MATEMATIKA, 2013, Volume 29, Number 2, 119–131

Performances Comparison of Information Criteria for Outlier Detection in Multiple Regression Models Having Multicollinearity Problems using Genetic Algorithms

Özlem Gürünlü Alma
Department of Statistics, Faculty of Science
Mugla University, Mugla, Turkey
e-mail: ozlem.gurunlu@hotmail.com

Abstract Multiple linear regression models are widely used in 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 multicollinearity or a serious outlier observation present in the data. Multicollinearity is a linear dependency between two or more explanatory variables in the regression models which can seriously affect the least squares estimated regression surface. The other important problem is outlier; 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. The purpose of this study is to performance comparison of Akaike Information Criterion (AIC'), Bayesian Information Criterion (BIC') and Information Complexity Criterion (ICOMP'(IFIM)) for detecting outliers using Genetic Algorithms when multiple regression model having multicollinearity problems.

Keywords Akaike Information Criterion; Bayesian Information Criterion; Information Complexity Criterion; Genetic Algorithms; multicollinearity; outlier detection.

2010 Mathematics Subject Classification 62J05

1 Introduction

Regression is one of the most commonly used statistical techniques for understanding the essential features of datasets. The purpose of regression analysis is to identify an appropriate model to relate a response variable to a set of independent variables, [1]:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \varepsilon, \tag{1}$$

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + ... + \hat{\beta}_p X_p, \tag{2}$$

where $Y \in \Re^n$ is a response variable, $\hat{Y}$ is the predicted value of the dependent variable, $X_1, \ldots, X_p \in \Re^n$ are independent variables, β_0 is the intercept on the Y axis, and $\beta_1, ..., \beta_p$ 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 minimizes the difference between Y and $\hat{Y}$values,

$$\sum e^2 = \sum \left(Y - \hat{Y}\right)^2.$$

The major disadvantage of OLS is when the error does not completely satisfy the classical assumptions, such as the presence of one or more outliers in the sample.

Outliers are defined as the observations or records which appear to be inconsistent with the remainder of the data in the group. A well quoted definition of outliers is given by Hawkins [2]. This definition described an outlier as an observation that deviates so much from other observations so as to arouse suspicion that it is generated by a different mechanism. Detected outliers are candidates for aberrant data that may otherwise adversely lead to model mispecification, biased parameter estimation and incorrect results. It is therefore important to identify them prior to modeling and analysis, especially if the data set contains more than one outlier, which is likely to be the case in most data sets. The problem of identifying such observations becomes more difficult because of the masking and swamping effects, [3,4]. Different ways to analyze the data with outliers have been suggested, using robust regression methods, by many statisticians [5–15]. So detection of outliers in regression is very important and should be studied more carefully.

A second problem is that of correlations 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, [16]. 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, [17].

Genetic Algorithms (GA) has been used for outlier detection and model selection of linear regression models or times series. A GA allowed simultaneous detection of outliers in data sets. Thus, this method is to overcome the problems of masking and swamping effects. The use of GA for outlier detection and variable selection can be found in Tolvi [18]. Ishibuchi *et al.* [19] 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 [20]. Gürünlü Alma *et al.* [27] derived AIC' and ICOMP'(IFIM) criteria for simultaneous outlier detection in multiple regression models using GA.

In this study, the focus is on the problem of detecting outliers in the dependent variable of multiple linear regression (MLR) model with multicollinearity using GA. AIC', BIC' and ICOMP'(IFIM) criteria have been used as the fitness function of genetic algorithms to detect outliers in MLR model. The scalability of information criterion is considered by generating simulation data. Simulation results of AIC', BIC' and ICOMP'(IFIM) criteria are obtained from for different number of sample sizes, and constant two levels percentages of contaminated outliers in the dependent variables. That is, the outliers are produced by adding a given amount to each dependent variable. The simulation results show that the performances of information criteria to accurately detect the outliers are affected by multicollinearirty problem in MLR.

