http://dx.doi.org/10.1590/1678-4162-13071

Arq. Bras. Med. Vet. Zootec., v.76, n.1, p.131-136, 2024

Multiple Linear Regression versus Automatic Linear Modelling

[*Regressão Linear Múltipla versus Modelagem Linear Automática*]

S. *Genç*[1], M. *Mendeş*[2]

[1]Kırşehir Ahi Evran University, Faculty of Agriculture, Department of Agricultural Biotechnology, 40100, Kirsehir, Türkiye
[2]Canakkale Onsekiz Mart University, Faculty of Agriculture, Biometry and Genetics Unit, 17100, Canakkale, Türkiye

ABSTRACT

In this study, performances of Multiple Linear Regression and Automatic Linear Modelling are compared for different sample sizes and number of predictors. A comprehensive Monte Carlo simulation study was carried out for this purpose. Random numbers generated from multivariate normal distribution by using RNMVN function of IMSL library of Microsoft FORTRAN Developer Studio composed the material of this study. Results of the simulation study showed that the sample size and the number of predictors are the main factors that lead to produce different results. Although both methods gave very similar results especially when studied with large sample sizes (n≥100), the Automatic linear modelling is preferred for analyzing data sets due to its simplicity in analyzing data and interpreting the results, ability to present results visually and providing more detailed information especially studying large complex data sets. It will be beneficial to use the Automatic linear modelling especially in analyzing massive and complex data sets for the purposes of investigating the relationships between one continuous dependent and 10 or more predictors and determine the factors that affect the response or target variable. At the same time, it will also be possible to evaluate the effect of each predictor with a more detailed response.

Keywords: multiple regression, automatic linear modelling, simulation, R^2

RESUMO

Neste estudo, os desempenhos da Regressão Linear Múltipla e da Modelagem Linear Automática são comparados para diferentes tamanhos de amostra e número de preditores. Para isso, foi realizado um estudo abrangente de simulação de Monte Carlo. Os números aleatórios gerados a partir da distribuição normal multivariada usando a função RNMVN da biblioteca IMSL do Microsoft FORTRAN Developer Studio compuseram o material deste estudo. Os resultados do estudo de simulação mostraram que o tamanho da amostra e o número de preditores são os principais fatores que levam à produção de resultados diferentes. Embora ambos os métodos tenham apresentado resultados muito semelhantes, especialmente quando estudados com amostras de tamanho grande (n≥100), a modelagem linear automática é preferida para a análise de conjuntos de dados devido à sua simplicidade na análise de dados e na interpretação dos resultados, à capacidade de apresentar os resultados visualmente e ao fornecimento de informações mais detalhadas, especialmente no estudo de conjuntos de dados grandes e complexos. Será vantajoso usar a modelagem linear automática, especialmente na análise de conjuntos de dados maciços e complexos com o objetivo de investigar as relações entre um dependente contínuo e 10 ou mais preditores e determinar os fatores que afetam a resposta ou a variável-alvo. Ao mesmo tempo, também será possível avaliar o efeito de cada preditor com uma resposta mais detalhada.

Palavras-chave: regressão múltipla, modelagem linear automática, simulação, R2

Corresponding author: serdargenc1983@gmail.com
Submitted: June 26, 2023. Accepted: September 20, 2023.

