Regression Model Selection Using Genetic Algorithms

SANDRA PATERLINI[a] and TOMMASO MINERVA[b]

[a] Dept. of Political Economics, Univ. of Modena and Reggio E., viale Berengario 51, 41100 Modena
[b] Dept. of Social Sciences, Univ. of Modena and Reggio E., viale Allegri 9, 42100 Reggio E.
ITALY

sandra.paterlini@unimore.it - tommaso.minerva@unimore.it

Abstract: The selection of independent variables in a regression model is often a challenging problem. Ideally, one would like to obtain the most adequate regression model. This task can be tackled with techniques such as expert based selection, stepwise regression and stochastic search heuristics, such as genetic algorithms (GA). In this study, we investigate the performance of two GAs for regressors selection (GARS) and for regressors selection with transformation of the regressors (GARST). We compare the performance with stepwise regression for the "Fat Measurement" and the "Cholesterol Measurement" datasets and use the AIC, BIC and SIC statistical criteria to quantify the adequacy of the models. The results for GARS are superior for all statistical criteria compared to both forward and backward stepwise regression, but not always when R^2 and RMSE statistics are considered. GARST turns out to be even better compared to GARS as variable transformations help to improve results further. Moreover, the type of transformations revealed the relationships between dependent and independent variables.

Key-words: regression model, genetic algorithms, stepwise techniques, regressors' selection and transformation.

1 Introduction

Regression analysis is a well-established method in data analysis with applications in various fields. Its main purpose is to determine the relationship between a so-called dependent variable and one or more independent variables. Different approaches to regression model selection have been proposed, such as expert-based selection, stepwise regression and stochastic search algorithms. The selection of the most *adequate* regression model can be stated as an optimization problem with the objective to select those independent variables that maximize the adequacy of the model according to a statistical criterion. Different statistical criteria, such as AIC, BIC or SIC (see section 2.2), have been proposed in the literature. Moreover, optimization can be used to determine the most appropriate transformation of these variables to obtain optimal adequacy.

A common approach is to use stepwise regression, which works in the following way: in each step all regression models are built and evaluated that differ from the current best regression model in just one variable. If the best of these models has a better adequacy than the current model, it becomes the starting point of the next step and the process is repeated, or otherwise the algorithm terminates. This approach is a local search process, and its main drawback is that it ultimately converges to local optima. A promising alternative to tackle optimization problems with local optima is to use genetic algorithms, since they explore the search space simultaneously by a population of candidate solution in which solutions compete and recombine. Apart from selecting the variables for a regression model, GAs can also be used to determine the most appropriate transformations of the independent variables.

In this paper, we consider two genetic algorithms for regression modelling. The first algorithm called GARS tackles the issue of variable selection. Based on this approach we developed a new genetic algorithm called GARST that selects the variables and also determines the most appropriate mathematical transformations to obtain optimal adequacy.

The paper is organised as follows: Section 2 gives a formal definition of the regression model selection problem, statistical criteria and algorithmic choices. Section 3 briefly introduces the main concepts in genetic algorithms and describes the two genetic algorithms for regression modelling. Section 4 describes the experimental set up and the implementation details. Section 5 reports and discusses the empirical results from GARS and

GARST, in comparison with the complete model and stepwise techniques, in the analysis of two real world datasets and section 6 concludes our study.

2 Regression Model Selection

2.1 The Model Selection Problem

Let $X \equiv \{X_1, X_2, ..., X_m\}$ the set of m independent variables (with n observations) and Y the dependent variable in a multivariate regression model. Let's suppose that the model

$$(1) \qquad Y = \beta_0 + \beta_1 \tilde{X}_1 + \beta_2 \tilde{X}_2 + ... + \beta_p \tilde{X}_p + \varepsilon$$

explains the relationship between the dependent and independent variables, where $\tilde{X} \subseteq X$ is the set of the $p \leq m$ independent variables chosen as regressors and $\mathrm{B} \equiv \{\beta_0, \beta_1, ..., \beta_p\}$ is the parameters set. If **B** is estimated as $\hat{\mathrm{B}} \equiv \{\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_p\}$ by using the Ordinary Least Square method (OLS), the remaining main task is to choose which independent variables should be included in $\tilde{X} \subseteq X$. This task is an optimisation problem, where the objective is to select $\tilde{X} \subseteq X$ such that the estimated model:

$$(2) \qquad \hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 \tilde{X}_1 + \hat{\beta}_2 \tilde{X}_2 + ... + \hat{\beta}_p \tilde{X}_p$$

has optimal adequacy with respect to some statistical criteria (see section 2.2).

