G.U. Journal of Science
22(3): 141-148 (2009)

GUJS
www.gujs.org

Genetic Algorithm Based Outlier Detection Using Bayesian Information Criterion in Multiple Regression Models Having Multicollinearity Problems

Özlem GÜRÜNLÜ ALMA [1♠], Serdar KURT [1], Aybars UĞUR [2]

[1] *Dokuz Eylül University, Faculty of Arts and Sciences, Department of Statistics, 35160, Izmir, Turkey*
[2] *Ege University, Faculty of Engineering, Department of Computer Engineering, 35040, Izmir, Turkey*

Received: 02.12.2008 Revised: 02.02.2009 Accepted: 02.02.2009

ABSTRACT

Multiple linear regression models are widely used applied statistical techniques and they are most useful devices for extracting and understanding the essential features of datasets. However, in multiple linear regression models problems arise when a serious outlier observation or multicollinearity present in the data. In regression however, the situation is somewhat more complex in the sense that some outlying points will have more influence on the regression than others. An important problem with outliers is that they can strongly influence the estimated model, especially when using least squares method. Nevertheless, outlier data are often the special points of interests in many practical situations. Another problem is multicollinearity in multiple linear regression (MLR) models, defined as linear dependencies among the independent variables. The purpose of this study is to define multicollinearity and outlier detection method using a Genetic Algorithm (GA) and Bayesian Information Criterion (BIC) in multiple regression models. Also, GA with BIC is to illustrate the algorithm with real and simulation data for outlier detection in MLR models having multicollinearity problems.

Key Words: Bayesian Information Criterion, Genetic Algorithms, Multicollinearity, Multiple Linear Regression, Outlier Detection.

1. INTRODUCTION

Regression is one of the most commonly used statistical techniques for understanding the essential features of datasets. However, there are a number of common difficulties associated with real datasets. The first involves detection and elimination of outliers in the original data. An outlier is one that appears to deviate so much from other observations of the sample [1, 4]. A problem with outliers in regression analysis is that they can strongly influence the regression model, especially when using least squares estimation criterion, so a multi step procedure is required, first to identify whether there are any samples that are atypical of the dataset, then to remove them, and finally to reformulate the model. Also, if there is a known distribution for the data, then using that distribution can aid in finding outliers [2, 8].

Influential outliers can bias parameter estimates and make the resulting analysis less useful. It is important to detect outliers since the outliers can provide misleading results. The classical identification method based on the sample mean or sample covariance matrix cannot always find them, because the classical mean and covariance matrix are themselves affected by outliers due to masking effects [1]. Several statistical estimates such as studentized residual, hat diagonal elements, Dffits, Cooks distance are available to identify both outliers and influential observations [7, 17, 20].

Statistical parametric methods for outlier detection either assume a known underlying distribution of the observations [4, 22] or, at least, they are based on statistical estimates of unknown distribution parameters. These methods flag as outliers those observations that deviate from the model assumptions. They are often

♠Corresponding author, e-mail: ozlem.gurunlu@deu.edu.tr

unsuitable for high dimensional data sets and for arbitrary data sets without prior knowledge of the underlying data distribution [21].

A second problem is that of correlations of between parameters in the model. The predictor variables in a regression model are considered orthogonal when they are not linearly related. But, when the regressors are nearly perfectly related, the regression coefficients tend to be unstable and the inferences based on the regression model can be misleading and erroneous, although the data may be predicted well. This condition is known as multicollinearity [19]. It is known that given strong multicollinearity the parameter estimates and hypotheses tests are affected more by the linear links between independent variables than by the regression model itself. The classical t-test of significance is highly inflated owing to the large variances of regression parameter estimates and the results of statistical analysis are often unacceptable [9].