INTRODUCTION

Investigating relationships between/among variables is of great interest for practitioners (Mendeş, 2019; Temizhan *et al*., 2022). Multiple Linear Regression (MLR) is the most used technique in investigating relations between one dependent and several independent variables. Despite its widespread use and a great tool, the presence of some disadvantages of the MLR limits its use. These disadvantages become more obvious especially when the number of variables is greater than the number of observations, high correlation between the predictors (multicollinearity problem), and presence of outliers in the data set. At the same time, although the MLR is a very beneficial technique to investigate the relationships between one dependent and several independent variables it isn't recommended for many cases in practice due to over-simplifying real world problems by assuming a linear relationship among the variables (Johnson, 1991, Yan and Su, 2009; Mendeş, 2009) Therefore, the MLR should not be used for such cases; otherwise it may lead to over fit. In case of such problems, either these problems are tried to be solved by using different methods and applying some transformations, or alternative methods that are not affected by such problems are applied. Due to its ease of application, its ability to visually present the results, and its ability to automatically determine the best sub-datasets and important independent variables it is possible to benefit from the Automatic Linear Modelling (ALM) Analysis efficiently when such problems exists (IBM..., 2012; Field, 2013; Yang, 2013; Rahnama and Rajabpour, 2016; Yakubu *et al.*, 2018; Genç and Mendeş, 2021a). This study aimed to compare the performance of the MLR and the ALM under different experimental conditions via a comprehensive simulation study.

MATERIAL AND METHODS

Random numbers generated from multivariate normal distribution by using the RNMVN function of the IMSL library of Microsoft FORTRAN Developer Studio composed the material of this study. Three goodness-of-fit criteria (i.e. R^2, accuracy level or $R^2_{adj.}$, and the rank of the place of importance of the independent variables) were used in evaluating the appropriateness of the models. To determine reference values for actual R^2 and R^2_{adj}, 1000,000 random numbers were generated from multivariate normal distribution for different variance-covariance matrixes. Randomly generated numbers were then transferred into SPSS package and the MLR and ALM were performed. Then, R^2 and R^2_{adj}. values were computed. These values were accepted as the reference or actual values for the R^2 and R^2_{adj}. Later, for p=4, 10, 15, and 20, different samples with the sizes of 20, 30, 50, 100, and 500 were taken from 1000,000 random numbers. The RT and ALM procedures were applied to those samples and the R^2 and R^2_{adj}. values were estimated. These processes were repeated 100 times. Therefore, each estimate was made based on 100 trials. Then, the numbers of trials given below were determined (Genç and Mendeş, 2021b).

a) The number of trials where both methods were found to have the same variables as important.
b) The number of trials where both the importance of variables and the order of importance were found to be the same.
c) The number of trials where the same variables were found to be important, but the order of importance was not the same.
d) The number of trials where both methods produced different results. These numbers were then converted to %.

Correlations between the predictors ranged from -0.20 to 0.90. Detailed information about experimental conditions simulated are given in Table 1.

Letter p denotes number of variables, n denotes number of observations, and X_{ij} is the i^{th} observation of the j^{th} variable. Then mean vector and variance-covariance matrix will be as below:

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ . \\ . \\ \mu_p \end{bmatrix}$$

Where $\mu_i = E(X_i) = \int X_i f(x) d(x)$ is the mean of the ith component of X.

Covariance between X_i and X_j is $\sigma_{ij} = E(X_i - \mu_i)(X_j - \mu_j) = E(X_i X_j) - \mu_i \mu_j$ and variance of each X_i is $\sigma_{ii} = E(X_i - \mu_i)^2 = E(X_i^2) - \mu_i^2$

In this case, the variance-covariance matrix will be as follow:

$$\sum = \begin{bmatrix} \sigma_{11} & \sigma_{12} \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} \ldots & \sigma_{2p} \\ \vdots & \vdots \quad \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \sigma_{pp} \end{bmatrix}$$

Table 1**.** Characteristics of Simulation Study

Sample sizes	Number of Variables (P)	Performance Criteria	Correlation range for predictors	Simulation Number
20, 30, 50, 100, 500 **Reference:** 1000,000	4, 10, 15, 20	1.R^2 2.Accuracy (R^2_{adj}) 3.Rank of the place of importance of predictors	[-0.20, 0.90]	100

Note: This study has been prepared in accordance with the Principles of the Declaration of Helsinki.

