مجلة

كلية التـــراث الجامعة

مجلة علمية محكمة

متعددة التخصصات نصف سنوية

العدد الثامن والثلاثون

15 شباط 2024

ISSN 2074-5621

رقم الايداع في دار الكتب والوثائق 719 لسنة 2011

مجلة كلية التراث الجامعة معترف بها من قبل وزارة التعليم العالي والبحث العلمي بكتابها المرقم (ب 3059/4) والمؤرخ في (2014/ 4/7)

The Using of Genetics Algorithm as a variable selection method for multiple Linear Regression model

الباحث /عمر رياض ناظم م.د.فاطمة عبدالحميد جواد

جامعة بغداد/ كلية الإدارة والاقتصاد

Abstract

Data gathered from the observation of a phenomenon are not all equally informative some of them may be noisy, correlated, or irrelevant. Variables selection aims at selecting a variable set that is relevant for a given task. This problem is complex and remains an important issue in many domains. This paper is a review of Genetics Algorithm (GA) as a variable selection method for the multiple linear regression model. Stepwise regression had been used as a traditional method for the selection of the best subset of variables to be included in the model. Adjusted R squared along with Akaike Information criterion (AIC) and Bayesian Information criterion (BIC) are used as a model's fit criteria. The research encompassed daily monitoring of nine variables, which included global horizontal irradiance, air temperature, relative humidity, dew point, wind speed, wind direction, precipitable water, azimuth, and zenith. The study sample was stratified by time into two seasons, namely the hot and cold seasons. The dataset for the hot season comprised 2571 readings, while the cold season dataset consisted of 2075 readings, encompassing the timeframe from January 1, 2021, to December 31, 2021, for Baghdad. The findings demonstrated the efficacy of the genetic algorithm approach in selecting model variables based on evaluation criteria (AIC, BIC, and adjusted R squared) surpassing the performance of stepwise regression.

Paper type: **Research paper**.

Keywords: Genetics Algorithm (GA), Multiple Linear Regression (MLR), Variable selection, Akaike Information criterion (AIC), Bayesian Information criterion (BIC)

Introduction

The process of selecting explanatory or independent variables is a very important task in building a multiple regression model. Therefore, a variable selection technique is used to identify the best subset among several variables to be included in the model. Unnecessary explanatory variables will increase the model parameter estimation noise. In addition, selecting important variables included in the model can improve the prediction accuracy of the model (Pavone et al. ,2023**)**. Moreover, the small subset of explanatory variables makes interpretation of the results easy. Therefore, we need to visualize the data in a perhaps simple way, where we select subsets of the original set of variables to get the smallest subset that can be used for modelling and reduce the cost. Therefore, redundant explanatory variables must be removed as this can help us save time and money. Traditional variable selection techniques, such as stepwise regression, forward selection, and backward elimination, have been suggested to achieve the goals. However, these methods are plagued by several drawbacks, including instability and sensitivity to violations of crucial model assumptions, which can ultimately result in inaccurate findings. The problem of selecting independent variables in linear regression has been addressed through the application of a genetic algorithm, as proposed by