In this study, we are interested with the problem of identifying outliers and detection of outliers in the dependent variable of MLR having multicollinearity problems using GA. GA has been used for outlier detection and model selection of linear regression models or times series. Also the use of GA for outlier detection and variable selection can be found in [24]. Ishibuchi and et al., (2001) proposed a genetic algorithm based approach for selecting a small number of instances from a given data set in a pattern classification problem. A robust simultaneous procedure is investigated for identification of outliers using Bayesian information criterion [16]. The scalability of information criterion is considered with a real data and also by generating simulation data. We have shown the behavior of our approach for constant sample sizes, two levels of correlations between independent variables and constant percentages of contaminated outliers by simulation. That is, the outliers were produced by adding a given amount to each dependent variable. We also studied on the affects of kappa coefficient which is the extra penalty value for Bayesian information criterion and we are obtained results for different values of it.

2. GA-BASED OUTLIER DETECTION IN MULTIPLE REGRESSION HAVING MULTICOLLINEARITY PROBLEMS

It is described detection of outliers in MLR models having multicollinearity problems based on GA and Bayesian information criterion in this section. These are as follows.

2.1. Outlier Detection in Multiple Regression

Regression analysis is to identify an appropriate transformation from sample to relate a response variable to a set of independent variables [10]:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k + \varepsilon \qquad (1)$$

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \dots + \hat{\beta}_k X_k \qquad (2)$$

where $Y \in \Re^n$ is a response variable, $\hat{Y}$ is the predicted value of the dependent variable, $X_1,\dots,X_k \in \Re^n$ are independent variables, β_0 is the intercept on the Y axis, and $\beta_1,\dots,\beta_k$ are the regression coefficients for each of the independent variables. The usual estimator of β coefficient $(\hat{\beta} = (X^T X)^{-1} X^T Y)$ comes from the method of Ordinary Least Squares (OLS) which is minimizes the difference between Y and $\hat{Y}$ values $\sum e^2 = \sum (Y - \hat{Y})^2$. The major disadvantage of OLS is performance when the error does not completely satisfy the classical assumptions. One of the most common violations of a normal distribution for the error terms is the presence of one or more outliers in the sample.

If outliers occur in the data, the errors can be thought to have a different distribution from normal. There are several possibilities, but perhaps the most intuitive one is the mixture model. We assume that the ε's in distinct cases are independent where,

$$\varepsilon \sim \begin{cases} N(0,\sigma^2) & (1-\pi) \\ N(0,K^2\sigma^2) & \pi \end{cases}$$

Here π is the probability of an outlier and K^2 is the variance inflation parameter. In practical works the data sets may have outliers. One outlying observation can destroy least squares estimation, resulting in parameter estimates that do not provide useful information for the majority of the data.

In this study, potential outliers can be incorporated into MLR model of equation (1) by the use of dummy variables. A dummy variable is N x 1 vector (N is the number of observations) that has a value of one for the outlier observation, and zero for all other observations. A dummy variable in this experimental study is equivalent to a detected outlier. The problem for outlier detection in MLR is to select of the best model. For this reason, the candidate MLR models have different combination of all possible dummy variables.

Bayesian information criterion (BIC) based on Bayesian method proposed by Schwarz [23]. The BIC will be used here for outlier detection. For MLR model with dummy variables the criterion can be calculated as,

$$BIC = \log(\hat{\sigma}^2) + m\log(N)/N \qquad (3)$$

where $\hat{\sigma}^2 = (e'e)/(N-k-1)$ is the estimated variance of regression model, and $m = 1 + k + m_d$, the total number of parameters in the estimated model, consists of parameters for the constant, the k independent variables and the number of outlier dummies m_d. Generally a good model has small residuals, and few parameters, then it is chosen with the smallest value of BIC is preferred [24].

A problem in using the BIC for outlier detection is that by itself it tends to include unnecessary outlier dummies. To circumvent this problem, a correction to the criterion is used. This takes the form of an extra penalty $(\kappa > 1)$ for the dummies. The corrected BIC is denoted BIC' which is given by [24],

$$BIC' = \log(\hat{\sigma}^2) + (1+k)\log(N)/N + \kappa m_d \log(N)/N, \quad (4)$$

where the kappa $(\kappa > 1)$ is the extra penalty given to outlier dummies. Simulation experiments are conducted to determine relevant different values of κ and true outlier detection.

2.2. Multicollinearity in Multiple Regression

In the applications of regression analysis, multicollinearity is a problem that always occurs when two or more predictor variables are correlated with each other. This problem can cause the value of the least squares estimated regression coefficients to be conditional upon the correlated predictor variables in the model.