Multiple linear regression (MLR) is one of the widely used statistical techniques to explain the relationship between one continuous dependent variable and two or more independent variables. If Y is a dependent or response variable and X_1, X_2, X_3…, X_p are independent or predictor variables, then the multiple regression model will be as follows, and it provides a prediction of Y values from the X values.

$Y_i=\beta_0+ \beta_1 X_1+ \beta_2 X_2+\ldots+ \beta_p X_p+\varepsilon_i$
where
Y_i is the response variable, X_1, X_2, …, X_p are the independent variables, β_0 is the constant term or intercept, β_1, β_2, …, β_p are the regression coefficients and ε is the random error.[3,4]

Automatic Linear Modeling (ALM) was introduced in version 19 of IBM SPSS, enabling researchers to select the best subset automatically. In ALM, to provide an improvement in data fit the predictors are directly converted. SPSS uses rescaling of time, other measurements, outlier trimming, category merging and other methods in performing ALM analysis. Although the ALM can be used on small and medium-sized data sets, it is more useful especially when working with large and complex data sets. Thus, it is possible to say that the advantages of the ALM become more obvious, especially in cases where there are many estimators. On the other hand, the fact that the ALM has the potential to be misused should not be overlooked due to its simplicity and the convenience it provides in automatically identifying important variables. It is because the ALM includes automatic data preparation steps. Therefore, after the final candidate models are determined, it is of great benefit to carefully evaluate these models by considering various criteria and asking some important questions (IBM..., 2012; Yang, 2013; Genç and Mendeş, 2021b; Mendeş, 2021).

RESULTS

Descriptive statistics for p=4, 10, 15, and 20 are given in Table 2 and the results for performances of the MLR and ALM are given in Table 3.

Table 2. Descriptive Statistics for R^2 estimates of MLR and ALL for the sample sizes of 20, 30, 50, 100, 500 when p=4, 10, 15, 20

		p=4		p=10		p=15		p=20	
Method	n	Mean	SE	Mean	SE	Mean	SE	Mean	SE
MLR	20	61.13	5.25	85.01	3.85	87.72	3.90	89.56	3.98
	30	67.19	4.45	78.60	5.48	81.20	5.51	82.91	5.62
	50	69.34	2.00	87.19	1.36	89.86	1.36	91.74	1.39
	100	68.36	1.35	85.39	2.16	88.04	2.23	89.89	2.28
	500	68.10	0.81	87.63	0.72	90.35	0.74	92.24	0.76
ALM	20	59.78	4.77	87.28	1.92	89.46	1.97	91.07	2.00
	30	64.07	4.24	83.95	1.80	86.05	1.85	87.60	1.88
	50	67.57	1.95	85.01	3.85	88.14	1.30	89.73	1.33
	100	67.62	1.34	78.60	5.48	87.69	2.28	89.27	2.32
	500	67.73	0.84	87.19	1.36	89.65	0.71	91.27	0.72

When Table 2 is examined, it is seen that the means of R^2 estimates of the MLR and ALM are generally similar, and this similarity becomes more evident as the sample size increases. It is seen that the estimations tend to increase as the number of predictors increases. However, this increase is more pronounced especially when the number of predictors is between 4 and 10.

Table 3. Simulation results for evaluating performances of MLR and ALM

p	n	C1	C2	C3	C4
4	20	95	80	15	5
	30	98	63	35	2
	50	98	61	36	2
	100	100	70	30	0
	500	100	85	15	0
10	20	90	70	20	10
	30	95	52	43	5
	50	97	62	35	3
	100	100	55	45	0
	500	100	47	53	0
15	20	92	58	34	8
	30	97	53	44	3
	50	100	47	53	0
	100	100	31	69	0
	500	100	34	66	0
20	20	93	40	53	7
	30	98	43	55	2
	50	100	41	59	0
	100	100	33	67	0
	500	100	26	74	0

p: The number of predictors, **n:** Sample size
C1: The percentage that the same variables were found to be important
C2: The percentage of the numbers of trials in which both variables and the order of importance were found to be the same.
C3: The percentage of the numbers of trials in which the same variables were found to be important, but the order of importance was not found to be the same.
C4: The percentage of trials where both methods produce different results.