Apart from selecting the best subset $\tilde{X} \subseteq X$ of independent variables, another task could be to determine which mathematical transformations (e.g.: exponential, logarithmic) should be applied to the independent variables in order to improve the adequacy of the model (Cook and Weisberg 1993). In this case, the problem consists of selecting the subset $\tilde{X} \subseteq X$ and the mathematical transformations $f : \tilde{X}_i \rightarrow T_i(\tilde{X}_i)$ such that the model $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 T_1(\tilde{X}_1) + \hat{\beta}_2 T_2(\tilde{X}_2) + ... + \hat{\beta}_p T_p(\tilde{X}_p)$ is optimal with respect to the used statistical criterion.

2.2 Statistical Model Selection Criteria

Different criteria could be chosen to quantify the degree of optimality of a regression model. In this work, we focus here on three statistical criteria:

- The *Asymptotic Information Criterion (AIC)* (Akaike, 1973):

$$(3) \qquad AIC(p) = n \log(S_p^2) + 2p$$

where n is the number of observations, p is the number of independent variables in the regression model, $S_p^2 = \frac{\sum_{i=1}^{n} (E_i)^2}{n - p - 1}$ is the variance of the residual when the model with p independent variables is considered, and E_i are the residuals.

The AIC provides an estimate of the "distance" between the estimated model and the unknown mechanism behind the data. The lower the AIC value, the better the model. It tends to overestimate the number of parameters that should be considered in the optimal model.

- The *Bayesian Information Criterion (BIC)* (Akaike, 1978)

$$(4) \quad BIC(p) = (n - p) \log\left(\frac{nS_p^2}{n - p}\right) + p \log\left(n \frac{S_0^2 - S_p^2}{p}\right)$$

where n is the number of observations, p is the number of independent variables in the regression model, S_p^2 is the variance of the residual when the model with p independent variables is considered and S_0^2 is the variance of the n observations of the dependent variable. The lower the BIC value, the better the model. In comparison with the AIC, it aims to consider the variance reduction by estimating a model with p covariates.

- The *Schwarz Information Criterion (SIC)* (Schwartz, 1978):

$$(5) \qquad SIC(p) = \log(S_p^2) + \frac{p}{n} \log(n)$$

where n is the number of observations, p is the number of independent variables in the regression model, S_p^2 is the variance of the residual when the model with p independent variables is considered. The lower the SIC value, the better the model.

2.3 Techniques for regression model selection

Different approaches to regression model selection have been proposed, such as expert-based selection,

classical stepwise approaches and stochastic search algorithms.
In expert-based selection, an expert tries to identify (manually) which variables should be included, which is a process that is often based on a blend of his past experience and knowledge of the problem and trial-and-error. For this, he investigates different models and then selects the one that he considers to be the most *adequate*. The main drawback of expert-based selection is that it is subjective and not guaranteed to yield regression models with optimal adequacy.
In contrast, classical stepwise regression techniques try to find a model that has optimal adequacy with respect to a statistical criterion by considering alternative models. At each step, all models are tested which differ from the currently best model by including/excluding of one independent variable. If the best of these models turns out to be better, it is used as the current best model and the search is continued from this point. The algorithm stops when no better model can be found that differs in one variable. The so-called *backward* stepwise regression method starts from a complete model and with each iteration reduces the number of variables, whereas the *forward* stepwise regression method starts by considering all the models with just one independent variable. Stepwise regression is a local search process (steepest-descent/ascent local search) that iteratively tries to refine the current solution by proceeding to its best neighbour if it is better, or terminates otherwise. Ultimately each single run of this process converges to a local optimum. Compared to that stochastic search heuristics can escape local optima.
Perhaps the most promising approach to deal with multiple local optima in non-linear optimization problems is to use population-based stochastic search heuristics, such as genetic algorithms (see section 3), since they explore the search space simultaneously by a population of candidate solution in which solutions compete and recombine. Apart from selecting the most appropriate variables in a regression model, GAs can also easily be used to determine the most appropriate transformations of the independent variables.

