Variable selection with genetic algorithm and multivariate adaptive regression splines in the presence of multicollinearity

Betul Kan Kilinc [1, *], Baris Asikgil [2], Aydin Erar [2], Berna Yazici [1]

[1]Department of Statistics, Science Faculty, Anadolu University, Eskisehir, Turkey
[2]Department of Statistics, Faculty of Science and Letters, Mimar Sinan Fine Arts University, Istanbul, Turkey

ARTICLE INFO

ABSTRACT

In this paper, it is aimed to determine the true regressors explaining the dependent variable in multiple linear regression models and also to find the best model by using two different approaches in the presence of low, medium and high multicollinearity. These approaches compared in this study are genetic algorithm and multivariate adaptive regression splines. A comprehensive Monte Carlo experiment is performed in order to examine the performance of these approaches. This study exposes that nonparametric methods can be preferred for variable selection in order to obtain the best model when there is a multicollinearity problem in the small, medium or large data sets.

1. Introduction

A problem occurring frequently in multiple regression processes is the violation of the independency assumption among regressors. The problem of near linear dependencies is a particular form of data weakness which is called as collinearity (Belsley, 1991). Generally, researchers either ignore this problem or eliminate one or more variables causing multicollinearity. On the other hand, ignoring this problem causes incorrect findings.

Some general approaches combating multicollinearity are given as collecting additional data, model respecification (create a new variable by using a function including nearly linearly dependent ones) and using some biased estimation methods (ridge regression, principal component regression etc.) Two of the corrective techniques dealing with multicollinearity are variable selection known at the same time as one stage of regression modeling and nonparametric estimation methods used frequently in recent years.

Selecting necessary variables to the model and eliminating insignificant variables from the model are known as variable selection or the best subset regression model selection. In a general regression process, an important stage for creating the best model is variable selection after examining residuals and regression assumptions. This stage is applied not only for creating a good model but also dealing with multicollinearity. There are two main reasons for selecting variables. One of them is the practical reasons including benefits in applications and economic sense. When there are more regressors in a model, the costs of data collection and model maintenance stand unreasonable. The other reason is known as theoretical reasons which mean that estimations and predictions should have some statistical properties in creating a model (Montgomery et al., 2012).

There has been a fast improvement in terms of modern regression and classification methods in recent years. There are many studies about smoothers, projection pursuit regression, additive models, genetic algorithm (GA), multivariate adaptive regression splines (MARS). Some of these methods make the ordinary regression assumptions flexible.

GA, first introduced by Holland (1975) and researched on different problems extensively by Goldberg (1989), is an evolutionary search algorithm based on the principles of natural selection and survival of the best. This method deals with a population of individuals which are defined as binary strings in the parameter space. It is tried to obtain solutions based on the fitness of the individuals. A new population is created from fitter individuals. The fitness of each individual in each generation is determined by the value of the objective function. In this method, there is a pressure

for the fitter individuals to be joined into the population (Pan et al., 1995).

There are two adaptive smoothing methods (Pittman, 2002). First of them is a group including adaptive splines and kernel methods used with a local adaptive smoothing parameter. For instance, variable bandwidth given by Fan and Gijbels (1995) is known as a kernel method. Second of them includes regression splines algorithms based on the assumption of representing the continuous function at low order derivatives with continuous piecewise polynomials approximately. Candidate models represent the space $S_{\{m,t\}}$ including m^{th} order spline functions, where t denotes some knots. The knot sequence is selected adaptively in curve fitting. In the first trials of adaptive selection, knots were added and removed iteratively like the algorithms given by De Boor (1978), but no optimal knot place was found. This algorithm only gave an optimal distribution for knots. Schwetlick and Schütze (1995) obtained the places of knots by using nonlinear optimization method when the number of knots was given. Also, Lindstrom (1999) defined a nonlinear optimization algorithm based on free knot spline modeling. A different algorithm, first introduced by Friedman (1991), is a strategy of adding/removing knots by using linear splines stepwisely. This algorithm is called MARS, a common method in recent years, also examines the linear and nonlinear correlations among variables. In MARS, one term is added to the model selection criterion, generalized cross validation (GCV), given by Wahba and Craven (1978) in order to adjust adaptive form of knot selection. GA is preferable because of the necessity of using stochastic optimization methods for knot selection Pittman (2002). Manela et al. (1993) used GA for available spline order and penalized least squares splines. Moreover, Rogers (1991) combined a modified GA with MARS. In this paper, the performance of GA and MARS is compared for obtaining the best and true model. The rest of this paper is organized as follows. In Section 2, the model and the methodology are given. In Section 3, a comprehensive Monte Carlo experiment is presented in order to make comparisons. Finally, the conclusion is given in Section 4.