2 GA Based Outlier Detection for MLR Models

Outlying observations can destroy least squares estimation, resulting in parameter estimates that do not provide useful information for the majority of the data. Belsley *et al.* [21]

described many of the well known outlier detection procedures for regression models. 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. It is assumed that the $\varepsilon's$ in distinct cases are independent and

$$\varepsilon \sim \begin{cases} N(0,\sigma^2), & (1-\pi), \\ N(0,k^2\sigma^2), & \pi, \end{cases} \tag{3}$$

where π is the probability of an outlier in data set and k^2 is the variance inflation parameter. To overcome non-normality, the detection of outliers are made possible by adding dummy variables to the regressor matrix of a regression model, [22]. Potential outliers can be incorporated into multiple regression models by the use of dummy variables, and this is what is done in this study. A dummy variable is $n \times 1$ (n, is the number of observations) vector that has a value of one for the outlying observation, and zero for all other observations. Each outlier would have a corresponding dummy variable and MLR can be written in a matrix form as in (4).

$$\begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & ... & x_{1p} & 1 & 0 & ... & 0 \\ 1 & x_{21} & ... & x_{2p} & 0 & 1 & ... & 0 \\ . & . & . & . & . & . & . & \\ . & . & . & . & . & . & . & \\ . & . & . & . & . & . & . & \\ 1 & x_{n1} & ... & x_{np} & 0 & 0 & ... & 1 \end{bmatrix} \tag{4}$$

A dummy variable in the regression model is therefore equivalent to the detected outlier, and the problem here is the selection of the best model, where the candidate models have different combinations of all possible dummy variables as explanatory variables. In this study, potential outliers can be incorporated into MLR model of equation (1) by the use of dummy variables. The problem for outlier detection in MLR is to select the best model. For this reason, the candidate MLR models have different combination of all possible dummy variables. In the next subsection, a brief description of detection of outliers in MLR models using information criteria is given, and then we will discuss information about GA for outlier detection when AIC', BIC' and ICOMP'(IFIM) criteria as a fitness function are used. These are as follows:

2.1 Outlier Detection in MLR using AIC', BIC' and ICOMP' Criteria

Numerous methods have been proposed in the literature for outlier detection using Akaike's information criterion (AIC) [23], Bayesian information criterion (BIC) [24], and Bozdogan's information complexity (ICOMP) criterion, [25,26]. In this study, AIC', BIC' and ICOMP'(IFIM)criteria were used for detecting outliers from MLR model having multicollinearity problem. AIC' and ICOMP'(IFIM)criteria were derived [27] and these are alternative criteria to BIC' approach for outlier detection in multiple regression model. They are given as follows:

- **AIC:** AIC was developed by Akaike [28]. It has played a significant role in solving problems in a wide variety of fields for analyzing actual data. The AIC is defined as

$$AIC = -2\log L(\hat{\theta}) + 2p, \tag{5}$$

where $\hat{\theta}$ is the maximum likelihood estimator of the parameter θ for an approximating model M, $L(\hat{\theta})$ is the maximized likelihood function, and p is the number of free parameters in M. Suppose that the observations y_α are independently and normally distributed with mean μ_α and variance σ^2. Then the density function of y_α can be written as

$$f(\,y_\alpha|\,\mu_\alpha,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{(y_\alpha-\mu_\alpha)^2}{2\sigma^2}\right\}. \tag{6}$$

Then, the AIC for a Gaussian linear regression model is given by

$$AIC = n\log(2\pi) + n\log(\hat{\sigma}^2) + n + 2(p+1), \tag{7}$$

where (p+1) is the number of estimated parameters ($\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_p$). The bias corrected in AIC is approximated by the number of parameters which are constant and have no variability. AIC criterion may be viewed as an asymptotically unbiased estimator of the Kullback-Leibler information which is a measure of discrepancy between statistical models, [29]. Thus, selection of a model minimizing AIC means that the selected model may be the best approximating model to the true model. For the data, the true model has infinite order, AIC provides an asymptotically efficient selection of a finite order model. AIC' for outlier detection in multiple regression models is given by [27],

$$\begin{aligned} AIC' &= AIC + \kappa m_d log(n) \\ &= nlog(2\pi) + nlog(\hat{\sigma}^2) + n + 2(p+1) + \kappa m_d log(n), \end{aligned} \tag{8}$$

where m_d is the number of outlier dummies, and kappa is extra penalty ($\kappa > 1$) for the dummies.