In this paper, we present two different genetic algorithms for regression modelling. The GARS (see section 3.1) tries to select the model variables that allow the optimal adequacy, whereas the GARST (see section 3.2) not only identifies the model variables, but also determines the most appropriate mathematical transformations for these variables to obtain optimal adequacy.

3 Genetic Algorithms for Regression Models

Genetic algorithms have been used in many different fields including statistics (see for a review Chatterjee et al 1996) for a variety of problems, such as time series analysis (Baragona et al 2004), AR/ARMA model selection (Minerva and Poli 2001), outliers detection (Baragona et al. 2001), graphical model selection (Poli and Roverato 1998, Roverato and Paterlini 2004), and clustering (Paterlini and Minerva 2003).

In this work, we consider two genetic algorithms: *Genetic Algorithm for Regressors' Selection* (GARS) and *Genetic Algorithm for Regressors' Selection and Transformation* (GARST).
GARS (see section 3.1), which has been proposed by Minerva and Paterlini (2002), aims to select which independent variables should be considered in the optimal linear regression model, where the optimality is determined with respect to the AIC, BIC or SIC criteria.
GARST (see section 3.2), which we propose here for the first time, aims not only to select the independent variables to be included in the regression model, but also to determine how such variables should be mathematically transformed by functions such as power, logarithm and exponential. Transforming the dependent variables and then estimating the parameters can in fact help improving the goodness of the model and pointing out the existence of nonlinear relationships. On the other hand, transformations may lead to complex models which the researcher could find difficult to interpret.

3.1 Genetic Algorithm for Regressors' Selection (GARS)

GARS (*Genetic Algorithm for Regressors' Selection*) uses binary encoding to identify which independent variables should be included in the model. No transformation is applied to the independent variables before including them.
Each GA individual consists of a string of m binary cells: if the *i-th* cell ($i=1,...,m$) has value 1, then X_i is included in the model, otherwise not.
Every candidate solution is then evaluated with respect to a fitness function. The AIC, BIC and SIC criteria (see section 2.2) have been considered as possible fitness functions. After randomly initialising the population and evaluating the population with respect to the chosen fitness function, the population

is evolved through generations using stochastic uniform sampling selection scheme, single point crossover with p_c=0.8, uniform mutation with p_m=1/NBITS and direct reinsertion of the best recorded candidate solution. The algorithm stops when the population has been evolved for MAXGEN generations. The best solution is then reported.
Section 5 reports the empirical results in the analysis of the Body Fat Measurement and Cholesterol Measurement datasets in comparison with the complete model and the backward and forward stepwise methods.

3.2 Genetic Encoding for Regressors' Selection and Trasformation

GARST (*Genetic Algorithm for Regressors' Selection and Transformation*) uses binary encoding. Binary strings are then converted to integers values that determine whether a variable should be included and which transformation shall be applied.
For each of the *m* independent variables, we encode two integers t_i and exp_i. Thus each candidate solution consists of a binary encoding of *2m* integer parameters.
Each t_i determines if variable *i* shall be included and, if yes, whether the power, logarithm, or exponential function shall be applied for transformation, whereas exp_i specifies the power exponent of the transformation if variable *i* is included. More specifically, $t_i \in \{1,2,3,4\}$, $exp_i \in \{-6,-4,-2,-1,1,2,4,6\}$)is such that:

if $t_i = 1$: $T_i(X_i)=0$, variable not included,
if $t_i = 2$: $T_i(X_i)= X_i^{exp_i/2}$
if $t_i = 3$: $T_i(X_i) = \ln(X_i)^{exp_i/2}$
if $t_i = 4$: $T_i(X_i) = (e^{X_i})^{exp_i/2}$

Note that such an encoding allows specifying all the models that GARS can explore. For example the string with all integer values equal to 2, selects the linear regression model with all the *m* independent variables without any transformation. The size of the search space is then $(4*8)^m$.
The algorithm starts by randomly generating the population. Every GA individual is then evaluated with respect to the AIC, BIC or SIC criterion. The population is evolved using a stochastic uniform sampling selection scheme, single point crossover with p_c=0.8, uniform bitflip mutation with p_m=1/NBITS, reinsertion of the best recorded individual (elite of size 1). The algorithm terminates when the population has been evolved for MAXGEN generations and reports best found solution.