2. Model and methodology

Let x_j, j = 1, 2, ..., k be regressors, y be the response and ε be the random error term. A multiple linear regression model is defined as (Eq. 1):

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i \tag{1}$$

where, β are unknown parameters and i=1,...n. To form this model by using the ordinary least squares (OLS) method classical assumptions related to regression analysis have to be satisfied. Also, the model given by (1) is in matrix notation as follows (Eq. 2):

$$Y = X\beta + \varepsilon \tag{2}$$

where, $\boldsymbol{Y}$ is an (n x 1) vector of responses, $\boldsymbol{X}$ is an (n x p) matrix of regressors, β is a (p x 1) vector of unknown parameters and $\boldsymbol{\varepsilon}$ is an (n x 1) vector of random errors for given $p = k + 1$.

In this section, the existing theory and operation of GA and MARS are outlined, respectively.

2.1. Genetic algorithm

Each parameter of a model is coded as a fixed-length string of binary numbers (genes) in GA. Genes is given as a cascade form called a chromosome. The initial values of parameters are determined randomly. Therefore, N chromosomes are converted to random binary strings at the beginning of the estimation process (Yao and Sethares, 1994). A binary string should be converted to base ten for decoding this string into a real number. The real values of parameters can be calculated in this way. If the number of parameters is greater than one, the strings of each parameter can be joined into a single string and operations can be made on this string. The operations used in GA are selection, crossover and mutation, respectively. They are called as genetic operators.

The selection operator is used for selecting chromosomes in the population. This operation is made according to the fitness level of chromosomes. There are many different selection strategies. Some of them are fitness-proportionate selection, rank selection and tournament selection. The originally used fitness-proportionate selection strategy is based on the division of the individual's fitness by the average fitness of the population (Mitchell, 1996). The rank selection strategy alternatively used for preventing too quick convergence is based on the rank of the fitness values of the objective function of chromosomes in the population. Let $S(\beta)$ be an objective function defined as (Eq. 3):

$$S(\beta) = (Y - f(X,\beta))'(Y - f(X,\beta)) \tag{3}$$

$S(\beta_i, u)$ is given as the fitness value of the chromosome β_i at generation u. The probability of selecting the chromosome β_i into the next generation of the population as a parent is given by (Eq. 4):

$$r(\beta_i, u) = \frac{2m_i(u)}{N(N+1)} \tag{4}$$

where, $m_i(u)$ denotes the rank of $S(\beta_i, u)$ by sorting descendingly and N is the number of chromosomes in the population (Pan et al., 995). Another alternatively used and computationally more efficient strategy is given as tournament selection. It is similar to rank selection in terms of selection pressure (Mitchell, 1996).

The crossover operator, main distinguishing feature of GA, provides exchanging randomly selected parts of two individuals in order to form two offspring which are evaluated in the next

generation of the population (Mitchell, 1996). If $P_1 = 0010100110$ and $P_2 = 1100011011$ and the crossover point is 5, then the two new offspring will become $P_1^* = 0010111011$ and $P_2^* = 1100000110$.

The mutation operator provides randomly changes on the gene from 0 to 1 or 1 to 0 with a small probability. It aims to inhibit premature convergence. If the probability of mutation is larger, then the convergence rate will be faster with a large steady state error (Yao and Sethares, 1994).

