Cilt: 06 Sayı: 01 Kategori: 02 Sayfa: 38-51 ISSN No: 1303-6319 Volume: 06 Number: 01 Category: 02 Page: 38-51

OUTLIER DETECTION IN MULTIPLE REGRESSION MODELS USING GENETIC ALGORITHMS AND BAYESIAN INFORMATION CRITERIA

Özlem GÜRÜNLÜ ALMA* **Serdar KURT****
Aybars UĞUR***

ABSTRACT

Statistical models, particularly regression models, are most useful devices for extracting and understanding the essential features of datasets. However, most of the databases in real-world include a particular amount of abnormal values, generally termed as outliers. An accurate identification of outliers plays a significant role in statistical analysis especially regression models. Nevertheless, many classical statistical models are blindly applied to data sets containing outliers, the results can be misleading at best. The appearance of outliers can exert negative influences on the fit of the multiple regression models. The aim of this study is to define outlier detection method using Genetic Algorithms (GA) with Bayesian Information Criterion (BIC) and to illustrate the algorithm with real and simulation data. We use a fitness function which is based on BIC in this algorithm. The criteria's value indicates a better model to fit data, the presence of one or more outliers will negatively impact the regression model and result in larger BIC values.

Keywords: Bayesian information criterion, Genetic algorithms, Multiple regression models, Outlier detection.

1. INTRODUCTION

According to Barnett and Lewis (1994), an outlier is one that appears to deviate so much from other observations of the sample. There are several statistical methods for outlier detection in different conditions. However, it may be difficult to decide which methods can be used in practical work. And if outliers are detected in the data, there are different ways of taking them into account in the analysis. For example, one can either remove the outlying observations from the data or incorporate the detected outliers into the statistical model.

A typical approach of detecting outliers is to characterize what normal observations look like, and then to single out samples that deviate from these normal properties. Existing methods for outlier detection include methods that classify a data point based on a distance from the expected value approaches that use information theoretical principles, such as selecting the subset of data points that minimize the prediction error. Outlier classification based on Mahalanobis distance can work quite well, but tends to

* Research Assistant, Dokuz Eylül University, Department of Statistics, Tınaztepe Campus, İzmir. e-mail: ozlem.gurunlu@deu.edu.tr
** Professor Dr., Dokuz Eylül University, Department of Statistics, Tınaztepe Campus, İzmir. e-mail: serdar.kurt@deu.edu.tr
*** Assistant Prof. Dr., Ege University, Faculty of Engineering, Department of Computer Engineering, İzmir. e-mail: aybars.ugur@ege.edu.tr

require the setting of some threshold that defines whether a point is an outlier or not. This threshold value typically needs to be tuned manually beforehand in order to determine its empirically optimal value for the system. Information theoretical approaches, outlier may be detected active learning (Abe et al., 2006), clustering (Barnett and Lewis, 1994; Breitenbach and Grudic, 2005; MacOueen, 1967) or mixture models (Scott, 2005). These methods may require sampling, the setting of certain parameters such as the optimal k in k-means, and may not all lend themselves to a real time implementation. There exist also a large number of outlier detection methods in literature (Ben-Gal, 2005). Traditionally, these can be categorized into three approaches: the statistical approach, the distance-based approach and the density based approach. But many of them are limited by assumptions of a distribution or limited in being able to detect only single outlier. If there is a known distribution for the data, then using that distribution can aid in finding outliers. Often, a distribution is not known, or the experimenter does not want to make an assumption about a certain distribution (Amidan et al., 2005). In addition to the basic problem of outlier detection mentioned above, there are additional problems in outlier detection for practical work. Data sets with multiple outliers are subject to masking and swamping effects. Although not mathematically rigorous, the following definition gives an intuitive understanding for these effects (Ben-Gal, 2005; Davies and Gather 1993).

According to Acuna and Rodriguez (2005), masking effect is said that one outlier masks a second outlier, if the second outlier can be considered as an outlier only by itself, but not in the presence of the first outlier. Thus, after the deletion of the first outlier the second instance is emerged as an outlier. Swamping effect is said that one outlier swamps a second observation, if the latter can be considered as an outlier only under the presence of the first one. In other words, after the deletion of the first outlier the second observation becomes a non-outlying observation (Acuna and Rodriguez, 2005). Sequential detection of outliers may therefore be misleading, if the detection of one outlier causes the subsequent detection of other outliers to be defective, because of either swamping or masking, or even both. Identification of outliers in Multiple Linear Regression (MLR) models is not trivial, especially when exist several outliers in data. 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. Therefore, simultaneous outlier detection method is important issue and in this work it is considered in MLR models.