4 Experimental Set-Up

4.1 Real World Data

GARS and GARST have been tested considering two real world datasets: the *"Body Fat Measurement"* dataset (Johnson 1996 - FAT) and the *"Cholesterol Measurement"* dataset (Purdie et al. 1992 - CHOLES).
The *"Body Fat Measurement"* dataset (*n=252, m=16*) consists of 252 observations of 16 independent variables: age, weight, height, density, net body weight and ten measurements of body circumferences of an individual. The dependent variable is the percentage of fat in the body.
The *"Cholesterol Measurement"* dataset (*n=264, m=21*) consists of 264 observations of 21 independent variables which are measures of optical absorption of blood samples with different frequencies. The dependent variable is the level of cholesterol in the blood.
The empirical results for GARS have been compared with the ones obtained for the complete model, backward and forward stepwise techniques, and GARST. AIC, BIC and SIC criteria have been considered with forecasting aims in order to select the most appropriate model. Each sample dataset has been partitioned in three disjoint (consecutive) subsets: a training set (first 40% of the data), a validation set (following 40% of the data) and a test set (remaining 20% of the data).
The regression parameters $\hat{B} = \{\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_p\}$ were estimated using the training set while the AIC, BIC and SIC values were computed for the validation set in order to improve the robustness of the variable selection mechanism. R^2 and the Residual Mean Squared Error (RMSE) statistics were computed with respect to the test set. The R^2 statistics refer to a simple linear regression model with intercept, where the dependent variable is the *Y* on the test set and the independent variable is the estimated $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1\tilde{X}_1 + \hat{\beta}_2\tilde{X}_2 + ... + \hat{\beta}_p\tilde{X}_p$ on the test set with $\hat{B}$ estimated on the training set. The Residual Mean Squared Error has been computed as $RMSE = \sqrt{\sum_{i=1}^{k}(Y_i - \hat{Y}_i)^2 / k}$, where *k* is the length of the test set.

Type of Model Selection	Stat. Criterion	Best Stat. Criterion Value	R^2	RMSE	Number of Independent Variables	Selected Independent Variables
Complete Model	AIC	1.042	0.956	0.264	16	1,2,...,16
	BIC	1.941	0.956	0.264	16	1,2,...,16
	SIC	1.456	0.956	0.264	16	1,2,...,16
Backward Stepwise Regression	AIC	0.826	0.960	0.253	8	2,3,5,6,8,9,11,15
	BIC	1.24	0.956	0.266	7	2 ,3,5,6,8,11,15
	SIC	1.012	0.956	0.266	7	2,3,5,6,8,11,15
Forward Stepwise Regression	AIC	0.693	*0.970*	*0.219*	5	3,4,6,9,11
	BIC	0.866	0.966	0.231	3	3,6,9
	SIC	0.801	0.967	0.229	4	3,4,6,9
GARS	AIC	**0.632**	0.967	0.231	5	3,5,6,9,14
	BIC	**0.803**	0.964	0.241	2	3,6
	SIC	**0.762**	0.967	0.230	5	3,5,6,9,14

Table 1: "Body Fat Measurement Dataset". Comparison of different approaches for model selection (Complete, Backward Stepwise, Forward Stepwise, GARS) for linear regression model when the Body Fat Dataset is considered. For each approach, three statistical criteria are considered (AIC, BIC and SIC).

4.2 Implementation Details

Both GARS and GARST use the same genetic operators: stochastic uniform selection, single-point crossover with p_c=0.8, uniform mutation with p_m=1/NBITS, reinsertion of the best recorded solution at each generation (elitism size=1). The number of individuals NIND is set equal to 100 and the algorithms terminate after 1000 generations (MAXGEN). Thirty runs have been performed for GARS and GARST for each dataset. The set of possible transformations considered for the "Cholesterol Measurement" dataset does not include the logarithmic operator, since some data are equal to zero, i.e., the set of feasible transformation includes only the exponential and power functions.

5 Empirical Results and Discussion

Section 5.1 reports the comparison of GARS results for thirty runs for the three statistical criteria (AIC, BIC and SIC), with the ones reported by backward and forward stepwise regression methods and the complete model, i.e., the model that includes all the regressors. Section 5.2 reports the comparison between GARS and GARST results after 30 runs for each statistical criterion.

5.1 GARS Empirical Results

Table 1 shows the empirical results when the *"Body Fat Measurement"* dataset is considered. Column 1 reports the model selection scheme, columns 2 and 3 the statistical criterion used and the corresponding value for the best models identified by each algorithm, columns 4 and 5 the R^2 and *Residual Mean Squared Error (RMSE)* statistics computed on the test set, columns 6 and 7 the number of regressors of the best selected models and the ordinal number of the selected regressors. The recorded minimum values of the statistical criteria are marked in bold, while the maximum R^2 and the minimum RMSE are underlined and in italics.