The existence of co linearity in the linear regression model can lead to a very sensitive least squares estimate. This implies that different samples taken at the same X levels could lead widely different coefficients and variances of the predicted values will be highly inflated. Least squares estimates of β_i are usually too large in absolute values with wrong signs. Interpretation of the partial regression coefficient is difficult when the regressor variables are highly correlated. Multicollinearity in multiple linear regression can be detected by examining variance inflation factors (VIF) and condition indices (CI).

Several methods have been developed to overcome the defficiencies of multicollinearity. These are Partial least squares regression (PLSR), Principal component regression (PCR), and Ridge regression (RR) as methods to handle multicollinearity [7]. PLSR is a method of modeling relationships between a response variable and other explanatory variables [11]. This method was developed by Wold (1966) [25]. According to Barker (1997) [6], the PCR performs the Principal Component Analysis on the explanatory variables and then runs a regression using the principal component scores as the explanatory variables with the response variable of interest. Hoerl and Kennard (1970)[13] developed ridge regression which is the modifications of the least squares method that allow biased estimators of the regression coefficients. However, as many author noted, the influence of the observations on ridge regression is different from the corresponding least squares estimate, and collinearity can even disguise anomalous data investigated the leverage in ridge regression [5].

2.3. A Genetic Algorithm for Outlier Detection

Genetic algorithms are search technique used in computing to find true or approximate solutions to optimization and search problems which are a particular class of evolutionary algorithms that use techniques inspired by evolutionary biology such as inheritance, mutation, selection, and crossover [12].

GAs has been implemented as a computer simulation in which a population of abstract representations to an optimization problem evolves toward better solutions. Traditionally, solutions are represented in binary as strings of 0s and 1s, but other encodings are also possible. The evolution usually starts from a population of randomly generated individuals (chromosomes) and happens in generations. In each generation, the fitness of every individual in the population is evaluated, multiple individuals are stochastically selected from the current population based on their fitness, and modified to form a new population. The new population is then used in the next iteration of the algorithm. In summary, the outline of the steps of GA is shown in Figure 1.

In this experimental study, GA was used to detect the outliers. A random population of chromosomes was created representing the solution space. Each member of this random population represents a different possible solution for the GA. The GA contains the following components.

- Parameter Encoding: The coding of the candidate models for outlier detection is straightforward. Each model also called a chromosome, is fully described by a binary vector "d", $d = (d_1, ..., d_N)$, where $d_i = 0$ indicates no outlier dummy and $d_i = 1$ indicates an outlier dummy for observation i, for each $i = 1, ..., N$. These dummy variables for outlier observations must be created before the GA is run on the data set.

In this study, the structure of a chromosome is shown in Figure2. Each chromosome consists of p genes, where p is the number of outliers (o_p, $(p = 1, ..., N)$) given in a model. For instance p=3; the first, second and N-1st observations are outliers in Figure 2.

- Fitness Function: The measure of fitness of a chromosome is evaluated by the fitness function, which has as its argument the string representation of the chromosome and returns a value indicating its fitness. The genes, which represent the serial number of outliers, are updated with each new population created and the fitness of a chromosome is computed by the BIC' (4) for MLR model with the corresponding dummy variables.

[Begin]

[Initialise] Generate random population of N_c chromosomes. These are candidate solutions for the detection of outliers in Y.
[Evaluate] Evaluate the fitness of each chromosome in the population using BIC'.
[Repeat Until Termination Condition is Satisfied **Do**

1. **Select** two parents chromosomes from a population according to their fitness value BIC'.
2. **Crossover** with a crossover probability crossover the parents to form a new offspring.
3. **Mutation** with a mutation probability mutate new offspring at each locus.
4. **Evaluate** new candidate solutions.
5. **Select** chromosomes for the next generation.

[Replace] Use new generated population for a further run of algorithm and look for minimum of the BIC'.
[Test] If the final condition is satisfied based on BIC' stop, and return te best solution in current population.
[Loop] Go to evaluate.

Figure 1. The Outline of GA.