When table 3 is examined, it is seen that the probabilities of finding the same variables as important for both techniques are quite similar and generally ranged from 92% to 100%. As the sample size increases, the probability of giving the same results for both techniques reaches a very high level. This probability reaches 100% for the sample sizes of 100 and more regardless of the number of predictors.

When the percentage of the experiments where the same predictors are found to be important, and the order of importance is the same, it can be easily seen that as the number of predictors increased this percentage decreased. This decrease is even more pronounced when p≥15.

When the percentage of the numbers of trials in which the same variables were found to be important but the order of importance was not found to be the same is examined this probability increases as the number of predictors and sample size increase. Increase in this probability becomes more prominent especially when p≥15 and n≥100.

On the other hand, as expected the percentage of trials where both methods produce different results is very close to each other. As can be seen from the last column of the Table 1, the MLR and ALM gave different results especially when sample size is small. For large sample sizes (n>100) no difference has been observed between the two methods regardless of number of predictors. When all results are evaluated

together it is possible to conclude that differences in the sample size and the number of predictors may lead to produce different results. However, in general, usage of the ALM provides more detailed information along with its simplicity in analyzing data and interpreting the results especially studying with large complex data sets. Therefore, it will be beneficial to use the ALM especially in analyzing massive and complex data sets for the purposes of investigating the relationships between one continuous dependent and 10 and more predictors and determine the factors that affect the response or target variable. At the same time, it will also be possible to evaluate the effect of each predictor's response in more detail.

DISCUSSION

Although MLR is widely used in practice and a great tool for investigating the relationships of dependent and independent variables, it isn't recommended for many cases due to over-simplifying real-world problems by assuming a linear relationship among the variables. There are some situations that limit it's use. Linear Regression is a great tool to analyze the relationships among the variables, but it isn't recommended for most practical applications because it over-simplifies real world problems by assuming a linear relationship among the variables. At the same time, it requires some assumptions to be provided in the data set, and these assumptions are generally not fulfilled in practice. Therefore, there are some situations that limit the use of the multiple linear regression analysis. The first limitation of Multiple Linear Regression (MLR) is the assumption of linearity between the dependent variable and the independent variables. In the real world, the relationship between dependent and independent variables is not linear in many cases. This assumption is not fulfilled for many cases and that limits the use of MLR in investigating the relationships between a dependent and several independent variables. It is because accuracy decreases as the linearity of the dataset decreases. The second limitation of the MLR appears when the number of observations is lesser than the number of predictors. Since it might cause overfitting or overestimating problem, the MLR should not be used for such cases ($n<p$). The third limitation of the MLR is that it assumes that there is no multicollinearity problem among the predictors. If this problem occurs in the dataset, there should be an attempt to handle it. The fourth limitation of the MLR is that since the MLR is very sensitive to outliers, it should be so careful against outliers and thus before performing MLR analysis it should be tested if there is an outlier. In cases where the number of variables is bigger than the number of observations the usage of MLR is not also appropriate even if all above assumptions are fulfilled (Johnson, 1991; Mendeş, 2009; Yakubu *et al.,* 2018).This case becomes more obvious especially when studied with complex data sets. The ALM which is a member of linear modelling might be used efficiently instead of MLR. The ALM has three main features: a) predictors can be both continuous and categorical b) ALM automatically finds the most important predictors and eliminates the predictors which are of little or no importance in predicting the dependent variable, and c) it automatically determines if the data set contains an outlier. One of the other important features of the ALM is that since it presents the results graphically it is very easy to interpret the results.