As a summary, the algorithm of GA can be given as follows:

1. Choose randomly initial population (t).
2. Determine the fitness of population (t).
3. Repeat the steps given below until the best individual is good enough.
4. Select parents from population (t).
5. Apply crossover on parents for creating population (t+1).
6. Apply mutation on population (t+1).
7. Determine the fitness of population (t+1).

2.2. Multivariate Adaptive Regression Splines

MARS is a nonparametric approximation of the relationship between a response and a set of regressors in reflected pairs of simple linear splines. The reflected pairs take the form of (Eq. 5):

$$(x - t)_+ = \begin{cases} x - t, & if \ x > t \\ 0, & else \end{cases} \qquad (5)$$

where, the subscript "+" means the argument is a truncated power function. The MARS model has the form (Eq. 6):

$$y = \beta_0 + \sum_{m=1}^{M} \beta_m h_m(x) \qquad (6)$$

where, $h_m(x)$ are basis functions and $\beta_0, \dots, \beta_m$ the unknown parameters for i=1,…, M. Once the basis functions (BFs) are investigated then $\beta_0, \dots, \beta_m$ are estimated by ordinary least squares method. The basis functions can be represented by (Eq. 7):

$$h_m(\boldsymbol{x}) = \prod_{k=1}^{K_m} [S_{k,m}(\boldsymbol{x}_{v(k,m)} - t_{k,m})]_+^q \qquad (7)$$

Where, K_m is the number of variables (interaction order) in the mth basis expansion. $\boldsymbol{x}_{v(k,m)}$ is the v th variable, $1 \le v(k, m) \le n$, $t_{k,m}$ is a knot on each of the corresponding variable.

MARS algorithm adaptively selects the basis function set by two iterative approaches: forward and backward selection. It uses the residual squared error in iterations to compare the partition points. The criterion used to set the final model is a modified generalized cross validation (GCV) of the first proposed one by Wahba and Craven (1978). The difference between the two criteria comes from a penalty term that reflects the complexity of the model in MARS by prohibiting a model with many variables that decreases only slightly the residual errors (Friedman, 1991).

MARS approach allows a nonlinear relationship over different intervals of x (the vector of regressors) for modeling y (Sephton, 2001). The relationship between the regressors and the response is fitted by basis functions that are basically splines. The main idea behind MARS is to explore the relationship by splitting the regressors over its region into several intervals and transform the original regressors over the intervals. MARS fits a spline based model in each interval. BFs include knot locations relating to the regressors. BFs can be a single or multivariable interaction term. The final model is a combination of BFs. Model selection is accomplished using GCV criterion given in Eq. 8:

$$GCV = \frac{1}{N} \sum_{i=1}^{N} [y_i - f_M(\boldsymbol{x})]^2 / [1 - \frac{C(M)}{N}]^2 \qquad (8)$$

here, $f_M(\boldsymbol{x})$ is the fitted response value with the observed regressors, M represents the maximum number of observations and $C(M)$ is the complexity parameter to the corresponding model. The final MARS model is the one with the smallest GCV value and the largest R_{adj}^2 value (Friedman, 1991; Hastie et al., 2001).

3. Monte Carlo simulation study

In this section, a simulation study was conducted in order to compare GA and MARS performance when the level of multicollinearity is low, moderate, and high. For this aim, two programs of the variable selection were generated for both MARS (Milborrow, 2011) and GA by using R Software (RDC Team, 2006) and MATLAB2010a, respectively. The percentages of selecting the best subset among others were summarized and the results were interpreted.

3.1. Design of the model

A comprehensive Monte Carlo experiment designed by using the popular asphalt data (Gorman and Toman, 1966) in the literature was performed in order to examine the performance of these approaches. The number of observations was considered as n = (50; 100; 500) regarding to the small, medium, and large size of samples. Two types of Press statistics were used in order to measure the discrepancy between each observation and its prediction. To illustrate how to compute them we might start for the data sets including 50 observations. $\sum_{i=1}^{50}(y_i - \hat{y}_i)^2$ was used to calculate Press1 statistic. After each data set was increased about extra 15 observations, $\sum_{i=51}^{65}(y_i - \hat{y}_i)^2$ was used to calculate Press2 statistic. For the data sets including 100 and 500 observations, the same computations were considered for Press statistics. When additional data set that was used for calculation of Press2 values and 100 times of repetition were considered, the total number of observations reached n = (6500, 11500, 51500) observation. $y = \beta_0 + \beta_1 x_1 + \dots + \beta_6 x_6 + \varepsilon$ was used for the corresponding model. Here, the random