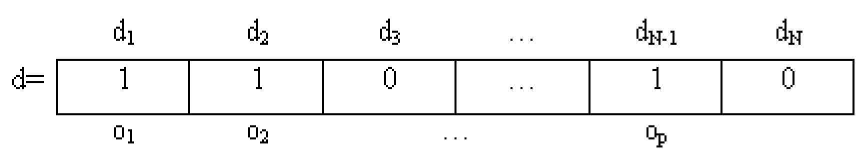

Figure 2. The Structure of a Chromosome.

- The Population and Generations: The population size in each generation is 40 chromosomes. MLR models corresponding to these chromosomes are then estimated using the observed data, and BIC' values for them computed. The chromoosmes with smallest values of the fitness function are more likely to pass their genes onto the next generation.

- Selection Operator: During selection operator, fitter individuals have a higher chance to be selected than less fit ones for next generation. Stochastic uniform selection function is used in GA. This function lays out a line in which each parent corresponds to a section of the line of length proportional to its scaled value.

- Crossover Operator: Crossover, the process whereby a new chromosome solution is created from the information contained within two parent solutions. The next generation of chromosomes from the previous one, is based on the BIC' values of the chromoosmes. The best chromosome has the smallest value of the fitness function BIC', are more likely to pass their genes onto the next generation. A crossover probability is selected as $p_c = 1$ and it is indicates that crossover always occurs between any two parent models chosen from the mating pool; thus the next generation will consist only of offspring models, not of any model from the previous generation.

- Mutation Operator: Mutation is applied to one candidate and results to build a new candidate chromosome Mating of the chromosomes from the previous one generation will not be enough for diversity of population. To this end, the chromosomes of each generation are also mutated before model estimation. Each gene of each individual is flipped, from zero to one or vice versa, with probability $p_m = 0.01$.

Executing crossover and mutation leads to a set of new candidates that compete based on their fitness value BIC' with the old ones for a place in the next generation. This process can be iterated until a candidate with sufficient a solution is found or a previously set computational limit is reached.

3. EXPERIMENTS AND DESIGN OF SIMULATION STUDY

In this study, the performance of BIC' information criterion to outlier detection is evaluated and performance of GA is demonstrated through real data and simulation experiments. These data sets demonstrated the effectiveness of our method. Data is generated for N=40 observations and number of outliers are inserted data set by taking into account of percentage of outliers in the dependent variable. Then this algorithm is applied for outlier detection in MLR having multicollinearity problems using synthetic data sets.

3.1. Experiments: Simultaneous Outlier Detection

In this study, two experimental data sets have been used to illustrate outlier detection in MLR modeling. References to these, and other information, including where to obtain the data can be found in [14][1]. In this section, it is investigated that detect outliers from these data sets with GA.

i. Scottish Hill Racing: The first example involves data supplied by Scottish Hill Runners Association. The purpose of the study is to investigate the relationship between record time of 35 hill races and two explanatory variables: distance is the total length of the race, measured in feet. One would expect that longer races and larger climbs would be associated with longer record times [3].

In this data set; races 7th and 18th observations are outliers. After removed observations 7 and 18, observation 33 is also an outlier. Thus observations 7 and 18 mask observation 33. After race numbers 7, 18, and 33 are removed from the data, standard diagnostic checking does not reveal any gross violations of the assumptions underlying MLR models [10].

The GA described earlier was run many times with this data; all runs result in the same outliers being detected, at observations 7, 18, and 33. The solution was always found quickly by the GA overcomes masking effects. The estimated model with the three outlier dummies has a BIC' value of 4.136 and the results of GA on the Matlab program as seen in Figure 3.