Holland in (1975). Genetic algorithms, considered a reliable metaheuristic, have demonstrated effectiveness in various challenging optimization problems.
Paterlini and Minerva (2010) presented an article about verifying the performance of two types of genetic algorithms. The two methods were compared with the stepwise regression method on sample data. The research relied on statistical standards (BIC, AIC) to determine the efficiency of the models, and the results showed the superiority of the genetic algorithm over traditional methods. Trejos and Villalobos-Arias (2016) proposed the use of GA in variable selection problem for the multiple regression model, and the adjusted R2 had been used as a fitness function, and the roulette wheel method was adopted in selecting individuals.GA operators such as crossover and mutation were applied, the results showed the ability of the genetic algorithm to solve the problem satisfactorily and promisingly. Zhang et al. (2018) presented a study about a medical trial that aims to provide a tutorial on how to implement a genetic algorithm to select clinically important and statistically significant variables, while excluding irrelevant variables or variables with noise in the logistics regression model. A simulated dataset was used as an example and the result showed the ability of the genetic algorithm to choose the best or closest classification model with a small set of variables. Metwally (2019) presented a research paper aiming at measuring and analysing the influence of the stock exchange on the economic development in the Kingdom. This is done through comparing the Gross Domestic Product (GDP) as a changeable factor affiliated with some independent variables in the KSA stock exchange using the Multiple Linear Regression (Stepwise). Redha and Hadia (2019) used the genetic algorithm to get optimal estimates for survival function. The genetic algorithm is employed in the method of moment, the least squares method and the weighted least squares method and getting on more efficient estimators than classical methods. Then, a comparison was made between the methods depending on the experimental side. The best method is evaluated based on mean square error of the survival function. Żogała-Siudem and Jaroszewicz (2020) presented an
approach to efficiently construct stepwise regression models in a very high dimensional setting using a multidimensional index. The approach is based on an observation that the collections of available predictor variables often remain relatively stable and many models are built based on the same predictors. They propose an approach where the user simply provides a target variable and the algorithm uses a pre-built multidimensional index to automatically select predictors from millions of available variables, yielding results identical to standard stepwise regression, but an order of magnitude faster. Robinson et al. (2021) presented a new genetic algorithm (GA) for partial least squares regression (PLSR) feature and latent variable selection is proposed to predict concentrations of 18 important bioactive components across three New Zealand horticultural products from infrared, near-infrared and Raman spectral data sets. Models generated using the GA-enhanced PLSR method have notably better generalization and are less complex than the standard PLSR method. GA-enhanced PLSR models are produced from each spectroscopic data set individually, and from a data set that combines all three techniques. Zhang ,Yang and Ding (2023) studied the Akaike information criterion (AIC) and Bayesian
information criterion (BIC), and revisit information criteria by summarizing their key concepts, evaluation metrics, fundamental properties, interconnections, recent advancements, and common misconceptions to enrich the understanding of model selection in general.

The research's objective is to assess and compare the effectiveness of variable selection between the stepwise regression method and the genetic algorithm method, considering (AIC,BIC and adjusted R squared) as model fit criteria.

1- Material and Methods

2.1 Multiple Linear Regression (MLR)

The idea of "regression" was introduced by the English scientist and statistician Sir Francis Calton (1885) (Ethington et al ;2002). Regression models constitute a foundational component within statistical analysis, serving as a critical tool for investigating the relationships between variables. Among these models, the multiple linear regression model, often referred to as the general linear model, facilitates the exploration of associations between a dependent variable (Y) and multiple independent variables ($X_1,X_2,\ldots,X_k$) . This becomes particularly pertinent when the behavior of the observed phenomenon exhibits complexity arising from the influence of multiple factors, each exerting varying degrees of impact. The multiple regression model, thereby, extends the fundamental principles of the simple regression model, accommodating the inclusion of multiple independent variables. While the simple model hinges on the interplay between two variables—one dependent and one independent (Shetty et al. 2020). the model equation can be written as follow: (Olive, 2017)

$$Y_i = \beta_1 X_{i,1} + \beta_2 X_{i,2} + \ldots + \beta_p X_{i,p} + u_i \quad \ldots (1)$$

for i = 1,2. . . , n. Here n is the sample size and the random variable u_i is the i^{th} error, in matrix terms, linear regression can be illustrated as follows:

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{bmatrix} 1X_{11}X_{12}\ldots X_{1k} \\ 1X_{21}X_{22}\cdots X_{2k} \\ \vdots\ \vdots\ \vdots \quad \vdots \\ 1X_{n1}X_{n2}\cdots X_{nk} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}$$

and it can be written as:

$$Y_{(n,1)} = X_{(n,\,k+1)}\ \beta_{(K+1\,,1)} + u_{(n\,,\,1)} \quad \ldots(2)$$

Where:

Y: is Vector of nx1 dimension, contains the dependent variable observation.

X: is matrix of n x K+1 contains the independent variable observation; its first column contains the values of the integer unit to represent the coefficients of B_0.

B: Is a vector with dimensions (k+1 x 1), containing the values of the features to be estimated.

U: A vector with dimensions (nx1), containing random error values.

2.1.2 Variable Selection methods

There are various approaches for selecting variables for a final model, and no consensus exists on the best method. One approach is the "full model approach," where all candidate variables are included in the model. This approach has its advantages because it eliminates the problem of selection bias, and the standard errors (SEs) and p-values of the variables are accurate. However, defining a full model can be challenging and may not always be practical due to complexity.