error term ε was generated from the normal distribution with mean 0 and standard deviation σ = (0.05,0.1,1) for all sample sizes. True model parameters were considered as β= (-1.02079, -0.64649, 0.55547, 0, 0, 0, 0.24479)'. The regressors were generated from normal distribution as well. In addition to that, x_4 was considered as dummy variable taking the values of 0 and 1. This experiment was repeated for 100 times for all sample sizes. In order to generate multicollinearity in the data set, two of regressors x_2 and x_3 were used. Hence, the multicollinearity yielded from x_5 = $2x_2$ + $0.5x_3$ + u where u was generated from the normal distribution with mean 0 and standard deviation h =(0.01,0.02,0.03).

4. Results and discussion

In the simulation study, the combinations of regressors (x_1, x_2), (x_1, x_6), (x_1, x_2, x_6), (x_1, x_5, x_6), (x_1, x_2, x_4, x_6), and $(x_1, x_2, x_4, x_5, x_6)$ were taken into account as the best subsets among the models. However, due to the limited size of the paper, only the best subset of x_1, x_2, x_6 namely (1,2,6) set was given and then interpreted. Hence, the frequencies of this selection were recorded in Tables 1-3. GA and

MARS were applied to 27 different scenarios with different values of σ =(0.05,0.1,1) and h=(0.01,0.02,0.03) respectively. Here, h=0.01 indicates high multicollinearity whereas h=0.03 indicates low multicollinearity. GA and MARS were devised for variable selection by using a computer based program code. Specifically, the frequencies of the best model (counts) that were able to select the best subset of regressors and Press values in GA and MARS were obtained from the code for interpretation. For this aim, the selection of the best subset of regressors was coded as 1 and 0 otherwise. Hence the total number of 1' s was the total count of the selection of the value 1 among repetitions. The Press2 values were estimated using a data set including 15 extra observations that were already generated at the beginning of the data set construction. Hence, n=15x100=1500 extra observations were used for validation of observations for each data set of n=6500, n=11500, and n=51500 of observations. Although Press1 and Press2 values were computed for two methods, only Press1 values and counts were summarized for both GA and MARS in Tables 1-3 to prevent duplication in explanations.

Table 1: Comparison of GA and MARS when high multicollinearity occurred

h = 0.01		σ = 0.05		σ = 0.1		σ = 1	
		GA	MARS	GA	MARS	GA	MARS
n = 50	Press	0.112	0.103	0.431	0.426	34.820	---
	Count	57	72	44	68	1	0
n = 100	Press	0.240	0.228	0.956	0.947	89.140	---
	Count	67	81	60	80	14	0
n = 500	Press	1.242	1.230	4.975	4.935	502.851	489.068
	Count	67	98	61	94	49	9

Among the data sets of n=50 observations given in Table 1 where high multicollinearity occurred, GA selected the best subset of (1,2,6) for 57 times among 100 experiments whereas MARS selected the best subset of regressors (1,2,6) for 72 times when σ = 0.05.

On the other hand, GA selected the best subset of regressors (1,2,6) for 44 times whereas MARS selected the best subset of (1,2,6) for 68 times when σ =0.1. The Press value was computed as 0.103 in MARS whereas it was computed as 0.112 in GA. The Press value increased from 0.103 in the change of the σ (=0.1) to 0.426 in MARS whereas it was also increased from 0.112 to 0.431 in GA. Next, GA selected the best subset of regressors (1,2,6) for 67 times among 100 experiments whereas the best subset of regressors (1,2,6) was selected by MARS for 81 times in the data sets of n=100 observations where high multicollinearty exits when σ = 0.05. On the other hand, GA selected the best subset of regressors (1,2,6) for 60 times whereas MARS selected the best subset of regressors of (1,2,6) for 80 times when σ = 0.1. The Press value was computed as 0.228 in MARS whereas it was computed as 0.240 in GA when σ = 0.05. The Press value increased from 0.2281 to 0.947 in MARS in the

change of the σ(=0.1) whereas it was also increased from 0.240 to 0.956 in GA. Note that MARS could not compute Press values for n=(50, 100) observations when σ=1 although it could compute for n=500 observations. The frequency of the selection of the best subset was very low (=9) whereas the Press value was quite large (=489.068) in MARS when n = 500.