- **BIC:** BIC is based on the Bayesian method proposed by Schwarz [24]. It is defined as,

$$\text{BIC} = -2\log L(\hat{\theta}) + p\log(n), \tag{9}$$

where $\hat{\theta}$ is the maximum likelihood estimator of the parameter θ for an approximating model M, $L(\hat{\theta})$ is the maximized likelihood function, and (p+1) is the number of estimated parameters. Bayesian approach for outlier detection in multivariate samples is proposed by several researchers [2], [30–33]. Ting *et al.* [34] introduced a Bayesian way of dealing with outlier infested sensory data and developed a block box approach to the removed outliers in real time. The BIC is used for outlier detection in MLR model with dummy variables and the criterion can be calculated as,

$$\text{BIC} = \log(\hat{\sigma}^2) + m\log(n)/n, \tag{10}$$

where $\hat{\sigma}^2 = (e'e)/(n-p-1)$ is the estimated variance of regression model, and $m = 1 + p + m_d$, the total number of parameters in the estimated model and it consists of parameters for the constant together with the number of dummies outlier m_d. Generally a good model has small residuals, and few parameters. Therefore the smallest value of BIC is preferred, [18]. 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 [18],

$$\text{BIC'} = \log(\hat{\sigma}^2) + (p+1)\log(n)/n + \kappa m_d \log(n)/n, \tag{11}$$

where the kappa ($\kappa > 1$) is the extra penalty given to dummies outlier.

- **ICOMP:** In the literature many consistency results on AIC and BIC criteria are based on the central assumption that one of the models considered is true model. However, notably in the context of multiple regression, this assumption often does not hold, since one or more variables have been omitted from the model, [25]. Bozdogan and Haughton [25] introduced a concept of consistency for this case, and established a consistency property for ICOMP criterion. Each formulation of ICOMP has the attractive feature of implicitly adjusting for the number of parameters, the sample size, and controlling the risks of both insufficient and over parameterized models. ICOMP inverse Fisher information matrix (ICOMP(IFIM)) is shown as for multiple regression, [26]; [35].

$$\text{ICOMP(IFIM)}_{\text{Mul.Reg}} = nlog(2\pi) + nlog(\hat{\sigma}^2) + n + C_1(\hat{F}^{-1}(\hat{\theta})) \tag{12}$$

with

$$\begin{aligned} C_1(\hat{F}_R^{-1}(\hat{\theta})) = (p+1)\log\left[\frac{tr(\hat{\sigma}^2(X'X)^{-1}) + \frac{2\hat{\sigma}^4}{n}}{(p+1)}\right] \\ - \log\left|\hat{\sigma}^2(X'X)^{-1}\right| - \log\left(\frac{2\hat{\sigma}^4}{n}\right), \end{aligned} \tag{13}$$

where $\hat{\sigma}^2$ is the estimated variance of regression model. As the number of parameters increases (i.e., as the size of X increases), the error variance $\hat{\sigma}^2$ gets smaller even though the complexity gets larger. Also, as $\hat{\sigma}^2$ increases, $(X'X)^{-1}$ decreases. Therefore $C_1(\hat{F}^{-1}(\hat{\theta}))$achieves a trade-off between these two extremes and guards against multicollinearity. To preserve scale invariance, the correlational form of information fisher information matrix (IFIM), $\hat{F}^{-1}$, is used, and the correlational form of ICOMP(IFIM) regression is given by (11), [35]. ICOMP'(IFIM) for outlier detection in multiple regression models is defined by

$$\begin{aligned} \text{ICOMP'(IFIM)}_{\text{Mul.Reg}} &= \text{ICOMP(IFIM)} + \kappa m_d log(n) \\ &= nlog(2\pi) + nlog(\hat{\sigma}^2) + n + C_1(\hat{F}^{-1}(\hat{\theta})) + \kappa m_d log(n). \end{aligned} \tag{14}$$