Another suggestion is to start with univariate analysis for each variable. Variables that show significance ($p \leq 0.25$) in univariate analysis, as well as those that are important for the case study, should be included for multivariate analysis. However, univariate analysis may overlook the fact that individually weakly associated variables can become significant when combined. To address this, a higher significance level can be set for univariate analysis to allow more variables to be considered. There are four major variable selection methods: backward elimination, forward selection, stepwise selection, and all possible subset selection. (Chowdhury And Turin,2020).

2.1.2 Stepwise Regression

The stepwise regression technique represents one of the most used methods in statistical modeling. Initially, it entails the computation of ordinary linear regressions for the dependent variable (Y) against each individual explanatory variable (xi) (Yilmazer And Kocaman, 2020). The selection of these explanatory variables is contingent upon the strength of their associations with the dependent variable. Therefore, a variable is incorporated into the model if its simple correlation coefficient (r) exhibits the highest value among all conceivable pairs of explanatory variables.

Subsequently, the model undergoes iterative modifications, wherein explanatory variables are either added or removed based on the outcomes of the F-test. This iterative process continues until the calculated F value no longer achieves statistical significance. The linear combination of the chosen explanatory variables then culminates in the formulation of the final model.

Figure (1) visually delineates the procedural steps involved in constructing the model through the stepwise regression method. (Sonoda et al, 2007).

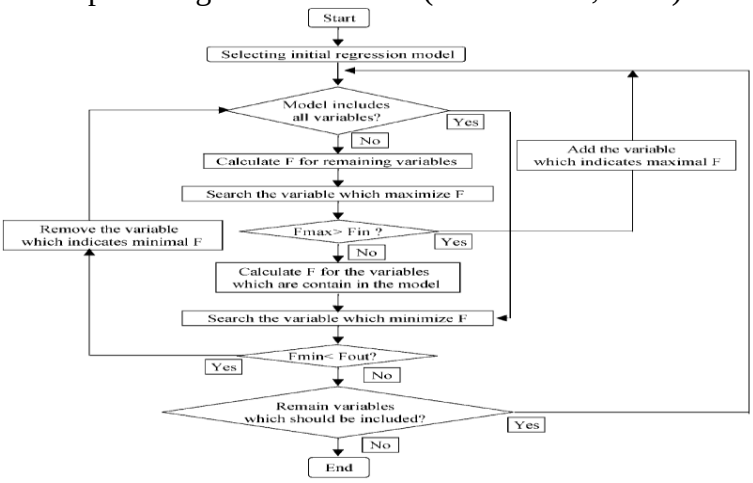

FIGURE 1 : STEPWISE REGRESSION DIAGRAM

2.2 Genetics Algorithm

Genetic algorithms (GAs) represent a type of evolutionary algorithm deeply rooted in the principles of natural selection, genetics, and artificial intelligence methodologies (Abdel Hadia and Redha,2020) . These algorithms find significant application in solving optimization problems characterized by vast or intricate solution spaces. Originating in the 1960s, genetic algorithms were conceptualized by computer science professor (JOHN H. HOLLAND) and subsequently refined by him, his students, and colleagues at the University of Michigan

throughout the 1960s and 1970s. (Katoch,Chauhan and Kumar,2021). Genetic algorithms draw inspiration from the foundational concepts outlined in Charles Darwin's theory of natural evolution, as expounded in his seminal work, "On the Origin of Species." This theory is predicated on the notion of "survival of the fittest" (Eliot, 1904) positing that organisms possessing traits conducive to their specific environment are more likely to thrive and propagate successfully. The mechanism behind this process involves the selective retention of advantageous characteristics and their inheritance across successive generations. Organisms endowed with traits that enhance their survival and reproductive capabilities are better positioned to pass on these favorable attributes to their progeny. Conversely, organisms burdened with traits that impede their ability to thrive and reproduce are less likely to perpetuate such traits to future generations.

2.2.1 Methodology

The genetic algorithm initiates its search process with a randomly generated set of potential solutions, represented as individuals or chromosomes. These chromosomes typically employ binary encoding, where each chromosome is a binary string There exists another type of encoding such as Octal Encoding, Hexadecimal Encoding and Value Encoding as shown in Figure (2).