In this experiment, GARS selects always the models with smaller AIC, BIC and SIC values than the complete model and the ones selected by the backward and forward stepwise regression. In fact, as already mentioned, stepwise regression converges to local optima. By exhaustive search, we checked that the models identified by GARS correspond to the global optima in correspondence of the different criteria. The analysis of the "Body Fat Measurement" dataset shows that GARS is robust and can explore effectively the search space. Thus GARS could be useful in the analysis of complex dataset, when the number of regression variables is not small.

Furthermore our investigation shows that even if GARS properly converges towards the smallest AIC, BIC and SIC values, the selected models are not the

Type of Model Selection	Stat. Criterion	Best Stat. Criterion Value	R^2	RMSE	Number of Independent Variables	Selected Independent Variables
Complete Model	AIC	7.284	0.757	7.533	21	1,2,3,...,21
	BIC	9.205	0.757	7.533	21	1,2,3,...,21
	SIC	7.812	0.757	7.533	21	1,2,3,...,21
Backward Stepwise Regression	AIC	6.938	*0.761*	7.612	16	2,3,4,6,7,8,9,11,12,13, 14,15,16,18,20,21
	BIC	6.947	0.731	5.500	3	3,4,14
	SIC	6.670	0.728	5.603	5	2,8,9,11,20
Forward Stepwise Regression	AIC	6.662	0.734	5.241	2	1,17
	BIC	6.828	0.731	5.218	2	1,14
	SIC	6.713	0.734	5.241	2	1,17
GARS	AIC	**6.480**	0.739	*5.186*	6	5,8,10,11,15,19
	BIC	**6.734**	0.747	5.509	2	4,13
	SIC	**6.601**	0.717	6.267	2	10,12

Table 2: "Cholesterol Measurement" dataset. Comparison of different approaches for model selection (Complete, Backward Stepwise, Forward Stepwise, GARS) for linear regression model when the Cholesterol Measurement Dataset is considered.

best ones in explaining and forecasting the dependent variable with respect to the test set. In fact, looking at table 1, one should note that the model with the largest R^2 *(=0.970)* and the smallest *RMSE (=0.219)* is the one identified by the forward stepwise regression scheme for AIC criterion, when the third, fourth, sixth, ninth and eleventh variables are considered.

Column 7 in Table 1 reports which independent variables have been included in the linear regression model. The models found by the forward stepwise regression and GARS use a smaller number of independent variables than the models identified by the backward stepwise regression. The third and sixth variables are always selected by all the approaches and the ninth by all the approaches except GARS for the BIC criterion. The models obtained by GARS and by the forward stepwise approach are quite similar.

The "Cholesterol Measurement" dataset consists of 264 observations of 21 independent variables. The size of the search space is equal to 2^{21}.

Even for a bigger search space, GARS is still capable of selecting the models with smaller AIC, BIC and SIC values (Table 2, column 3, numbers in bold) than the ones selected by the backward and the forward stepwise methods and the complete model.

As for the *"Fat Measurement"* dataset, none of the models selected by GARS for AIC, BIC and SIC *(AIC-R^2=0.739, BIC-R^2=0.747, SIC-R^2=0.717)* has the largest R^2 among all the reported models. The backward stepwise method for the AIC criteria selects the model with the largest R^2, which is equal to 0.761 (column 4 underlined in italics). However, this model has inferior forecasting capabilities in terms of *RMSE (=7.612)* than all the other reported models. The model with the smallest RMSE is the one with minimum AIC selected by GARS *(RMSE=5.186).*

In contrast to the first dataset, the set of selected variables varies greatly for the different algorithms and for different statistical criteria. When we further investigated the *"Cholesterol Measurement"* dataset, it turned out that all the dependent variables are strongly correlated: the minimum correlation coefficient among all the ones is in fact 0.8623. Hence, the selection of the regressors and the model selection are negatively influenced by it.

The forward stepwise regression and GARS for BIC and SIC values select parsimonious models with only two regressors. Moreover, the forward stepwise regression always selects the first variable, and then the seventeenth in correspondence of AIC and SIC criteria and the fourteenth in correspondence of BIC criterion. The inclusion of the first variable in all the three models but its exclusion in all the models selected by the other approaches might be because of stagnation at a local optimum. GARS selects the fourth and the thirteen for AIC and the tenth and the twelveth for SIC.

