$ follows:

$$\varphi_j = \bar{R}(X_j, Z)R(X_j, Z) \tag{2.5}$$

The larger the fitness φ_j is, the better the implicit estimate $\hat{\alpha}_j$ is, and the better the regressor X_j is. Good regressors X_j (also α_j or $\hat{\alpha}_j$) which have large enough fitness values φ_j are called robust regressors. Consequently, Regressive Expectation Maximization with RObust regressors (REMRO) algorithm searches for robust regressors and sorts them according to descending ordering with their fitness values φ_j as searching criterion. Another problem is how to produce K models to calculate the averaged local correlation $\bar{R}_k(X_j, Z)$. Fortunately, Sala-I-Martin [2] generated a set of K combinations of doubtful regressors which need to be checked their fitness. Each model in K models is estimated with each combination of doubtful ones and estimation method can be least squares method as usual. Moreover, REMRO can resist incomplete data because it applies Regressive Expectation Maximization (REM) algorithm into filling missing values for both regressors and responsor by estimated values based on ideology of expectation maximization (EM) algorithm. Let free set A be the set of regressors which is compulsorily included in the regression model and let focus set $B = X\backslash A$ be the complement of A with subject to the entire set X. Let d be the number of regressors in each combination set D_k taken from doubtful set $C = B\backslash\{X_j\}$ where X_j is current focused regressor, the following is flow chart of REMRO algorithm.

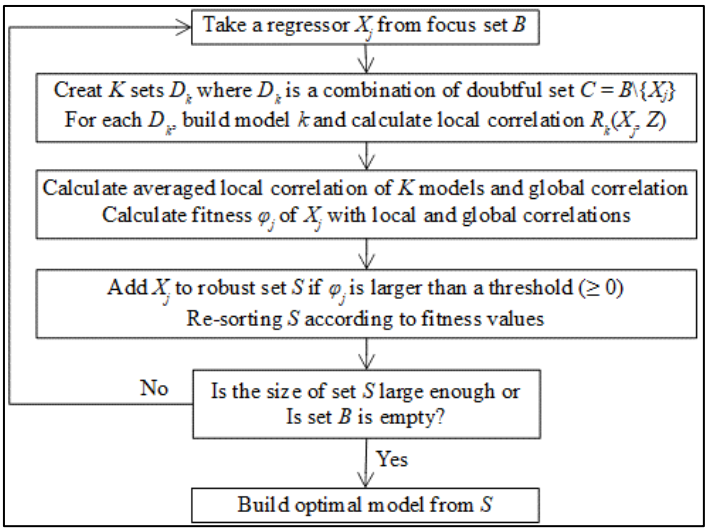

Figure-2.1. Flow chart of REMRO

Indeed, REMRO estimates fitness values of focused regressors in B and then builds up regression model with high fitness regressors. The final regression model estimated by REMRO with only robust regressors is called optimal regression model. Each combination suggested in some literature includes three doubtful regressors, $d = 3$.

Because the exhausted number of combinations will get huge as $2^{|C|}-1$ if d is browsed from 1 to the cardinality $|C|$ of doubtful set, I suggest the size d of each combination is half the cardinality of doubtful set C and hence, the number of models is determined as follows:

$$d = \left\lfloor \frac{|C|}{2} \right\rfloor$$
$$K = \frac{|C|!}{(c!)^2} \tag{2.6}$$

Note, the notation $\lfloor \rfloor$ represents lower integer of given real number. The accuracy of fitness computation is decreased when the number of target models is limited by such new d but this reduction will make REMRO faster and its decrease in accuracy will be alleviated by the global correlation $R(X_j, Z)$ which does not concern any model.