GA has been used for outlier detection and model selection of linear regression models or times series. Jann (2000) describes a GA for the detection of level shifts in a time series, the problems caused by change points are similar to those caused by outliers. Ishibuchi et al., (2001) were used GA for the feature selection in data mining and they give a lot of references about this literature. Additionally, the use of GA for outlier detection and variable selection can be found in (Tolvi, 2004).

In this work, we are interested with the problem of identifying outliers and detection of outliers in the dependent variable of MLR models using GA. A robust simultaneous procedure is investigated for identification of outliers using Bayesian information criteria (Kullback, 1996). The scalability of information criteria is considered with a real data and also by generating experimental data. We have shown the behavior of our approach for different sample sizes and different 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 a penalized value of Bayesian information criteria and obtained results for different values of it.

2. METHOD

As mentioned in the first section, outliers can be described as; given a set of n data points and k the expected number of outliers, find the top k outliers that are considerably different, inconsistent with the respect to the remaining data. The outlier detection problem can be viewed as two sub problems:

- which observations data can be considered as inconsistent or exceptional in a given data set,
- finding an efficient method to detection of outliers.

Based on the above sub problems, the purpose of this work is to investigate detection of outliers in MLR models based on GA and Bayesian information criteria, which are described in the next subsections.

2.1 Outlier Detection in Multiple Linear Regression

The purpose of regression analysis is to fit equations to observed variables. The MLR equation takes the following type:

$$\mathbf{Y} = \beta_0 + \beta_1 \mathbf{X}_1 + \beta_2 \mathbf{X}_2 + ... + \beta_k \mathbf{X}_k + \boldsymbol{\varepsilon} \quad (1)$$

$$\hat{\mathbf{Y}} = \hat{\beta}_0 + \hat{\beta}_1 \mathbf{X}_1 + \hat{\beta}_2 \mathbf{X}_2 + ... + \hat{\beta}_k \mathbf{X}_k \quad (2)$$

where:
$Y \in \Re^n$ is a response variable,
$\hat{Y}$ is the predicted value of the dependent variable,
$X_1, \ldots, X_k \in \Re^n$ are different explanatory variables,
β_0 is the intercept on the Y axis, and
$\beta_1, ..., \beta_k$ are the regression coefficients for each of the independent variables.

Ordinary Least Squares (OLS) remains the most often utilized regression coefficient estimation method. This method optimizes the fit of the model by minimizing the sum of the squared deviations between the actual and predicted Y values $\sum e^2 = \sum (Y - \hat{Y})^2$. Computing an intercept term and estimating a set of β coefficient is calculated by $\hat{\beta} = (X^T X)^{-1} X^T Y$. However, some researchers began to realize that real data usually do not completely satisfy the classical assumptions. These are for errors:

- normally distributed,
- have equal variance at all levels of the independent variables and
- uncorrelated with both the independent variables and with each other.

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. Because of these reasons, the detection of outliers is important for multiple regression analysis.

In this work 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. For example, we assume that the last observation is an outlier, then one dummy variable to be added to the model (2), and the independent variable could be below.

$$X_{N\times(k+1)} = \begin{bmatrix} x_{11} & x_{1k} & 0 \\ \dots & \dots & \dots \\ x_{N1} & x_{Nk} & 1 \end{bmatrix}$$

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.

The BIC criteria 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 for outlier detection in multiple regression (Tolvi, 2004).

A problem in using the BIC for outlier detection is that by itself tends to include unnecessary outlier dummies. To circumvent this problem, a correction to the criterion is used. The corrected BIC takes into account the different nature of outlier dummies and other variables, and has a different penalty term for different variables. This takes the form of an extra penalty (κ) for the dummies. The corrected BIC, denoted BIC′ (Tolvi, 2004), is given by

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

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

2.2 A Genetic Algorithm for Outlier Detection

GA is a stochastic search technique that guides a population of solution towards an optimum using the principles of evolution and natural genetics. The algorithm starts with a randomly generated initial population consisting of sets of chromosomes that represent the solution of the problem. These are evaluated for the fitness function or one of the objective functions, and then selected according to their fitness (Bozdogan, 2004; Goldberg, 1989; Rothlauf, 2006). To perform its optimization like process, the GA employs three operators to propagate its population from one generation to another. The first operator is the selection operator, which mimics the principal of the survival of the fittest. The second operator is the crossover operator, which mimics mating in biological populations. It propagates features of good surviving designs from the current population into the future population, which will have better fitness value on average. The last operator is the mutation operator, which promotes diversity in population characteristics.