	Stat. Criterion	Best Stat. Criterion Value	R^2	RMSE	Number of Independent Variables	Selected Independent Variables
FAT	AIC-GARS	0.632	0.967	0.231	5	3,5,6,9,14
	AIC-GARST	**-2.484**	*0.973*	0.233	5	2,3,4,6,16
	BIC- GARS	0.803	0.964	0.241	2	3,6
	BIC- GARST	**-2.573**	*0.973*	*0.228*	3	3,4,6
	SIC- GARS	0.762	0.967	0.230	5	3,5,6,9,14
	SIC- GARST	**-2.380**	*0.973*	0.232	4	2,3,6,16
CHOLES	AIC- GARS	6.480	0.739	5.186	6	5,8,10,11,15,19
	AIC- GARST	**6.405**	*0.800*	*3.985*	7	3,6,7,14,17,19,20
	BIC- GARS	6.734	0.747	5.509	2	4,13
	BIC- GARST	**6.699**	0.725	5.225	2	4,11
	SIC- GARS	6.601	0.717	6.267	2	10,12
	SIC- GARST	**6.538**	0.754	4.603	4	4,8,9,11

Table 3: "Fat Measurement" and "Cholesterol Measurement" datasets. Comparison of GARS (cells in white) and GARST (cell in grey) approaches for model selection for regression models. For each approach, three statistical criteria are considered (AIC, BIC and SIC).

Finally, the results could serve as a starting point for further expert-based selection. For instance, if the expert is interested in selecting a parsimonious model, then models selected by GARS for BIC and SIC and the models selected by the forward stepwise method could be a good starting point for further investigation. In case that the expert is interested in a model with good forecasting capabilities, the model selected by GARS for AIC and the models selected by the forward stepwise methods should be considered first.

5.2 GARST Empirical Results

As mentioned earlier, GARST not only identifies which variables should be included in the model, but also which transformations (from a given set of mathematical functions) should be applied before being included in the model such that the selected regression models have optimal AIC, BIC and SIC values.

Table 5 reports the empirical results from the analysis of the *"Fat Measurement"* dataset (2^{nd}-7^{th} rows) and of the *"Cholesterol Measurement"* dataset (8^{th}-13^{rd} rows) of GARS (cells in white) and GARST (cells in grey) algorithms in 30 runs. For each approach, as reported in column 2, three statistical criteria were considered (AIC, BIC and SIC). The values of the statistical criteria for the best selected model is reported in column 3, the R^2 and the Return Mean Squared Error (RMSE) respectively in columns 4 and 5, the number and the ordinal identification of the regressors included respectively in columns 6 and 7. The best fitness values are reported in bold, while the biggest R^2 and the smaller RMSE are in italics and underlined.

GARST always selects models with smaller fitness values than the ones selected by GARS in correspondence of all the three criteria (AIC, BIC and SIC). The fitness values of the best recorded models are much smaller for the *"Fat Measurement"* dataset, while the decrease in the fitness values is quite small for the "Cholesterol Measurement" dataset. As Table 5 reports, GARST does not always converge to the same fitness value for all the thirty simulations. It is important to notice that, for the "Fat Measurement" dataset, the maximum fitness values are always much smaller than the ones in correspondence of the optimal models selected by GARS, while this is not true when the *"Cholesterol Measurement"* dataset is considered.

The results from the analysis of the "Fat Measurement" dataset show that the optimal models selected by GARST for three criteria (AIC, BIC and SIC) still include the third and the sixth variables. R^2 values (=0.973 for AIC, BIC and SIC) are always larger than the ones previously computed for GARS, while RMSE is smaller only for the BIC criterion (RMSE=0.228). While the R^2 is larger than the best one reported in Table 1 (i.e.: Forward Stepwise Regression – AIC R^2*=0.970*), the smaller RMSE reported in Table 5 (i.e.: *GARST-BIC=0.228*) is not

smaller than the best result reported in Table 1 (i.e.: Forward Stepwise Regression – *AIC RMSE=0.219*).
Using the three criteria leads to select quite similar models. The best model identified by GARST in correspondence of AIC, BIC and SIC criteria are reported below. Note that the third and sixth variables are included in all the three models after being mathematically transformed in the same way. Moreover, the BIC and SIC models are all sub-models of the best one identified in correspondence of the AIC.