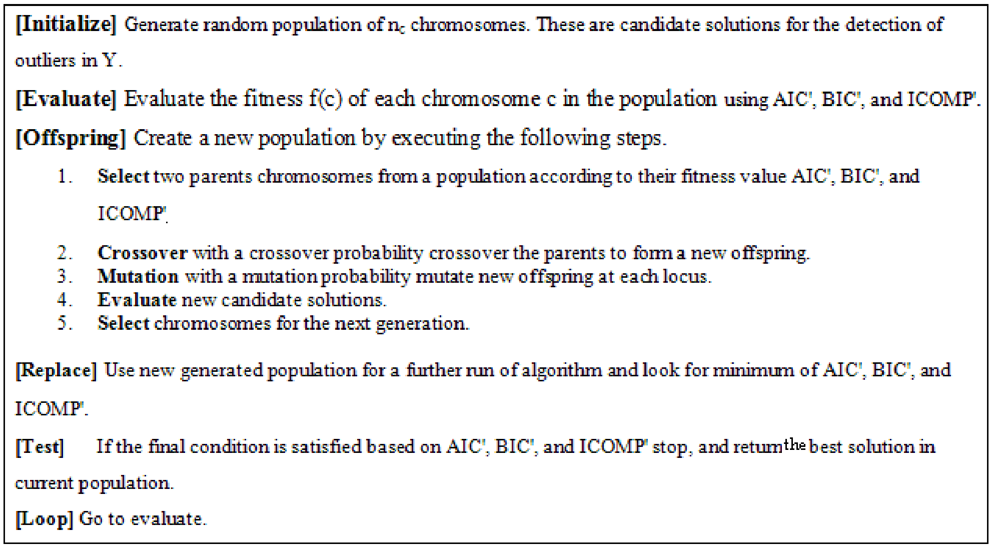
where,

$$\begin{aligned} C_1(\hat{F}_R^{-1}(\hat{\theta})) = (p+1)\log\left[\frac{tr(\hat{\sigma}^2(X'X)^{-1}) + \frac{2\hat{\sigma}^4}{n}}{(p+1)}\right] \\ - \log\left|\hat{\sigma}^2(X'X)^{-1}\right| - \log\left(\frac{2\hat{\sigma}^4}{n}\right). \end{aligned} \tag{15}$$

AIC' and ICOMP'(IFIM) criteria provide a more judicious penalty term $(\kappa m_d log(n))$ than BIC', since counting and penalizing the number of parameters for outlier dummies in a model is necessary. Gürünlü Alma *et al.* [27] illustrates the practical utility and the importance of AIC' and ICOMP'(IFIM) criteria by providing simulation examples for comparing their performance against BIC'.

2.2 Genetic Algorithms for Outlier Detection

GA is a 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, [36]. GA 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.



[Initialize] Generate random population of n_c chromosomes. These are candidate solutions for the detection of outliers in Y.

[Evaluate] Evaluate the fitness f(c) of each chromosome c in the population using AIC', BIC', and ICOMP'.

[Offspring] Create a new population by executing the following steps.

1. **Select** two parents chromosomes from a population according to their fitness value AIC', BIC', and ICOMP'.
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 AIC', BIC', and ICOMP'.

[Test] If the final condition is satisfied based on AIC', BIC', and ICOMP' stop, and return the best solution in current population.

[Loop] Go to evaluate.

Figure 1: The Outline of GA

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 Figure 2. Each chromosome consists of p genes, where p is the number of outliers $((o_p, (p = 1, ..., n))$ given in a model. For instance for $p = 3$; the first, second and n-1'st observations are outliers in Figure 2.