In this paper, for the given set of objects located in the space, GA was used to detect the outliers. There are five primary elements in the GA, and the parameter setting of GA was shown as following in details.

- **Parameter Encoding:** The coding of the candidate models for outlier detection is straightforward. Each model also called an individual or 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 $i = 1,...,N$. For example, a model with a dummy variable for the last observation is described by the vector $d = (0\ 0...1)$. These dummy variables for outlier observations must be created before the GA is run on a data set.

In this study, the structure of a chromosome or an individual is shown in Figure1. It has N genes which is the number of observations in data set. Each chromosome consists of p genes, where p is the number of outliers given in a model. For instance, if the second and N-1th observations are outliers in data, the chromosome structure will be as seen in Figure 1.

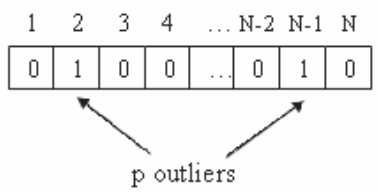

Figure 1. The structure of a chromosome in GA

• **Fitness Function:** The genes, which represent the serial number of outliers, are updated with each new population created. The random population is sorted based on the least fitness is considered to be the elite chromosome within population. The fitness of an individual is computed as the BIC' which is given equation 4 for MLR model with the corresponding dummy variables.

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

• **Selection Operator:** Stochastic uniform selection function is used in our 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. The algorithm moves along the line in steps of equal size. At each step, the algorithm allocates a parent from the section it lands on. The first step is a uniform random number less than the step size. It is noted that the results can be improved if a small number of the best individuals. These are kept the same from one generation to the next. In our GA the best two individuals are kept as elite population.

• **Crossover Operator:** The next generation of individuals from the previous one, is based on the BIC' values of the individuals. The best individuals has the smallest value of the fitness function BIC', are more likely to pass their genes onto the next generation. This procedure is repeated to create the same number of individuals as existed in the previous generation. Scattered crossover model is used in our approach and the crossover probability is defined as one. A crossover probability $p_c = 1$ 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:** Mating of the individuals from the previous one generation will not be enough for diversity of population. In evolutionary terms, more genetic variation in the population is needed. To this end, the individuals 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 0.01.

In addition to crossover and mutation, a condition for the maximum number of dummies is used to alter the population. This condition is used in order to keep the candidate models from having too many variables, because only a few dummies will be allowed in the final model. The rule states that if a candidate model has more than N/2 dummy variables, or outliers is more than 50% of the number of observations, it is dropped from consideration. Depending on the particular crossover and mutation rates, the second generation will be composed entirely of offspring models or of a mixture of offspring and parent models. In summary, the outline of the GA is shown in Figure 2.

1. **[Start]** Generate random population of N_C chromosomes. These are suitable solutions for the problem.
2. **[Fitness]** Evaluate the fitness of each chromosome in the population using BIC'.
3. **[New population]** Create a new population by repeating following steps until the new population is complete.
   (a) **[Selection]** Select two parent chromosomes from a population according to their fitness value BIC'. The better fitness, the bigger chance to be selected.
   (b) **[Crossover]** With a crossover probability cross over the parents to form a new offspring. If no crossover was performed, offspring is an exact copy of parents.
   (c) **[Mutation]** With a mutation probability mutate new offspring at each locus
   (d) **[Accepting]** Place new offspring in a new population.
4. **[Replace]** Use new generated population for a further run of algorithm and look for the minimum of the BIC'.
5. **[Test]** If the final condition is satisfied based on the BIC' stop, and return the best solution in current population
6. **[Loop]**Go to step 2.

Figure 2. The outline of the GA

In the approach, the number of outliers was specified firstly in dependent variable of MLR model, and a random population of chromosomes was created representing the solution space. Each chromosome of this random population represents a N observations in data set and each locus in the chromosome is a binary code indicating the outlier observation (1) or non-outlying observation (0) in data set. The GA proceeded to find the optimal solution as fitness function value (BIC') of each chromosome. The process continues one generation after another for a specified number of generations controlled by the researcher.