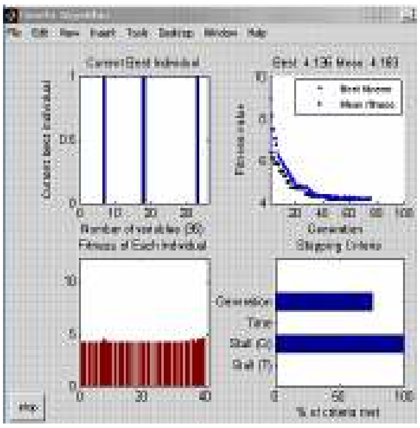

Figure 3. The Results of Scottish Hill Data by GA.

ii. The Stack Loss Data: The stack loss data consist of 21 days of operation form a plant for the oxidation of ammonia as a stage in the production of nitric acid. The response is called stack loss which is the percent of unconverted ammonia that escapes from the plant. There are three explanatory variables. The air flow is first independent variable which measures the rate of operation of the plant. The second independent variable measures the inlet temperature of cooling water circulating through coils in this tower and the last independent variable is proportional to the concentration of acid in the tower. Small values of the respond correspond to efficient absorption of the nitric oxides. In earlier research [3, 14] been identified as outliers four observations. These are 1, 3, 4, and 21 observations. This data set provides an interesting extreme example of masking [3]. The detection of any of these outliers is very difficult if only one observation at a time is examined. But the simultaneous methods such as GAs are able to detect all of four outliers at a time.

The GA was run a lot of times with this data. The entire run gives to result in the same outliers being detected, at observations 1, 3, 4, and 21. The best outlier combination was always found quickly by the GA. The estimated model with the four outlier dummies has a BIC' value 2.30 and the results of GA on the Matlab program as seen in Figure 4.

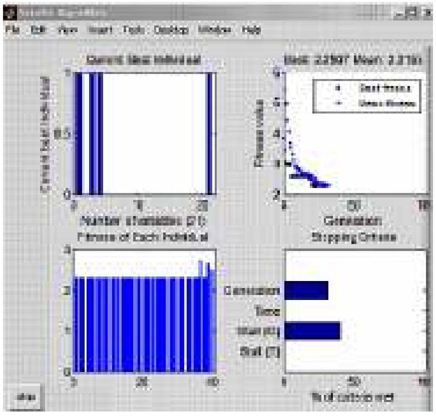

Figure 4. The Results of Stack Loss Data by GA.

3.1. Data Generation and Outlier Detection in MLR Having Multicollinearity Problems

In order to study the performance of the BIC' criterion and the role of κ values for in MLR having multicollinearity problems, it is conducted a simulation study. Following McDonald and Galarneau (1975) [18], it is explained how to generate a suitable design: firstly, it is generated the values for Y and X_1, X_2, X_3 using the following simulation protocol.

- the first three predictors are generated from;

$$X_1 = 2 + \varepsilon_1,$$

$$X_2 = 2 + 0.3\varepsilon_1 + \alpha\varepsilon_2$$

$$X_3 \sim \text{Normal}(3,1)$$

where $\varepsilon_1, \varepsilon_2$ are independent and identically distributed (i.i.d.) according to $N(0, \sigma^2 = 1)$. The parameter α controls the degree of collinearity between predictors X_1 and X_2 and it is determined

[1] These data sets are available from one of the author's website; http://www.stat.colostate.edu/~jah/index.html

α value for degree of correlations between predictors. Thus, two levels are selected for α values, these are 0.60 and 0.80. Then, it is generated the response variable Y_i from;

- $Y_i = 0.2 + X_{1i} + 0.5X_{2i} + 0.3X_{3i} + \varepsilon_i$,

  $\varepsilon_i \sim N(0, \sigma^2 = 1)$

  for i=1,…,40 observations.
- Percentage of outliers in the dependent variable is selected as %5. The outliers are generated from the uniform distribution which lies at least 3σ from the mean of Y_i and, the kappa values are selected as $\kappa = 2, 3, ..., 10.$

Under these conditions, firstly it is simulated the explanatory variables and the error terms for (i= 1,…,N) observations and N=40. Then, it is generated the response variable. After Y_i are generated from normal distribution, outlier observations are generated from uniform distribution take into account of percentage of outliers. For example, for the sample size N=40 and percentage of outlier for the %5, it can be generated 2 outliers observation.

Outliers are then added to the dependent variables. The iteration number for each combination of experiments is 100. The following table shows that the parameters of GA with BIC' as the fitness function for the simulated models. The best models chosen most of the generations of GA can detect the outliers.

Table 1. The Parameters of the GA for the Simulated Model.