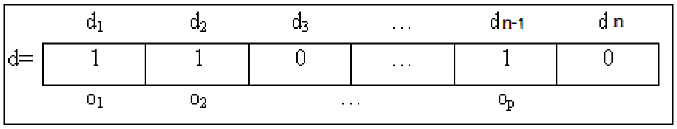


Figure 2: The Structure of a Chromosome (c)

- **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 AIC', BIC' and ICOMP'(IFIM) in (7, 10, 13) for MLR model with the corresponding dummy variables.

- **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 AIC', BIC' and ICOMP'(IFIM) values for them computed. The chromosomes 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 is 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 AIC', BIC' and ICOMP'(IFIM) values of the chromoosmes. The best chromosome has the smallest value of the fitness function AIC', BIC' andICOMP'(IFIM), are more likely to pass their genes onto the next generation. A crossover probability is selected as $p_c = 1$ and it 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 AIC', BIC' and ICOMP'(IFIM) 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.

Table 1 shows the parameters of GA with AIC', BIC' and ICOMP'(IFIM) 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

Number of Generations	250
Population Size	40
Fitness Value	AIC', BIC', and ICOMP'(IFIM)
Crossover Probability	1
Mutation Probability	0.01
Elitism	For two parents

3 Generating of Simulation Data Sets and Experimental Results

The performance of AIC', BIC' and ICOMP'(IFIM) information criteria to outlier detection for MLR model having multicollinearity problem was evaluated and performance of GA was demonstrated by simulation experiments. The next subsection shows the steps of data generation and results of AIC', BIC' and ICOMP'(IFIM) information criteria to outlier detection for MLR model having multicollinearity problem.

3.1 Data Generation

In this subsection, simulation is performed to evaluate the performance comparisons of information criteria. The data sets were generated based on McDonald and Galarneau [37], which explained how to generate a suitable design. A detailed description of simulation protocol for the regression model can be summarized in Table 2.

As seen from Table 2, $\varepsilon_{1i}, ..., \varepsilon_{4i}$ are independent and identically distributed (i.i.d.) according to normal distribution. The α parameter controled the degree of collinearity (>0.70) between predictor variables, and α values were selected as $\alpha_{1,...,3} = 0.15$. The correlation and variance inflation factor (VIF) values of predictor variables for $n \quad = 30$ are shown in Table 3.

Percentage levels of outliers in the dependent variable were selected at 5% and 10%. The outliers were generated from the uniform distribution which lies at least $+3\sigma$ from the mean of Y_i. Under these conditions, the simulation was done on the explanatory variables and the error terms for ($i \quad = 1,\ldots,n$) observations. Then, the response variable was generated. After Y_i were generated from normal distribution, outlier observations were generated from uniform distribution which take into account the percentage of outliers.

Table 2: Multiple Regression Model and Variates

$X_{1i,\ldots,5i}$	$\varepsilon_{1i,\ldots,4i}$	$\alpha_{1,\ldots,3}$
$X_{1i} = 1 + \varepsilon_{1i}$	$\varepsilon_{1i} \sim N(0,1)$	$\alpha_1 = 0.15$
$X_{2i} = 1 + 0.3\varepsilon_{1i} + \alpha_1\varepsilon_{2i}$	$\varepsilon_{2i} \sim N(0,1)$	$\alpha_2 = 0.15$
$X_{3i} = 1 + 0.3\varepsilon_{1i} + 0.3\varepsilon_{2i} + \alpha_2\varepsilon_{3i}$	$\varepsilon_{3i} \sim N(0,1)$	$\alpha_3 = 0.15$
$X_{4i} = 1 + 0.3\varepsilon_{1i} + 0.3\varepsilon_{2i} + 0.3\varepsilon_{3i} + \alpha_3\varepsilon_{4i}$	$\varepsilon_{4i} \sim N(0,1)$	
$X_{5i} \sim N(1,4)$	$\varepsilon_i \sim N(0,1)$	
Multiple Regression Model: $Y_i = 0.8 + 0.8X_{1i} + 0.8X_{2i} + 0.8X_{3i} + 0.8X_{4i} + 0.8X_{5i} + \varepsilon_i$		
$i = 1, \ldots, n, n =$ 30, 50, 100, 300		
Percentage of Outliers: 5%- 10%		