3. FINDING

A comprehensive performance study has been conducted to evaluate our algorithm. This algorithm was implemented in Matlab. We ran this algorithm on some real life data sets: Scottish Hill Racing and Stack Loss. These data sets demonstrated the effectiveness of our method against other algorithms. Data is generated for N=20, 30, 40, 50 and 100 observations and different number of outliers are inserted for each data set by taking into account of percentage of outliers in the dependent variable.

3.1 Experiments: Simultaneous Outlier Detection

In this paper, 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 (Hoeting et al., 1996)[§]. In this subsection, it is investigated that detect outliers from these data sets with GA. Some information on the data sets and results are following;

[§] These data sets are available from one of the authors' website. This web address is http://www.stat.colostate.edu/~jah/index.html, access date: 30.04.2009

i. Scottish Hill Racing: The first example involves data supplied by Scottish Hill Runners Association (Atkinson, 1986). 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. Several authors have examined these data sets using both predictors in their analysis (Atkinson, 1986; Hadi, 1986; Hoeting et al., 1996). They concluded that races 7^{th} and 18^{th} observations are outliers. After they removed observations 7 and 18, their methods indicated that 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 (Fox, 1997; Hoaglin and Tukey, 1983; Hoeting et al.,1996). The scatter plot of this data set is shown in Figure 3.

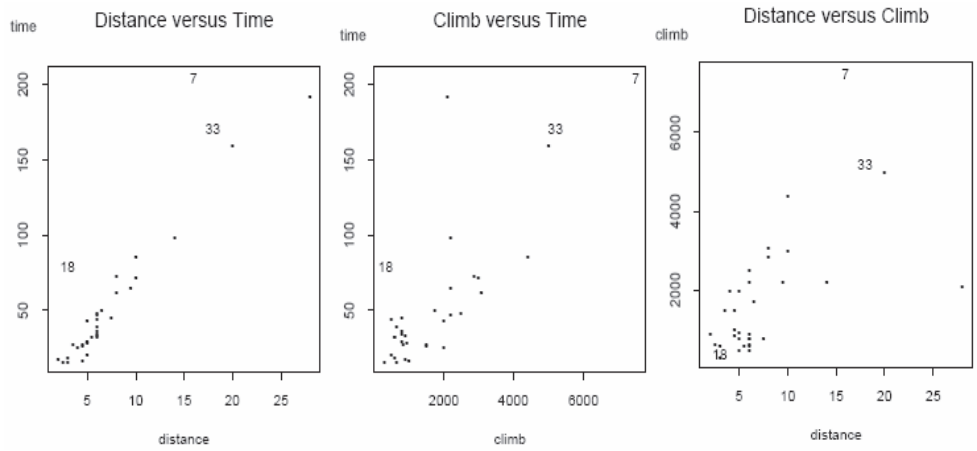

Figure 3. Scatter plot of Scottish hill racing data[**]

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. The estimated model and estimated variance with the three outlier dummies are

$$y = -8{,}45 + 6{,}63x_1 + 0{,}00661x_2 + 57{,}1d_1 + 64{,}6d_2 + 24{,}8d_3 \text{ and } \hat{\sigma}^2 = 22\,.$$

Then, the optimal fitness function value of GA has a BIC' value 4.30.

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

[**] Numbers correspond to race numbers 7, 18, 33. Distance is given in miles, time is given in minutes, and climb is given in feet.

the tower. Small values of the respond correspond to efficient absorption of the nitric oxides. In earlier research (Atkinson, 1986; Hoeting et al., 1996) been identified as outliers four observations. These are 1, 3, 4, and 21 observations. This data set provides an interesting extreme example of masking (Atkinson, 1986). The detection of any of these outliers is very difficult if only one observation at a time is examined. But the simultaneous methods 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 regression model and variance with the four outlier dummies are

$$y = -37{,}7 + 0{,}798x_1 + 0{,}577x_2 - 0{,}0671x_3 + 6{,}22d_1 + 6{,}43d_2 + 8{,}17d_3 - 8{,}63d_4$$

and $\hat{\sigma}^2 = 1{,}57$. The optimal fitness function value of GA has a BIC' value 2.69.