Chromosome 1	1 1 0 1 0 0 1 1 1 0
Chromosome 2	1 0 1 0 1 1 0 1 0 1

Binary Encoding

Chromosome 1	01674237
Chromosome 2	51056801

Octal Encoding

Chromosome 1	2.12, 3.63, 0.21
Chromosome 2	ADHGTGHBSADJ

Value Encoding

Chromosome 1	3DA8
Chromosome 2	9FBC

Hexadecimal Encoding

FIGURE 2: ENCODING

Subsequently, the performance of each individual chromosome is assessed based on specific criteria determined by a designated fitness function. The fitness function is closely tied to the objective function of optimization or search problem under consideration. (Dharma et al, 2020). Following the evaluation phase, natural selection mechanisms such as (Roulette Wheel, Rank selection, Elitism and Tournament selection) are applied to the chromosomes. Chromosomes with superior performance, akin to the concept of "survival of the fittest" in natural selection, are chosen to reproduce and pass on their genetic information (AlKhafaji and AlBakri,2021). Less fit individuals are removed from the population. The next step involves reproduction, where a new generation of chromosomes is generated through processes such as crossover and mutation.

Crossover involves exchanging subparts of two chromosomes (Mirjalili,2019), there are several mechanisms of crossover such as single point crossover, tow point crossover and uniform crossover (see figure 3) , while mutation entails altering the genetic factors at specific positions within the chromosome. (Hassanat et al,2019) These processes iteratively continue for a specified number of generations or until a satisfactory solution to the problem is achieved.

With each new generation, there is an evolution of chromosomes, where favorable traits are progressively propagated from one generation to the next (Ismael and Ghanawi,2019). In this manner, genetic algorithms mimic the iterative biological processes observed in natural evolution. They serve as a powerful approach for finding optimal solutions to complex problems and adapting to varying environmental conditions.

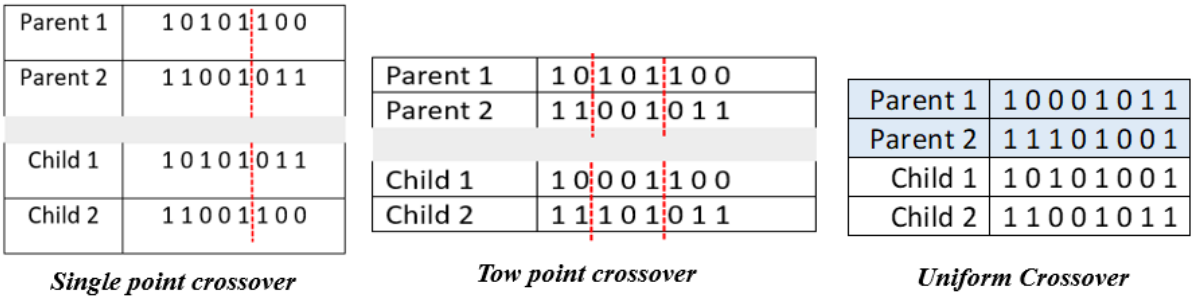

FIGURE 3 CROSSOVER MECHANISM

2.3 Model Fit Criteria

2.3.1 Adjusted R^2

The adjusted R^2 is a statistical measure used to evaluate the performance of a linear regression model. The adjusted coefficient of determination is similar to the R-squared but adds an adjustment to correct for imprecision resulting from adding predictor variables to the model. The coefficient of determination measures the extent to which a model can explain the variance in the dependent variable (the variable that is predicted) through the explanatory variables (the variables that are used to predict outcomes). (Karch ,2020) The criteria formula as follows:

$$\text{adjusted } R^2 = 1 - \left(\frac{n-1}{n-k}\right)(1 - R_k^2) \qquad \ldots(3)$$

$$\text{where: } R^2 = 1 - \frac{RSS}{TSS} \qquad \ldots(4)$$

Where:

R^2 : is the coefficient of determination.

N: is the sample size.

K: is the number of explanatory variables in the model.

RSS: Residual sum of squares.

TSS: Total sum of squares.