Table 3: Correlation and VIF Values of Predictor Variables

Correlation Values				
	X_1	X_2	X_3	X_4
X_2	0.91			
X_3	0.68	0.89		
X_4	0.54	0.74	0.926	
X_5	-0.08	-0.09	-0.07	-0.13

VIF Values				
X_1	X_2	X_3	X_4	X_5
14.68	52.90	51.72	14.37	1.09

For example, for the sample size $n \quad = 30$ and percentage of outliers equal to 5%, it can generate 2 outlying observation. Outliers were then added to the dependent variables. For each of the combinations of parameters in Table 2. 100 data sets were generated taking into account the regression model, so that 800 data sets were generated. Then, these data sets were applied to AIC', BIC' and ICOMP'(IFIM) information criteria with their penalized values of kappa equal to 3.

3.2 Performance Comparison of AIC', BIC' and ICOMP'(IFIM) for Outlier Detection in MLR Model Having Multicollinearity Problems using GA

The GA is used to find the optimal solution through for each combination of experiments. Each dataset contains a known percentage of outliers, and the GA successfully detected these outliers in all of the datasets tested. GA can detect the outliers by simultaneously searching in the solution space, therefore the GA based outlier detection method allows for detection of multiple outliers, not just one at a time. The simulation results are shown in Table 4; the values are the percentage of outliers for 100 replicates. It tests the performance of information criteria under two components. These are numbers of incorrectly identified observations as outliers (swamping) (I) and numbers of failure to identify any of the outliers (masking) (F) in all iterations for each subsets. Percentage of outliers (P) is calculated by

$$P = \Big(\frac{I+F}{T}\Big)100\%,$$

where T is the total number of outliers for all iterations in each subset.

As seen from Table 4 the true results for experiments are obtained for sample sizes $n =$ 30, 50,100, 300 and percentage of outliers 5% and 10%. A simulation study is carried out to support the good behavior of the AIC' and ICOMP'(IFIM) criteria. It is clear that from simulation results of AIC' and BIC' for $n = 50$ and 100, the percentage of outliers with 10% give true information about how many observations are found as outlier. Therefore, it is concluded that the best performance for outlier detection using AIC' and ICOMP'(IFIM) in MLR models is when by n <100 except for BIC'.

Table 4: Percentage of Outliers Finding by GA as the Fitness Function AIC', BIC', and ICOMP'(IFIM)

	5%			10%		
n	AIC'	BIC'	ICOMP'(IFIM)	AIC'	BIC'	ICOMP'(IFIM)
30	3.50	66.50	3.50	3.33	59.00	2.67
50	0.67	5.00	1.33	3.00	5.80	1.40
100	9.80	12.20	10.80	4.70	6.20	5.50
300	15.20	18.74	12.84	11.78	22.45	12.63
Total	29.17	102.44	28.47	22.81	93.45	22.20

In Figure 3, one important result from these comparisons was that the run time of information criterion tends to increase linearly as both the number of observations and the number of finding outliers is increased. In the case that the number of observations goes to infinity, the criterion will estimate the right percentage of outliers or even detects successfully most of outliers. These results suggest that the ICOMP' based method is less affected by changes in the dimension of regression models, percentage of contamination of data, and sample sizes.