3.2 Data Generation and Outlier Detection in MLR Models

In order to study the performance of the BIC' criterion and also the role of κ values for outlier detection, we conduct a simulation study. The conditions under which the simulation is performed are;

- the linear regression model is selected as $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$,
- the first explanatory variable X_1 is generated from Normal (3,1), and the second explanatory variable X_2 is generated from Normal (2,1),
- the elements of $\beta_0 = 0$, and β_1, β_2 are generated from Uniform (1,2),
- the error terms are independent and identically distributed according to standard normal distribution N(0,1),
- the sample size N is determined as different sizes N=20, 30, 40, 50, and 100,
- percentages of outliers (P_O) in the dependent variable each of sample size are between %5-%10,
- the outliers are generated from the uniform distribution which lie at least 3σ from the mean of y_i and,
- the Kappa values are selected as $\kappa = 2, 3, 4, \text{and } 5$.

Under these conditions, firstly we simulate the explanatory variables and the error terms for i= 1,…,N observations and N=20, 30, 40, 50, 100. Then, we generate the response variable from y_i each of different sample size. After we generated y_i from normal distribution, we generated outlier observations from uniform distribution take into account of percentage of outliers. For example, for the sample size N=20 and percentage of outlier for the %5, it can be generated 1 outlier observation. However, two outliers must be added for the sample size 30 and 50 for the percentage of outlier 5, because of rounding problems.

Then, the percentage of outliers must be 6% for the sample sizes 30 and 50. The number of outliers for different sample sizes and the percentages are given in Table 1.

Table 1. Number of outliers for different sample size and percentage of outliers in dependent variable

	N				
P_O	20	30	40	50	100
5	1	2	2	3	5
10	2	3	4	5	10

Outliers are then added to the dependent variables. The iteration number for each combination of experiments is 100. Table 2 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 2. The parameters of the GA for the simulated model

Sample Size of Simulation Data	**N=20, 30, 40, 50, 100**
Number of Generations	**250**
Population Size	**40**
Fitness Value	BIC′
Crossover Probability	**1**
Mutation Probability	**0.01**
Elitism	**For two parents**

The computational capacity in terms of the number of generations needed to find the true model is increased by an increase in the sample size. GA can simultaneously search in the solution space and find the outliers. The simulation results are shown in Table 3, where the value $T_{Outliers}$ is defined as total numbers of outliers in all iterations and $P_{Outliers}$ is defined as percentage of outliers in dependent variable finding with GA.

Table 3. Generating descriptions of data sets and total number of outliers found

Generating Data Sets Descriptions			Results Finding with GA κ							
			2		3		4		5	
N	$T_{Outliers}$	$P_{Outliers}$	$T_{Outliers}$	$P_{Outliers}$	$T_{Outliers}$	$P_{Outliers}$	$T_{Outliers}$	$P_{Outliers}$	$T_{Outliers}$	$P_{Outliers}$
20	100	5	283	14	105	5	105	5	105	5
30	200	6	296	10	202	6	202	6	202	6
40	200	5	383	10	215	5	215	5	215	5
50	300	6	323	7	300	6	300	6	300	6
100	500	5	683	7	502	5	502	5	502	5
20	200	10	285	14	210	10	210	10	210	10
30	300	10	410	14	302	10	302	10	302	10
40	400	10	660	17	414	10	414	10	414	10
50	500	10	536	11	505	10	505	10	505	10
100	1000	10	1207	12	1002	10	1002	10	1002	10

As seen in Table 3 the true results for experiments are obtained for values of $\kappa = 3, 4,$ and 5 for sample size is N=20, 30, 40, 50, 100, and percentage of outliers %5-10. A simulation study is carried out to support the good behavior of the BIC′ when different percentage of outlier and different sample size. It is clear that from simulation results for high values of Kappa coefficient ($\kappa > 2$) 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 two. The important issue is that the BIC′criteria can not be affected masking or swamping effects finding outliers so we also said that this criteria is robust than other outlier detection methods.

In Figure 4, it is seen that the Kappa coefficient good results when the dependent variable Y containing of %5 outlier observation for all of sample size.