The adjusted R^2 increases when a new independent variable improves the model's fit and decreases when adding a variable does not contribute much explanatory power. It provides a more balanced assessment of the model's goodness of fit, considering the trade-off between model complexity (number of variables) and explanatory power.

2.3.2 The Akaike Information Criterion (AIC)

AIC is a statistical measure used in the process of selecting and evaluating models. It was developed by Japanese statistician Hirotugu Akaike in 1974. It is used to evaluate the relative performance of different models by comparing several models and selecting the model that provides the best balance between data characterization and model complexity. (Cavanaugh and Neath. 2019). The calculation of the Information Criterion (AIC) is based on the principle of the need to characterize data with the least amount of complexity. It is calculated from the following formula:

AIC = 2*k - 2*ln(L) … (5)

Where :

k: is the number of explanatory variables in the model (number of coefficients).

ln(L): is the natural logarithm of the standard likelihood function of the model.

2.3.3 The Bayesian Information Criterion (BIC)

BIC is one of the most popular and widespread tools in statistical model selection. Its popularity derives from its computational simplicity and effective performance in various modeling scenarios; The BIC was developed by (Gideon E. Schwarz). The BIC is designed to encourage the avoidance of unnecessary variables in a model, thereby reducing model complexity and enhancing its ability to generalize to new data. Similar to the Akaike Information Criterion (AIC), the BIC aids in comparing different models and selecting the one that strikes the best balance between providing a good statistical description of the data and keeping model complexity in check. (Pho et al., 2019).

BIC = k * ln(n) - 2 * ln(L) . . . (6)

Where :

n is the sample size

k: is the number of explanatory variables in the model (number of coefficients).

ln(L): is the natural logarithm of the standard likelihood function of the model.

2.4. Data Descriptive :

The study dataset consists of hourly readings for (8) meteorological variables in Baghdad, located at a longitude of 44.3661, latitude of 33.3152, and altitude of 37. The data spans from January 1, 2021, to December 31, 2021. To enhance data homogeneity, the 'Time-Stratified-TS' method was applied (Mohammed and Basheer Hannon ,2019) , dividing the dataset into two seasons: the hot season comprising the months (April, May, June, July, August, and September) with 2571 readings and the cold season including the months (January, February, March, October, November, and December) with 2075 readings. Readings with zero solar radiation intensity (indicating nighttime hours) were excluded from the study sample. The variables are as follows: The global horizontal irradiance (GHI) represents the dependent variable Y, the independent variables are represents as follows (Air temp(X1), Azimuth (X2) Dewpoint (X3), Precipitable water (X4), Relative humidity (X5), wind direction (X6) , wind speed (X7), Zenith (X8)) .

These data were obtained from:

Solcast, 2019. Global solar irradiance data and PV system power output data. URL https://solcast.com/"

3. Discussion of Results

Statistical software (SPSS and MATLAB) were used to perform the stepwise regression and calculate the three comparing criteria (AIC , BIC, adjusted R squared). For genetics algorithm, R studio software were used, and the GA inputs where as follows:

Initial Population size	216
Number of Generations	30
Fitness Function	AIC, BIC, Adjusted R squared
Encoding	Binary
Selection	Tournament Selection
Crossover	Single point

The Result were as follows:

Table (1) **AIC results according to Model, Method and Season with deference variables included**

Model	Method	Season	Variables Included	AIC
1	Stepwise Regression	Hot	X1 X2 X7 X8	27660.61
		Cold	X1 X2 X4 X8	23261.94
	GA	Hot	X1 X2 X3 X4 X6 X7 X8	22286.86
		Cold	X1 X2 X3 X4 X5 X6 X7 X8	19045.18
2	Stepwise Regression	Hot	X1 X2 X8	27662.91
		Cold	X1 X2 X8	23268.98
	GA	Hot	X1 X2 X3 X4 X5 X6 X7 X8	22287.56
		Cold	X1 X2 X3 X4 X6 X7 X8	19048.61
3	Stepwise Regression	Hot	X2 X8	27685.61
		Cold	X2 X8	23291.09
	GA	Hot	X1 X2 X4 X6 X7 X8	22295.66
		Cold	X2 X3 X4 X5 X6 X7 X8	19066.16

From Table (1) we noticed that all the three models of GA have the lowest AIC value comparing to the stepwise regression models, and the model number (1) has the lowest AIC value which indicates that it is the best model, then the model number (2) and the last one is model number (3). According to the best model, the stepwise regression includes 4 variables and excludes 4. While the genetics algorithm included 7 variables in the hot season model and excludes one variable which is (X5) , and for the cold season model all the eight variables were included.