Sala-I-Martin [2] estimated the fitness of estimate $\hat{\alpha}_j$ as the value of cumulative density function of α_j at 0, denoted as $\text{cdf}(\alpha_j = 0 \mid \bar{\alpha}_j, \sigma_{\sigma_j}^2)$ followed by calculating the mean $\bar{\alpha}_j$ and the variance $\sigma_{\sigma_j}^2$ of α_j based on likelihood function over K models.

$$\varphi_j = \text{cdf}\left(0 \middle| \bar{\alpha}_j, \sigma_{\sigma_j}^2\right) \tag{2.7}$$

Especially, Sala-I-Martin mentioned the variance $\sigma_{\sigma_j}^2$ as averaged variance of K models. When REMRO is tested with Sala-I-Martin method, I improve Sala-I-Martin formulation by estimating $\sigma_{\sigma_j}^2$ only based on K distributed values of $\hat{\alpha}_j$ because the averaged variance of K models does not reflect variation of regressors. For instance, give K models and suppose each estimate of α_j within model k is $\hat{\alpha}_j(k)$, the variance $\sigma_{\sigma_j}^2$ is calculated as follows:

$$\sigma_{\sigma_j}^2 = \frac{\sum_{k=1}^{K}\left(\hat{\alpha}_j(k) - \bar{\alpha}_j\right)^2 L_k}{\sum_{k=1}^{K} L_k} \tag{2.8}$$

Where L_k is likelihood function of model k with assumption that regressor instances are also mutually independent random variables, as follows:

$$L_k = \prod_{i=1}^{N} P_k(\boldsymbol{z}_i | \boldsymbol{x}_i, \boldsymbol{\alpha}_k)$$

Where $P_k(\mathbf{z}_i \mid \boldsymbol{x}_i, \boldsymbol{\alpha}_k)$ is the PDF of $\mathbf{z}_i$ given model k:

$$P_k(\boldsymbol{z}_i | \boldsymbol{x}_i, \boldsymbol{\alpha}_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left(-\frac{(\boldsymbol{z}_i - \boldsymbol{\alpha}_k^T \boldsymbol{x}_i)^2}{2\sigma_k^2}\right)$$

The variance σ_k^2 of model k is estimated as follows:

$$\sigma_k^2 = \hat{\sigma}_k^2 = \frac{1}{N}\sum_{i=1}^{N}\left(z_i - \hat{z}_i(k)\right)^2$$

Where $\hat{z}_i(k)$ is the estimate of z_i with model k. The mean $\bar{\alpha}_j$ is still followed Sala-I-Martin formulation[2].

$$\bar{\alpha}_j = \frac{\sum_{k=1}^{K} \hat{\alpha}_j(k) L_k}{\sum_{k=1}^{K} L_k} \tag{2.9}$$

According to formulation of $\sigma_{\sigma_j}^2$ here, when $\bar{\alpha}_j$ is a mean with likelihood distribution, the variance $\sigma_{\sigma_j}^2$ is estimated with likelihood distribution too, which is slightly different from sample mean and sample variance as usual. In practice, L_k is replaced by logarithm of likelihood function $l_k = \log(L_k)$ in order to prevent producing very small number due to large matrix data with many rows.

REMRO applies REM algorithm into computing regressive estimates $\hat{\boldsymbol{\alpha}} = (\hat{\alpha}_0, \hat{\alpha}_1, \hat{\alpha}_2, \ldots, \hat{\alpha}_n)^T$ and REM, in turn, applies EM algorithm to resist missing values. It is necessary to describe shortly REM. REM [3] builds parallelly an entire regressive function and many partial inverse regressive functions so that missing values are estimated by both types of entire function and inverse functions. The model construction process of REM follows ideology of EM algorithm, especially EM loop but it is a bidirectional process. Recall that z_i^- and x_{ij}^- denote missing values of responsor Z and regressor X_j at the i^{th} row of matrix $(\boldsymbol{X}, \mathbf{z})$, which are estimated by REM as follows [1, 3]:

$$x_{ij}^- = \beta_{j0}^{(t)} + \beta_{j1}^{(t)} z_i^-$$
$$z_i^- = \frac{\sum_{j \in U_i} \alpha_j^{(t)} \beta_{j0}^{(t)} + \sum_{k \notin U_i} \alpha_k^{(t)} x_{ik}}{1 - \sum_{j \in U_i} \alpha_j^{(t)} \beta_{j1}^{(t)}} \tag{2.10}$$

Note, U_i is a set of indices of missing values x_{ij} with fixed i and β_{jk} (s) are regressive coefficients of partial inverse regressive functions. Although the ideology of REM is interesting, the pivot of this research is the association of local correlation and global correlation for computing fitness values of regressors. The source code of REM and REMRO is available at.

https://github.com/ngphloc/rem/tree/master/3_implementation/src/net/rem

3. Experimental Results and Discussions

In this experiment, REMRO is tested with Sala-I-Martin [2] given absolute mean error (MAE) as testing metric. MAE is absolute deviation between original response Z in matrix data and estimated response $\hat{Z}$ produced from regression model.