$$AIC \quad \hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 \frac{1}{e^{X_2}} + \hat{\beta}_2 \frac{1}{\log X_3} + \hat{\beta}_3 \frac{1}{\sqrt{e^{X_4}}} + \hat{\beta}_4 \frac{1}{\sqrt{\log X_6}} + \hat{\beta}_5 \frac{1}{\sqrt{e^{X_{16}}}}$$

$$BIC \quad Y = \hat{\beta}_0 + \hat{\beta}_1 \frac{1}{\log X_3} + \hat{\beta}_2 \frac{1}{\sqrt{e^{X_4}}} + \hat{\beta}_3 \frac{1}{\sqrt{\log X_6}}$$

$$SIC \quad Y = \hat{\beta}_0 + \hat{\beta}_1 \frac{1}{e^{X_2}} + \hat{\beta}_2 \frac{1}{\log X_3} + \hat{\beta}_3 \frac{1}{\sqrt{\log X_6}} + \hat{\beta}_4 \frac{1}{\sqrt{e^{X_{16}}}}$$

GARST allows to identify models with better AIC, BIC and SIC values. The identification of appropriate transformations of the variables could also help in revealing non-linear relationships between variables, even if the interpretation of the results could be more difficult.
The results from the analysis of the *"Cholesterol Measurement"* dataset shows better fitness values in correspondence of the selected optimal models from GARST, but the improvement in the fitness value is relatively small compared to the values reported by GARS analysis. The models selected by GARST are quite different from each other for the three criteria and from the ones previously identified by GARS. As mentioned earlier, one of the possible reasons could be that all the independent variables in the "Cholesterol Measurement" dataset are strongly correlated with each other.

Regarding the AIC criterion, GARST selects the model with both larger R^2 *(=0.800)* value and smaller *RMSE (=3.985)* than all other models obtained by GARS for BIC and SIC criteria and the backward and forward stepwise methods. Such model, reported below suggests a nonlinear relationship between the dependent and independent variables.

$$AIC \quad \hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 \frac{1}{e^{3X_3}} + \hat{\beta}_2 \frac{1}{e^{X_6}} + \hat{\beta}_3 \frac{1}{e^{2X_7}} + \hat{\beta}_4 \frac{1}{\sqrt{e^{X_{16}}}} + \ldots$$

$$\ldots + \hat{\beta}_5 \frac{1}{e^{3X_{17}}} + \hat{\beta}_6 \frac{1}{e^{3X_{19}}} + \hat{\beta}_7 \frac{1}{e^{2X_{20}}}$$

$$BIC \quad \hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 \frac{1}{\sqrt{e^{X_4}}} + \hat{\beta}_2 \frac{1}{e^{X_{11}}}$$

$$SIC \quad \hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 \frac{1}{e^{3X_3}} + \hat{\beta}_2 \frac{1}{e^{-2X_6}} + \hat{\beta}_3 \frac{1}{e^{X_9}} + \hat{\beta}_4 \frac{1}{e^{X_{11}}}$$

6 Conclusions

Genetic algorithms are a promising approach to tackle the regression model selection problem. In this paper, we compare two genetic algorithms for regressors' selection (GARS) and regressors' selection and transformation (GARST) with stepwise regression in the analysis of two real world datasets *("Fat Measurement"* and *"Cholesterol Measurement"* datasets), for three different statistical criteria (AIC, BIC and SIC). The experiments showed that results obtained with stepwise regression were inferior compared to GARS for all statistical criteria regarding the "*Cholesterol Measurement*" dataset and that GARS converges to the global optimum when considering the "*Fat Measurement*"dataset. The models selected by GARS are superior in term of AIC, BIC and SIC but not always when R^2 and RMSE statistics are considered.
The empirical results from analysing the "*Cholesterol Measurement*" dataset with GARS shows that it is not possible, as in the analysis of "*Fat Measurement*" dataset, to identify some independent variables that are included in all or most of the selected optimal models in correspondence of the three statistical criteria. The presence of highly correlated independent variables might partially explain this result.
GARS can be a valuable alternative to stepwise regression and help in suggesting new models that could be worthy to examine further.
GARST allows not only selecting which dependent variables should be considered but also which mathematically transformations should be applied to improve the adequacy of the model. The models obtained with GARST are better compared to those obtained with GARS with respect to AIC, BIC and SIC criteria. The results related to the "Fat Measurement" dataset seem to suggest that mathematical transformations of dependent variables already included in GARS optimal models should be considered, indicating the possible existence of non-linear relationships. The models that we obtained with