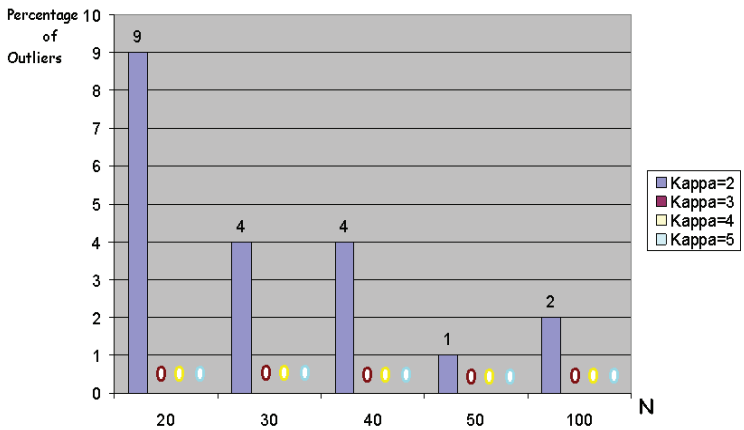

Figure 4. Results for percentage of outliers= 5%

Additionally, the same results for the dependent variable Y containing of %10 outlier observation for all of sample size can be observed from the simulation study. These results are shown in Figure 5.

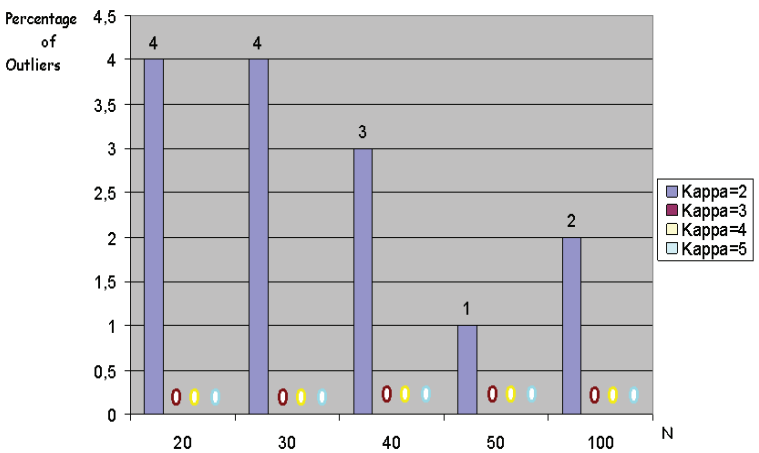

Figure 5. Results for percentage of outliers= 10%

Experiments with real and synthetic data sets show that the information criteria based on outlier detection method using GA in MLR models find the outlier automatically. 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 and the second is the scalability against the power of different Kappa coefficients for a given sample size and number of outliers. Figure 4 and 5 show the results of using GA to find diversity number of outliers on data set. One important observation from these figure was that the

GA based on information criteria can be found accurately outliers especially the kappa coefficient bigger than two. The GA was also run with Kappa value bigger than 5 and the same result are obtained for $\kappa = 2,\ 3, 4$ and 5. Therefore the results have been the same with a wide range of Kappa values.

4. DISCUSSION AND RESULT

In this paper, it is demonstrated that Bayesian information criteria and developed a GA for outlier detection in MLR models. The value of BIC' is calculated for each observation as a measure of the fitness of dependent variable in MLR models using GA. GA can simultaneously search in the solution space and find the outliers. The main advantage of this method is that on does not have to bother the distribution of the observed residuals, which has proved to be complicated for the simple reason that the estimated residuals do not have a constant variance. Nevertheless, exact distributions for appropriate test statistics based on these adjusted residuals become intractable (Barnett and Lewis, 1994). The simulation results are shown in Table 3, especially Kappa coefficient ($\kappa > 2$) gives true information about how many observations are found as outlier. Hence, it is confident to claim that the GA based on BIC' criteria is suitable for MLR models.

We are working on comparing other applications of the GA for detection of outliers in MLR models as the future work.

5. REFERENCES

Abe, N., Zadronzy, B., and Langford, J., 2006. Outlier detection by active learning. ACM. Proceedings of the 12th ACM SIGKDD International conference on Knowledge Discovery and Data Mining, 767-772, New York, USA.

Acuna, E., and Rodriguez, C., 2005. On detection of outliers and their effect in supervised classification, http://academic.uprm.edu/~eacuna/vene31.pdf, 30 April 2008

Amidan, B., Ferryman, and T., Cooley S., 2005. Data outlier detection using the Chebyshew theorem. IEEE Aerospace Conference Proceedings, IEEE, Piscataway NJ USA, 3814-3819.