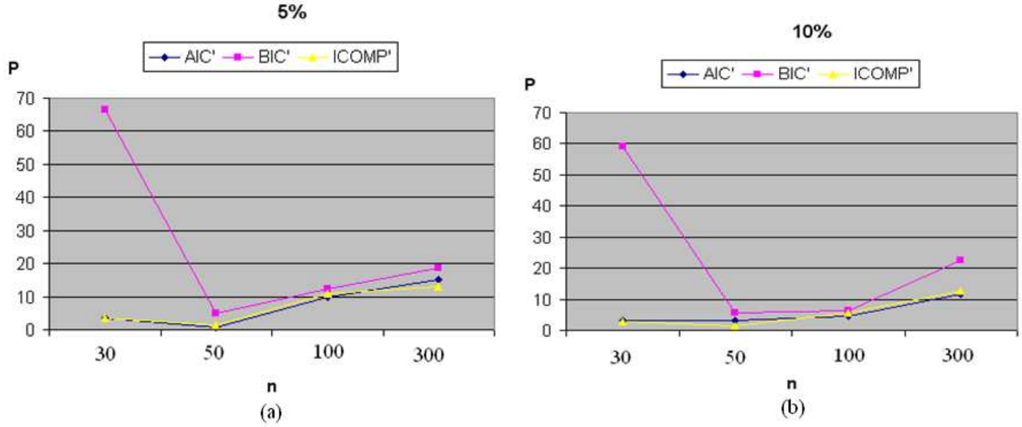


Figure 3: P Values When AIC', BIC', and ICOMP'(IFIM) as the Fitness Function of GA

4 Conclusions

The purpose of this experimental study was to test the scalability of the GA based on fitness function as information criteria when MLR models having multicollinearity problems in handling different sample sizes and different contaminated data with outliers. GA mimics evolution is also a useful optimization tool for statistical modeling. In this study, it is demonstrated that AIC', BIC', and ICOMP'(IFIM) criteria and a GA outline for outlier detection in MLR have multicollinearity problems. The value of information based selection criterion is calculated for observation as a measure of the fitness of dependent variable in MLR. GA can simultaneously search in the solution space and also find the outliers for multicollinearity problems. The simulation results are shown in Table 4, where the values in cells are defined as extra total number of finding outliers in all iterations in the dependent variable with GA. 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 percentage of outliers in MLR models having multicollinearity problems and the second is the scalability against the power of different sample sizes. Figure 4 shows the results of using GA to find diversity number of outliers on data set. One important observation from this figure is that the GA can accurately finds the outliers especially when the sample size is smaller than 100 observations for MLR models with multicollinearity problems handling. However, we note that numerical results clearly demonstrate that all criteria performances are affected by multicollinearity problems in MLR model when sample sizes increases. BIC' criterion performance are greatly affected as compared to AIC' and ICOMP' when it is used in outlier detection in multiple regression.

References

[1] Fox, J. *Applied Regression Analysis, Linear Models and Related Methods.* 3th Edition. New York: Sage Publication. 1997.

[2] Hawkins, D. *Identification of Outliers.* Reading, London: Chapman and Hall. 1980.

[3] Acuna, E. and Rodriguez, C. A Meta analysis study of outlier detection methods in classification. *Technical paper, Department of Mathematics, University of Puerto Rico at Mayaguez, In proceedings IPSI* 2004

[4] Shekhar, S. and Chawla, S. *A Tour of Spatial Databases.* New York: Prentice Hall. 2002

[5] Atkinson, A. C. *In Plots, Transformations, and Regression: An Introduction to Graphical Methods of Diagnostic Regression Analysis.* Oxford: Clarendon Press. 1985.

[6] Barnett, V. and Lewis T. *Outliers in Statistical Data.* 3rd Edition. USA: John Wiley and Sons. 1994.

[7] Cambell, N. A. Robust procedures in multivariate analysis I: robust covariance estimation. *Applied Statistics.* 1980. 29: 231-237.

[8] Hadi, A. S., Simonoff, J. S. Procedures for the identification of multiple outliers in linear models. *Ammer. Statist. Assoc.* 1993. 88: 1264-1272.

[9] Huber, P. J. *Robust Statistic.* New York: John Wiley & Sons. 1981.

[10] Lopuhaa, H. P. and Rousseeuw, P. J. *Breakdown Point of Affine Equivariant Estimators of Maultivariate Location and Covariance Matrice.* Technical Report, Faculty of Mathematics and Informatics, Netherlands: Delft University of Technology. 1987.

[11] Kianifard, F. and Swallow, W. A monte carlo comparison of five procedures for identifying outliers in linear regression. *Communication in Statistics Theory Methods.* 1990. 19: 1913-1938.

[12] Marona, R. A. Robust M-estimates of multivariate location and scatter. *Ann. Stat.* 1976. 4: 51-67.