In Table 2, the data sets of n=(50,100,500) observations where medium multicollinearity existed were compared in terms of Press and count values for MARS and GA. The results show that MARS selected the best subset of regressors of (1,2, 6) more than GA for each sample size when σ = (0.05,0.1).

In Table 3, the data sets of n=(50,100,500) observations where low multicollinearity existed were compared in terms of Press and count values for MARS and GA. As it can be seen from the results that the Press values and counts obtained by MARS were smaller than the ones obtained by GA at σ = (0.05, 0.1). On the other hand, MARS could not compute the Press values for n=(50, 100) at σ = 1 as given in the previoustables. The Press value obtained by GA was lower (=496.423) for n=500 than the ones in the previous multicollinearity levels.

In Table 4, AIC and GCV values obtained by GA and MARS were summarized for 27 scenarios, respectively. According to the results, the larger the sample size, the better AIC values obtained by GA. Similarly, the larger the sample size, the better GCV values obtained by MARS.

Table 2: Comparison of GA and MARS when medium multicollinearity occurred

h = 0.02		σ = 0.05		σ = 0.1		σ = 1	
		GA	MARS	GA	MARS	GA	MARS
n = 50	Press	0.106	0.104	0.429	0.419	39.945	---
	Count	58	75	52	66	2	0
n = 100	Press	0.238	0.236	0.956	0.940	91.816	---
	Count	61	85	63	71	11	0
n = 500	Press	1.230	1.229	4.960	4.885	491.964	466.984
	Count	63	100	68	95	49	6

Table 3: Comparison of GA and MARS when low multicollinearity occurred

h = 0.03		σ = 0.05		σ = 0.1		σ = 1	
		GA	MARS	GA	MARS	GA	MARS
n = 50	Press	0.112	0.105	0.446	0.437	33.609	---
	Count	58	82	43	68	4	0
n = 100	Press	0.233	0.223	0.898	0.882	96.6145	---
	Count	60	85	58	80	13	0
n = 500	Press	1.237	1.234	5.007	4.942	496.423	480.496
	Count	62	100	62	98	37	4

Table 4: AIC and GCV values obtained by GA and MARS

h	GA					
	n; σ	AIC	n; σ	AIC	n; σ	AIC
0.01	50; 0.05	-154.264	100; 0.05	-310.022	500; 0.05	-1571.295
	50; 0.1	-87.307	100; 0.1	-172.690	500; 0.1	-877.284
	50; 1	134.014	100; 1	280.549	500; 1	1429.372
0.02	50; 0.05	-157.116	100; 0.05	-311.775	500; 0.05	-1576.024
	50; 0.1	-87.398	100; 0.1	-172.432	500; 0.1	-878.923
	50; 1	139.051	100; 1	282.452	500; 1	1418.199
0.03	50; 0.05	-154.483	100; 0.05	-313.956	500; 0.05	-1573.117
	50; 0.1	-85.284	100; 0.1	-178.867	500; 0.1	-874.568
	50; 1	128.593	100; 1	288.080	500; 1	1423.096
h	MARS					
	n; σ	GCV	n; σ	GCV	n; σ	GCV
0.01	50; 0.05	0.003	100; 0.05	0.003	500; 0.05	0.002
	50; 0.1	0.014	100; 0.1	0.012	500; 0.1	0.010
	50; 1	---	100; 1	---	500; 1	1.016
0.02	50; 0.05	0.003	100; 0.05	0.003	500; 0.05	0.002
	50; 0.1	0.014	100; 0.1	0.012	500; 0.1	0.010
	50; 1	---	100; 1	---	500; 1	0.973
0.03	50; 0.05	0.003	100; 0.05	0.002	500; 0.05	0.002
	50; 0.1	0.015	100; 0.1	0.011	500; 0.1	0.010
	50; 1	---	100; 1	---	500; 1	0.998