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|z_i - \hat{z}_i|$$

The smaller the MAE is, the better the method is. The traditional data "1974 Motor Trend" (mtcars) available in R data package [4] measuring fuel consumption based on technical parameters is tested dataset, in which responsor is the vehicle's miles per gallon (mpg) and 8 numeric regressors are number of cylinders (cyl), displacement in cubic inches (disp), gross horsepower (hp), rear axle ratio (drat), weight in thousands of pounds (wt), quarter-mile time in seconds (qsec), number of forward gears, and carburetors (carb). Only 4 robust regressors are extracted, which takes fifty percent of doubtful set. Table 3.1 shows the experimental results, in which second column lists sorted fitness values of robust regressors and third column shows optimal regression models whereas fourth column shows the evaluation metric MAE of REMRO method and Sala-I-Martin method.

Table-3.1. Evaluation of REMRO and Sala-I-Martin

Method	Fitness	Optimal model	MAE
REMRO	fit(cyl) = 0.7262 fit(disp) = 0.7200 fit(hp) = 0.6435 fit(wt) = 0.6133	mpg = 40.8285 – 1.2933*(cyl) + 0.0116*(disp) – 0.0205*(hp) – 3.8539*(wt)	1.771
Sala-I-Martin	fit(cyl) = 0.9913 fit(disp) = 0.7055 fit(hp) = 0.6908 fit(qsec) = 0.6545	mpg = 49.2352 – 1.6137*(cyl) – 0.0119*(disp) – 0.0288*(hp) – 0.6827*(qsec)	2.245

According to table 3.1, the robust regressors of REMRO and Sala-I-Martin method are (cyl, disp, hp, wt) and (cyl, disp, hp, qsec) along with sorted fitness values (0.7262, 0.7200, 0.6435, 0.6133) and (0.9913, 0.7055, 0.6908, 0.6545), respectively. Because MAE metric of REMRO as 1.771 is smaller than the one of Sala-I-Martin method as 2.245, REMRO is better than Sala-I-Martin method. Moreover, REMRO and Sala-I-Martin method share the three same regressors such as cyl, disp, and hp but their last robust regressors are different and hence, such difference makes REMRO better than Sala-I-Martin method in this test.

It is easy to deduce from experimental result, the strong point of REMRO is to appreciate the important level of strongly independent regressors from their global correlation when such regressors can explain well responsor without associating with other regressors. However, Sala-I-Martin method can work well in cases of binary data and multinomial data because the computing likelihoods for estimating fitness values does not depend directly on data types of regressors whereas arithmetic formulation of correlation coefficients requires strictly numerical regressors. Therefore, Sala-I-Martin method is more general than REMRO when it can be applied in many data types of regressors. Sala-I-Martin method can even be used for logit regression model because probabilistic applications are coherent aspects of such logistic model with note that likelihood function is essentially probability of random variable and prior/posterior functions are probabilities of parameter in Bayesian statistics.

4. Conclusions

From experimental results, REMRO is more accurate for modeling numeric regressors and responsor but it is not general and common like Sala-I-Martin method and other ones. In the future, I will try my best to improve REMRO by researching methods to approximate or replace numeric correlation by similar concepts within mixture of nonnumeric variables and numeric variables.

References

[1] Hlavac, M., 2016. "Extreme Bounds: Extreme Bounds Analysis in R. (B. Grün, T. Hothorn, R. Killick, and A. Zeileis, Eds.)" *Journal of Statistical Software, 72*(9), 1-22. doi:10.18637/jss.v072.i09

[2] Nguyen, L., and Ho, T.-H. T., 2018. "Fetal Weight Estimation in Case of Missing Data. (T. Schmutte, Ed.)." *Experimental Medicine (EM), 1*(2), 45-65. doi:10.31058/j.em.2018.12004

[3] Sala-I-Martin, X. X., 1997. "I Just Ran Two Million Regressions." *The American Economic Review, 87*(2), 178-183. http://www.jstor.org/stable/2950909