Atkinson, A.C., 1986. Influential observations, high leverage points, and outliers in linear regression. Statistical Science, 1, 397-402.

Barnett, V., and Lewis, T., 1994. Outliers in statistical data. John Wiley and Sons, USA.

Ben-Gal I., 2005. Outlier detection.,131-146. In: Maimon O. and Rokach L., Data mining and knowledge discovery handbook. Springer, USA.

Bozdogan, H., 2004. Statistical data mining and knowledge discovery. Chapman and Hall/CRC, USA.

Breitenbach, M., and Grudic, G.Z., 2005. Clustering through ranking on manifolds. Proceedings of the 22nd International Conference on Machine Learning, 73-80, New York, USA.

Davies L., and Gather U., 1993. The identification of multiple outliers. Journal of the American Statistical Association, 88, (423), 797-801.

Fox, J., 1997. Applied regression analysis, linear models and related methods. Sage Publication, USA.

Goldberg, D.E., 1989. Genetic algorithms in search, optimization, and machine learning. Addison-Wesley, USA.

Hadi, A., 1986. Influential observations, high leverage points, and outliers in linear regression. Journal of the American Statistical Association, Statistical Science, 1 (3), 379-393

Hoaglin, D., and Tukey, J., 1983. Understanding robust and exploratory data analysis. John Wiley and Sons, Canada

Hoeting, J., Raftery, A.E., and Madigan, D., 1996. A method for simultaneous variable selection and outlier identification in linear regression. Computational Statistics and Data Analysis, 22, 251-270.

Ishibuchi, H., Nakashima, T., and Nii, M., 2001. Genetic algorithm based instance and feature selection. In: Liu, H., and Motoda, H., Instance selection and construction for data mining, Kluwer Academic.

Jann, A., 2000. Multiple change point detection with a genetic algorithm. Soft Computing , 4, 68-75.

Kullback, S., 1996. Information theory and statistics. Dover Publications, USA.

MacQueen, J.B., 1967. Some methods for classification and analysis of multivariate observations. IProceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, 1, 281-297.

Rothlauf, F., 2006. Representations for genetic and evolutionary algorithms. Springer, Netherlands.

Scott, D.W., 2005. Outlier detection and clustering by partial mixture modeling. Phsica-Verlag. In COMPSTAT 2004 Symposium, 453-465, Heidelberg.

Tolvi, J., 2004. Genetic algorithms for outlier detection and variable selection in linear regression models. Soft Computing, Springer, 527-533.

ÇOKLU REGRESYON MODELLERİNDE GENETİK ALGORİTMA VE BAYES BİLGİ KRİTERİ KULLANARAK SAPAN DEĞERLERİN BELİRLENMESİ

ÖZET

İstatistiksel modeller; özellikle regresyon modelleri, veri setlerinin önemli özelliklerinin anlaşılması ve ortaya çıkarılmasında en çok kullanılan araçlardandır. Bununla birlikte, gerçek hayatta birçok veri seti genellikle sapan değer olarak adlandırılan belirli miktardaki anormal değerler içerebilmektedir. Sapan değerlerin doğru bir şekilde tespit edilmesi, istatistiksel çözümlemelerde özelliklede regresyon modellerinde önemli bir rol oynar. Buna rağmen, birçok klasik istatistiksel modeller sapan değer içeren veri setlerine de uygulanmakta, nihayetinde de sonuçlar yanıltıcı olmaktadır. Sapan değerler, uygun olan çoklu regresyon modelinin belirlenmesini de güçleştirir.

Bu çalışmanın amacı, Genetik Algoritma (GA) ve Bayes Bilgi Kriteri (BIC) kullanarak sapan değer belirleme yöntemini tanımlamak ve algoritmayı gerçek ve benzetim verisi ile göstermektir. Genetik algoritmada BIC tabanlı uygunluk fonksiyonu kullanılmıştır. BIC değeri, veri için en uygun modeli göstermekte olup, bir veya daha çok sapan değerin varlığında regresyon modeli bu gözlemlerden olumsuz yönde etkilenecek ve daha büyük BIC değerli sonuçlar verecektir.

Anahtar Kelimeler: Bayes bilgi kriteri, Genetik algoritmalar, Çoklu regresyon modelleri, Sapan değer belirleme.