5. Conclusion

In this study, a comprehensive experiment designed by using the popular methods GA and MARS was applied in order to examine the performance of these approaches. The number of the sample sizes was considered from small to large data sets. It can be concluded that Press values obtained by MARS are smaller than the ones in GA. The frequency of the selection of the best subset is better for MARS than GA except for σ = 1. It can be also seen that the larger sample sizes, the more success in the selection of the best subset for MARS and GA. Finally, this study exposes that nonparametric methods can be preferred for variable selection in order to obtain the best model when there is a multicollinearity problem in the small, medium or large data sets.

Acknowledgment

References

Belsley DA (1991). Conditioning diagnostics collinearity and weak data in regression. John Wiley and Sons, New York, USA.

De Boor C (1978). A Practical guide to splines. Springer-Verlag, New York, USA.

Fan J and Gijbels I (1995). Data-driven bandwidth selection in local polynomial fitting: variable

bandwidth and spatial adaptation. Journal of the Royal Statistical Society Series B, 57(2): 371-394.

Friedman JH (1991). Multivariate adaptive regression splines. The Annals of Statistics, 19(1): 1-67.

Goldberg DE (1989). Genetic algorithms in search, optimization and machine learning. Addison-Wesley, Reading, Massachusetts, USA.

Gorman JW and Toman RJ (1966). Selection of variables for fitting equations to data. Technometrics, 8(1): 27-51.

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

Holland J (1975). Adaptation in natural and artificial systems. University of Michigan Press, Ann Arbor, USA.

Lindstrom MJ (1999). Penalized estimation of free-knot splines. Journal of Computational and Graphical Statistics, 8(2): 333-352.

Manela M, Thornhill N and Campbell J (1993). Fitting spline functions to noisy data using a genetic algorithm. In Proceedings of the 5th International Conference on Genetic Algorithms, Morgan Kaufmann Publishers Inc., San Francisco, USA: 549-556

Milborrow S (2011). Derived from mda: MARS by Trevor Hastie and Rob Tibshirani. Earth: Multivariate Adaptive Regression Spline Models. R package version 3.2-0. Available online at: http://CRAN.R-project.org/package=earth.

Mitchell M (1996). An introduction to genetic algorithms. MIT Press, Cambridge, UK.

Montgomery DC, Peck EA and Vining GG (2012). Introduction to linear regression analysis. John Wiley and Sons, New Jersey, USA.

Pan Z, Chen Y, Kang L and Zhang Y (1995). Parameter estimation by genetic algorithms for nonlinear regression. In the Proceedings of International Conference on Optimization Technique and Applications '95, 2. World Scientific, Singapore: 946–953.

Pittman J (2002). Adaptive splines and genetic algorithms. Journal of Computational and Graphical Statistics, 11(3): 615-638.

RDC Team (R Development Core Team) (2006). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. Available online at: http://www.R-project.org

Rogers D (1991). G/SPLINES: A hybrid of Friedman's multivariate adaptive regression splines (MARS) algorithm with Holland's genetic algorithm. In the Proceedings of the Fourth International Conference on Genetic Algorithms. San Diego, USA.

Schwetlick H and Schütze T (1995). Least squares approximation by splines with free knots. BIT Numerical Mathematics, 35(3):361-384.

Sephton P (2001). Forecasting recessions: Can we do better on MARS?. Review Federal Reserve Bank of St. Louis, 83(2): 39-49.

Wahba G and Craven P (1978). Smoothing noisy data with spline functions. Numerische Mathematik, 31(4): 371-403.

Yao L and Sethares W A (1994). Nonlinear parameter estimation via the genetic algorithm. IEEE Transactions on Signal Processing, 42(4): 927-935.