[13] Riani, M. and Atkinson, A. C. Robust diagnostic data analysis: transformations in regression. *Technometrics.* 2000. 44: 384-391.

[14] Sebert, D. M. *Identifying Multiple Outliers and Influential Subsets: A Clustering Approach.* AZ: Unpublished Dissertation, Arizona State University. 1996.

[15] Woodruff, D. L. and Rocke, D. M. Computable robust estimation of multivariate location and shape in high dimension using compound estimator. *J. Ammer. Statist. Assoc.* 1994. 89: 888-896.

[16] Mason, R. L., Gunst, R. F. and Webster, J. T. Regression analysis and problem of multicollinearity. *Communication in Statistics.* 1975. 4(3): 277-292.

[17] Dempster, A.P., Schatzoff, M., Wermuth, N. A simulation study of alternatives to ordinary least square. *Journal of American Statistical Association.* 1977. 72: 77-91.

[18] Tolvi, J. Genetic Algorithms for outlier detection and variable selection in linear regression models. *Soft Computing.* 2004. 527-533.

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

[20] Kullback, S. *Information Theory and Statistics.* New York: Dover Publications. 1996.

[21] Belsley, D. A., Kuh E., Welsch R. E. *Regression Diagnostics: Identifying Influential Data and Source of Collinearity.* New York: John Wiley. 1980.

[22] Polasek, W. Multivariate regression and ANOVA models with outliers: a comparative approach. *Reihe Ökonomie / Economics Series.* 2003. 136, ISSN: 1605-7996.

[23] Akaike, H. Information theory and an extension of the maximum likelihood principle. *Proc. 2nd Inter.Symposium on Information Theory, 267-281, Budapest.* 1973.

[24] Schwarz, G. Estimating the dimension of a model. *The Annals of Statistics.* 1978. 6(2): 461-464.

[25] Bozdogan, H., Haughton D.M.A. Informational complexity criteria for regression models. *Computational Statistics and Data Analysis.* 1998. 28: 51-76.

[26] Bozdogan, H. Akaike's information criterion and recent developments in information complexity. *Journal of Mathematical Psychology.* 2000. 44: 62-91.

[27] Gürünlü Alma, Ö., Kurt, S. and Uur, A. Genetic algorithms for outlier detection in multiple regression with different information criteria. *Journal of Statistical Computation and Simulation.* 2009. d**oi:** 10.1080/00949650903136782

[28] Akaike, H. A new look at the statistical model identification. *IEEE Transaction on Automatic Control.* 1974. 19: 716-723.

[29] Kullback, S., and Leibler, R.A. On information and sufficiency. *Annals of Mathematical Statistics* 1951. 22: 79-86

[30] Guttman, I. Care and handling of univariate or multivariate outliers in detecting spuriosity a bayesian approach. *Technometrics.* 1973. 15: 723-738.

[31] Gnanadesikan, R., Kettering, J. Robust estimates, residuals and outlier detection with multiresponse data. *Biometrics.* 1972. 28: 81-124.

[32] Rousseeuw P. J. and Leroy, A. M. *Robust Regression and Outlier Detection.* New York: Wiley-Interscience. 1987.

[33] Varbanov, A. Bayesian approach to outlier detection in multivariate normal samples and linear models. *Communications in Statistics-Theory and Methods.* 1998. 27(3): 547-557.

[34] Ting, J., A., Souza, D. and Schaal, S. Automatic outlier detection: a Bayesian approach. *IEEE International Conference on Robotics and Automation, Roma-Italy.* 2007. 2489-2494.

[35] Bozdogan, H. *Statistical Data Mining and Knowledge Discovery.* USA: Chapman and Hall/CRC. 2004.

[36] Goldberg, D. E. *Genetic Algorithms in Search, Optimization, and Machine Learning.* USA: Addison-Wesley. 1989.

[37] McDonald, G. C. and Galarneau, D. I. A monte carlo evaluation of some ridge type estimators. *Journal of. American Statistics.* 1975. 70: 407-416.