Table 2: BIC results according to Model, Method and Season with deference variables included				
Model	**Method**	**Season**	**Variables Included**	**BIC**
1	Stepwise Regression	Hot	X1 X2 X7 X8	27689.86
		Cold	X1 X2 X4 X8	23290.13
	GA	Hot	X1 X2 X3 X4 X6 X7 X8	22333.67
		Cold	X1 X2 X3 X4 X6 X7 X8	19093.71
2	Stepwise Regression	Hot	X1 X2 X8	27686.31
		Cold	X1 X2 X8	23291.52
	GA	Hot	X1 X2 X3 X4 X5 X6 X7 X8	22340.23
		Cold	X1 X2 X3 X4 X6 X7 X8	19095.91
3	Stepwise Regression	Hot	X2 X8	27703.16
		Cold	X2 X8	23308
	GA	Hot	X1 X2 X4 X5 X6 X7 X8	22342.47
		Cold	X2 X3 X4 X5 X6 X7 X8	19110.77

The Table (2) shows that all the three models of GA have the lowest BIC value comparing to the stepwise regression models, the best model is the model number (1) which has the lowest BIC value, the second-best model is number (2), while model number (3) has the largest BIC values which indicate the lowest accuracy of the model fit. According to the best model, the stepwise regression includes 4 variables and excludes 4. While the genetics algorithm included 7 variables in both seasons and excludes one variable which is (X5).

Table 3: Adjusted R squared results **according to Model, Method and Season with deference variables included**

Model	**Method**	**Season**	**Variables Included**	**Adjusted R^2**
1	Stepwise Regression	Hot	X1 X2 X7 X8	0.8695
		Cold	X1 X2 X4 X8	0.726
	GA	Hot	X1 X2 X3 X4 X5 X6 X7 X8	0.944
		Cold	X1 X2 X3 X5 X4 X6 X7 X8	0.828
2	Stepwise Regression	Hot	X1 X2 X8	0.8694
		Cold	X1 X2 X8	0.7249
	GA	Hot	X1 X2 X3 X4 X6 X7 X8	0.944
		Cold	X1 X2 X3 X4 X6 X7 X8	0.8276

3	Stepwise Regression	Hot	X2 X8	0.8681
		Cold	X2 X8	0.7218
	GA	Hot	X1 X2 X4 X5 X6 X7 X8	0.9438
		Cold	X2 X3 X4 X5 X6 X7 X8	0.8262

The results of table (3) shows that the Adjusted R squared for all of the GA models are the larger comparing to the stepwise method, and the largest adjusted R squared was for the hot season dataset in model number (3) followed by the cold season dataset for the same model where 7 variables were included and only one variable was excluded in both seasons.

4.Conclusion

From the results that have been obtained, selecting the best variable subset to be included in the multiple linear regression model using (AIC,BIC and Adjusted R squared) as a model fit criteria. The Genetics algorithm shows the superiority over the traditional method represented be the stepwise regression by giving the lowest value of (AIC and BIC) and the largest value of the adjusted R squared and for all models.