Sample Size of Simulation Data, N=40	
Number of Generations	250
Population Size	40
Fitness Value	BIC'
Crossover Probability	1
Mutation Probability	0.01
Elitism	For two parents

4. RESULTS

GA mimics evolution is also a useful optimization tool for statistical modeling. In this study, it is demonstrated that Bayesian information criterion and a GA outline for outlier detection in MLR having multicollinearity problems. The value of information based selection criterion is calculated for observation as a measure of the fitness of dependent variable in MLR models using different values of κ . GA can simultaneously search in the solution space and find the outliers also for multicollinearity problems. The simulation results are shown in Table 2, where the values in cells are defined as extra total numbers of finding outliers in all iterations in dependent variable with GA.

As seen in Table 2 the true results for experiments are obtained for values of $\kappa = 3, 4, ..., 10$ for sample size is N= 40 and percentage of outliers %5. A simulation study is carried out to support the good behavior of the BIC'. It is clear that from simulation results for high values of kappa coefficient (κ >3) gives true information about how many observations are found as outlier. Therefore, we concluded that the best performing for outlier detection using BIC' in MLR models is taken by the kappa coefficient is bigger than three. The important issue is that the BIC' criterion can not be affected masking or swamping effects finding outliers so we also said that this criterion is robust than other outlier detection methods.

In Figure 5 it is seen that the kappa coefficient gives good results when the dependent variable Y containing of %5 outlier observation and MLR model having multicollinearity problems for degree of correlations 0.60 and 0.80 between two predictors.

Experiments with real and synthetic data sets show that the information criterion based on outlier detection method using GA in MLR models having multicollinearity problems find the outlier automatically.

The purpose of this experimental study was to test the scalability of the GA based on fitness function as BIC' criterion when MLR models having multicollinearity problems handling constant sample size and contaminated data with outliers.

We tested two types of scalability of the GA for outlier detection on data sets. The first one is the scalability of the GA against the given number of outliers in MLR models having multicollinearity problems and the second is the scalability against the power of different kappa coefficients for a given sample size and number of outliers. Figure 5 shows the results of using GA to find diversity number of outliers on data set. One important observation from this figure was that the GA based on information criterion can be found accurately outliers especially the kappa coefficient bigger than six for MLR models having multicollinearity problems handling.

Hence, we are confident to claim that the GA based on BIC' criterion is suitable for MLR models having multicollinearity problems.

Table 2. Total Number of Outliers Finding with GA.

Correlations Between X_1 and X_2	Total # of Outliers in Data	Kappa Values								
		2	3	4	5	6	7	8	9	10
0.60	200	193	22	5	1	0	0	0	0	0
0.80	200	168	13	3	1	0	0	0	0	0

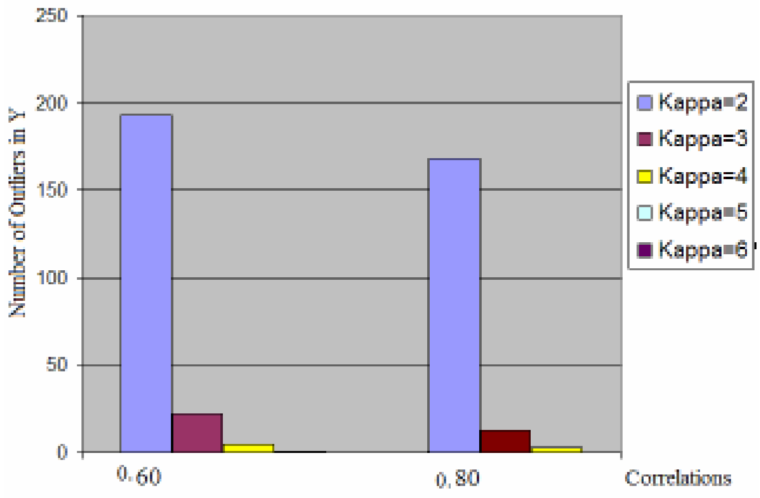

Figure 5. Results for Percentage of Outliers = 5% and Correlations 0.60, 0.80 Between X_1 and X_2.