GARST for the "Cholesterol Measurement" dataset had better adequacy than the ones identified by GARS, but the improvement was small and the GARST models were more complex.
The empirical results show that GARST can indeed be useful in selecting model variables and proposing transformations that link them with the dependent variable. Both GARS and GARST are useful tools that can provide the researcher with useful information that could not be obtained from classical stepwise regression.

Acknowledgements

This work was supported by MIUR (Rome PRIN 2007). The authors would like to thank Thiemo Krink for helpful suggestions and Stefano Favaro for his help in the implementation of the algorithms.

References

Akaike A (1969) Statistical predictor identification. Annal of the Institute of Statistical Mathematics 22, pp. 203-217.

Akaike A (1978) Bayesian analysis of the minimum AIC procedure. Annals of the institute of Statistical Mathematics 30, part A, pp. 9-14.

Baragona R, Battaglia F, Calzini C (2001) Genetic Algorithms for the identification of additive and innovation outliers in time series. Computational Statististics and Data Analysis 37, pp. 1- 12.

Baragona R, Battaglia F, Cucina D, (2004) Fitting piecewise linear threshold autoregressive models by means of genetic algorithms. Computational Statististics and Data Analysis 47, pp. 277- 295.

Chatterjee S, Laudato M, Lynch LA, (1996) Genetic algorithm and their statistical applications: an introduction. Computational Statististics and Data Analysis 22, pp. 633–651.

Chipperfield A, Fleming P, Pohlheim H, Fonseca C, (1994) Genetic Algorithm TOOLBOX For Use with MATLAB, User's Guide. Version 1.2, Department of Automatic Control and Systems Engineering,University of SheFeld.

Cook R, Weisberg S, (1999) Applied Regression Including Computing and Graphics, Wiley Series in Probability and Statistics.

Davis L, (1991) Handbook of Genetic Algorithm. Van Nostrand Reinhold, New York.

Holland JH, (1975) Adaptation in Natural and Artificial Systems. University of Michigan Press, Ann Harbor.

Johnson RW, (1996) Fitting Percentage of Body Fat to Simple Body Measurements. Journal of Statistics Education 4, 1.

Kass RE, Raftery AE, (1995) Bayes Factors. Journal of the American Statistical Association 90, 430.

Michalewicz Z, (1996)Genetic Algorithms + Data Structures = Evolution Programs, Springer-Verlag.

Minerva T, Paterlini S, (2002) Evolutionary Approaches for Statistical Modelling. In: Fogel DB, El-Sharkawi MA, Yao X., Greenwood G, Iba H, Marrow P, Shackleton A, IEEE Press, Piscataway NJ (2002) Proc.of the Fourth Congress on Evolutionary Computation 2, pp. 2023-2028.

Minerva T, Poli I. (2001) Building ARMA Models with Genetic Algorithms. EvoWorksho, pp. 335-343.

Paterlini S, Minerva T, (2003) Evolutionary Approaches for Cluster Analysis. In Bonarini A., Masulli F., Pasi G., Soft Computing Applications, Springer-Verlag, Berlin, pp. 167-178.

Pattarin F, Paterlini S, Minerva T, (2004) Clustering Financial Time Series: An Application to Mutual Funds Style Analysis. Computational Statistics & Data Analysis 47/2 pp 353-372.

Poli I, Roverato A, (1998) A genetic algorithm for graphical model selection. J. Italian Statist. Soc. 7, 2, pp. 197–208.

Purdie , Lucas EA, Talley MB, (1992) Direct measure of total cholesterol and its distribution among major serum lipoproteins. Clinical Chemistry 38, 9, pp. 1645-1647.

Roverato A, Paterlini S, (2004) Technological Modelling for Graphical Models: An Approach Based on Genetic Algorithms. in: Computational Statistics & Data Analysis 47/2, pp 323-337.

Schwartz G, (1978) Estimating the dimension of a model. Annals of Statistics 6, pp. 461-464.

Sen A, Srivastava M, (1990) Regression Analysis: Theory, Methods, and Application, New York, Springer-Verlag.