References

1	**Abdel Hadia, A. T. and Redha, S. M. (2020) "Estimate The Survival Function By Using The Genetic Algorithm", Journal of Economics and Administrative Sciences, 26(122), pp. 440–454. doi: 10.33095/jeas.v26i122.2018.**
2	**AlKhafaji,M.A. and AlBakri, R.A. (2021) "Using Iterative Reweighting Algorithm and Genetic Algorithm to Calculate The Estimation of The Parameters Of The Maximum Likelihood of The Skew Normal Distribution", Journal of Economics and Administrative Sciences, 27(127), pp. 253–264. doi: 10.33095/jeas.v27i127.2148.**
3	**Cavanaugh, J. E., & Neath, A. A. (2019). The Akaike information criterion: Background, derivation, properties, application, interpretation, and refinements. Wiley Interdisciplinary Reviews: Computational Statistics, e1460. doi:10.1002/wics.1460**
4	**Chowdhury, M.Z.I. and Turin, T.C., 2020. Variable selection strategies and its importance in clinical prediction modelling. Family medicine and community health, 8(1).**
5	**Dharma, F., Shabrina, S., Noviana, A., Tahir, M., Hendrastuty, N. and Wahyono, W., 2020. Prediction of Indonesian inflation rate using regression model based on genetic algorithms. Jurnal Online Informatika, 5(1), pp.45-52.**
6	**Eliot, C.L,1904," THE ORIGIN OF SPECIES BY CHARLES DARWIN", P F COLLIER & SON,NEW YORK. 146**
7	**Ethington, C.A., Thomas, S.L. & Pike, G.R.. (2002). Back to the Basics: Regression as It Should Be. In: Smart, J.C., Tierney, W.G. (eds) \ Higher Education: Handbook of Theory and Research, vol 17. Springer Science+Business Media , 263-264**
8	**Hassanat, A., Almohammadi, K., Alkafaween, E.A., Abunawas, E., Hammouri, A. and Prasath, V.S., 2019. Choosing mutation and crossover ratios for genetic algorithms—a review with a new dynamic approach. Information, 10(12), p.390.**
9	**Holland, J.H. (1975). Adaptation in Natural and Artificial Systems. The University of Michigan Press, Ann Arbor**

10	Ismael,M.M and Ghanawi, H.A. (2019) "Building the optimal portfolio for stock using multi-objective genetic algorithm - comparative analytical research in the Iraqi stock market", Journal of Economics and Administrative Sciences, 25(113), pp. 45–78. doi: 10.33095/jeas.v25i113.1688.
11	Karch, J., 2020. Improving on Adjusted R-squared. Collabra: Psychology, 6(1).
12	Katoch, S., Chauhan, S.S. and Kumar, V., 2021. A review on genetic algorithm: past, present, and future. Multimedia tools and applications, 80, pp.8091-8126.
13	Metwally,F.R. (2019) "Stock Exchange and its Impact on Economic Development In the Kingdom of Saudi Arabia (KSA)", Journal of Economics and Administrative Sciences, 25(114), pp. 367–389. doi: 10.33095/jeas.v25i114.1740.
14	Mirjalili, S., 2019. Genetic algorithm. Evolutionary Algorithms and Neural Networks: Theory and Applications, pp.43-55.
15	Olive, D.J. and Olive, D.J., 2017. Multiple linear regression (pp. 17-83). Springer International Publishing.
16	Paterlini, S. and Minerva, T., 2010, June. Regression model selection using genetic algorithms. In Proceedings of the 11th WSEAS international conference on nural networks and 11th WSEAS international conference on evolutionary computing and 11th WSEAS international conference on Fuzzy systems (pp. 19-27). Lasi, Romania: World Scientific and Engineering Academy and Society (WSEAS).
17	Pavone, F., Piironen, J., Bürkner, PC. et al. Using reference models in variable selection. Comput Stat 38, 349–371 (2023). https://doi.org/10.1007/s00180-022-01231-6
18	Pho, K.H., Ly, S., Ly, S. and Lukusa, T.M., 2019. Comparison among Akaike information criterion, Bayesian information criterion and Vuong's test in model selection: A case study of violated speed regulation in Taiwan. Journal of Advanced Engineering and Computation, 3(1), pp.293-303.
19	Shetty, D.V., Rao, B.P., Prakash, C. and Vaibhava, S., 2020, December. Multiple regression analysis to predict the value of a residential building and to compare with the conventional method values. In Journal of Physics: Conference Series (Vol. 1706, No. 1, p. 012118). IOP Publishing.
20	Sonoda, S., Takahashi, Y., Kawagishi, K., Nishida, N., & Wakao, S. (2007). Application of Stepwise Multiple Regression to Design Optimization of Electric Machine. IEEE Transactions on Magnetics, 43(4), 1609–1612.
21	Yilmazer, S. and Kocaman, S., 2020. A mass appraisal assessment study using machine learning based on multiple regression and random forest. Land use policy, 99, p.104889.
22	Zhang, Z., Trevino, V., Hoseini, S.S., Belciug, S., Boopathi, A.M., Zhang, P., Gorunescu, F., Subha, V. and Dai, S., 2018. Variable selection in logistic regression model with genetic algorithm. Annals of translational medicine, 6(3).
23	Redha, S.M. and Hadia, A.T.A., 2020. Employment of the genetic algorithm in some methods of estimating survival function with application. Periodicals of Engineering and Natural Sciences, 8(1), pp.481-490.