REFERENCES

[1] Acuna, E., Rodriguez, C., "On Detection Of Outliers And Their Effect In Supervised Classification", http://academic.uprm.edu/~eacuna/vene31.pdf

[2] Amidan, B., Ferryman, T., Cooley, S., "Data Outlier Detection Using the Chebyshew Theorem, In: Aerospace", ***IEEE Aerospace Conference Proceedings***, IEEE, Piscataway NJ USA, 3814-3819 (2005).

[3] Atkinson, A.C., "Influential Observations, High Leverage Points, and Outliers in Linear Regression", ***Statistical Science***, 1: 397-402 (1986).

[4] Barnett, V., Lewis T., "Outliers in Statistical Data 3rd ed.", ***John Wiley and Sons***, USA (1994).

[5] Belsley, D.A., "Conditioning Diagnostics", ***Wiley***, New York (1991).

[6] Barker, M., "A Comparisons of Principal Component Regression and Partial Least Squares Regression", ***Multivariate Project*** (1997).

[7] Birkes and Dodge, "Alternative Methods of Regression", 3th ed., ***John Wiley & Sons***, Canada (1993).

[8] Davies, L., Gather, U., "The Identification of Multiple Outliers", ***Journal of the American Statistical Association***, 88(423): 797-801 (1993).

[9] Dempster, A.P., Schatzoff, M., Wermuth, N., "A Simulation Study of Alternatives to Ordinary Least Square", ***Journal of American Statistical Association***, 72: 77-91 (1977).

[10] Fox, J., "Applied Regression Analysis, Linear Models and Related Methods", 3th ed., ***Sage Publication***, USA (1997).

[11] Garthwaite, P.H., "An Interpretation of Partial Least Squares", ***Journal of the American Statistical Association***, 89: 122-127 (1994).

[12] Goldberg, D.E., "Genetic Algorithms in Search, Optimization, and Machine Learning", ***Addison-Wesley***, USA (1989).

[13] Hoerl, A.E., Kennard, R.W., "Ridge Regression: Biased Estiamtion to Nonorthogonal Problems", ***Technometrics***, 12: 56-67 (1970).

[14] Hoeting, J., Raftery, A.E., Madigan, D., "A Method for Simultaneous Variable Selection and Outlier Identification in Linear Regression", ***Computational Statistics and Data Analysis***, 22: 251-270 (1996).

[15] Ishibuchi, H., Nakashima, T., Nii, M., "Genetic Algorithm Based Instance and Feature Selection", In: Liu H, Motoda H, Instance Selection and Construction for Data Mining, ***Kluwer Academic***, (2001).

[16] Kullback, S., "Information Theory and Statistics", ***Dover Publications***, USA (1996).

[17] Neter, J., Wasserman, W., Kutner, M.H., "Applied Linear Regression Models", 2nd ed., ***Irwin Homewood IL*** (1989).

[18] McDonald, G.C., Galarneau, D.I., "A Monte Carlo Evaluation of Some Ridge Ttype Estimators", ***J. American Statistics***, 70: 407-416 (1975).

[19] Mason, R.L., Gunst, R.F., Webster, J.T., "Regression Analysis and Problem of Multicollinearity", ***Communication in Statistics***, 4(3): 277-292 (1975).

[20] Montgomery, D.C., Peck, E.A., "Introduction to Linear Regression Analysis", 2nd ed., ***John Wiley***, New York (1992).

[21] Papadimitriou, S., Kitawaga, H., Gibbons, P.G., and Faloutsos, C., "LOCI: Fast Outlier Detection Using the Local Correlation Integral", ***Intel research Laboratory Technical report*** **no. IRP-TR-02-09** (2002).

[22] Rousseeuw, P., Leory, A., "Robust Regression and Outlier Detection", ***Wiley Series in Probability and Statistics*** (1987).

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

[24] Tolvi, J., "Genetic Algorithms for Outlier Detection and Variable Selection in Linear Regression Models", Soft Computing, ***Springer***, 527-533 (2004).

[25] Wold, H., "Estimation of Principal Components and Related Models by Iterative Least Squares. In Multivariate Analysis", Ed. Krishnaiah, P.R., ***New York: Academic Pres***, 391-420 (1966).