24	**Żogała-Siudem, B. and Jaroszewicz, S., 2021. Fast stepwise regression based on multidimensional indexes. Information Sciences, 549, pp.288-309.**
25	**Robinson, D., Chen, Q., Xue, B., Killeen, D., Fraser-Miller, S., Gordon, K.C., Oey, I. and Zhang, M., 2021, June. Genetic algorithm for feature and latent variable selection for nutrient assessment in horticultural products. In 2021 IEEE Congress on Evolutionary Computation (CEC) (pp. 272-279). IEEE.**
26	**Zhang, J., Yang, Y. and Ding, J., 2023. Information criteria for model selection. Wiley Interdisciplinary Reviews: Computational Statistics, p.e1607.**

استعمال الخوارزمية الجينية كطريقة لاختيار المتغيرات لأنموذج الانحدار الخطي المتعدد

الباحث /عمر رياض ناظم [1]
جامعة بغداد/ كلية الإدارة والاقتصاد/قسم الاحصاء
بغداد، العراق
Omarriadh9@gmail.com

م.د.فاطمة عبدالحميد جواد [2]
جامعة بغداد/ كلية الإدارة والاقتصاد/ قسم الاحصاء
بغداد، العراق
Fatimah.a@coadec.uobaghdad.edu.iq

المستخلص

تُعد مشكلة اختيار المتغيرات ذات الصلة بظاهرة معينة من القضايا المعقدة وذات اهمية في كثير من المجالات. اذ ان البيانات التي تم جمعها من مراقبة ظاهرة ما ليست جميعها مفيدة بالقدر نفسه، فقد يكون بعضها ضوضائياً، أو قد تكون مترابطة، أو غير ذات صلة. ويعُد الانحدار الخطي المتعدد اسلوباً راسخاً في تحديد وتفسير العلاقة بين ظاهرة ما والمتغيرات التي ترتبط بها. تهدُف هذه الورقة الى استعمال طريقة الانحدار المتدرج (Stepwise Regression) في اختيار المتغيرات التي تدخل ضمن النموذج ومقارنتها مع طريقة الخوارزمية الجينية بالاعتماد على معايير المقارنة (Akaike Information Criterion, Bayesian Information Criterion, Adjusted R2) وشملت الدراسة القراءات اليومية لتسعة عوامل تمثلت بـ(الاشعاع الشمسي الافقي العالمي، درجة حرارة الجو ، الرطوبة النسبية، نقطة الندى، سرعة الرياح، اتجاه الرياح، الماء القابل للتساقط، الزاوية السُمتية وبُعد السمت) وتم استخدام التراصف الزمني (Time-Stratified) لتقسيم عينة الدراسة الى موسمين (حار وبارد) بلغ حجم العينة للموسم الحار (2571) قراءة وحجم العينة للموسم البارد (2075) قراءة، حيث تمثلت القراءات لمدينة بغداد للفترة (2021\1\1 – 2021\12\31) واظهرت النتائج دقة طريقة الخوارزمية الجينية في اختيار المتغيرات التي تدخل في النموذج بالاعتماد على المعايير (AIC,BIC adjusted R2) مقارنة مع الانحدار المتدرج ، اما الشبكات العصبية فكانت نتائجها اكثر دقة في احتساب معايير المقارنة الثلاثة (AIC,BIC, adjusted R2) وان البرامج الإحصائية المستخدمة في تحليل البيانات وإيجاد النتائج هي (SPSS,MATLAB,R studio) .

نوع البحث: ورقة بحثية
الكلمات الرئيسة: الانحدار الخطي المتعدد ، الخوارزمية الجينية ، اختيار المتغيرات ، معيار معلومات اكاكي (AIC) ، معيار المعلومات البيزي(BIC) ، معامل التحديد المعدل.