The results of this study showed that the MLR and ALM gave different results especially when sample size is small. But any difference has not been observed between two methods regardless of number of predictors when studied with large sample sizes ($n>100$). However, the usage of ALM provides more detailed information along with its simplicity in analyzing data and interpreting the results especially studying with large complex data sets in general. A few previous studies where the ALM was used in investigating relationships between dependent and independent variables also reported that the ALM could be considered as a great tool especially when studying with large and complex data sets. For example, Oshima and Dell-Ross reported that the Automatic Linear Modeling can be an indispensable screening tool especially when there are many predictors (Oshima and Dell-Ross, 2016). However, once a handful of final candidates are chosen, it is the researcher's responsibility to carefully evaluate those models with various criteria along with substantive questions. Likewise, Yakubu et al. used ALM to predict heat stress index in Sasso hens and they reported that the ALM can be used efficiently in predicting heat stress index (Yakubu *et al.*, 2018). Genç and Mendeş used ALM to model

the factors affecting the 305-day milk yield of dairy cows by using Automatic Linear Modeling Technique (ALM). They reported that the ALM can be efficiently used for investigating the relationships between one continuous response and more predictors which had different measurement scale (Genç and Mendeş, 2021a). Mendeş used ALM for evaluating results of Monte Carlo Simulation Studies and he informed that the ALM could be used efficiently to determine the factors that affect the response variable when there is a large and complex data set (Mendeş, 2021). When results of this study and previous studies are evaluated together it is possible to conclude the following: a) usage of the ALM in analyzing large and complex data sets might enable us to interpret the results easily, b) Preferring the ALM enables us to evaluate higher order interactions among the independent variables and, c) The ALM can be efficiently used in analyzing data sets obtained from all kinds researches as long as there are many predictors. However, a potential threat of misuse of the ALM due to its simplicity should not be ignored.

CONCLUSION

As a result, it is possible to conclude that the ALM can be efficiently used in investigating relationships between one dependent and several predictors especially when used on large and complex data sets.

REFERENCES

FİELD, A. *Discovering statistics using IBM SPSS statistics*. 4.ed. Los Angeles: SAGE, 2013. 952p.

GENÇ, S.; MENDEŞ, M. Evaluating performance and determining optimum sample size for regression tree and automatic linear modeling. *Arq. Bras. Med. Vet. Zootec*., v.73, p.1391-1402, 2021b.

GENÇ, S.; MENDEŞ, M. Linear modeling analysis using for determining the factors affecting 305-day milk yield. *Arq. Bras. Med. Vet. Zootec*., v.73, p.949-954, 2021a.

IBM SPSS statistics 21 algorithms. Chicago: IBM SPSS Inc., 2012.

JOHNSON, J.D. *Applied multivariate data analysis*. New York: Springer-Verlag, 1991.

MENDEŞ, M. Determination of minimum sample size for testing effect of ındependent variables in multiple linear regression analysis: a Monte Carlo simulation study. *Türkiye Klinikleri Biyoistatistik*, v.1, p.38-44, 2009.

MENDEŞ, M. Re-evaluating the Monte Carlo simulation results by using graphical techniques. *Türkiye Klinikleri J. Biostatistics*, v.13, p.28-38, 2021.

MENDEŞ, M. *Statistical methods and experimental design*. İstanbul: Kriter Yayınevi, 2019.

OSHİMA, T.C.; DELL-ROSS, T. All possible regressions using IBM SPSS: a practitioner's guide to automatic linear modeling. 2016. In: GEORGİA EDUCATİONAL RESEARCH ASSOCİATİON CONFERENCE. *Proceeding...* Georgia: GERA, 2016.

RAHNAMA, H.; RAJABPOUR, S. Identifying effective factors on consumers' choice behavior toward green products: the case of Tehran, the capital of Iran. *Environ. Sci. Pollut. Res.*, v.24, p.911-925, 2016.

TEMİZHAN, E.; MİRTAGİOĞLU, H.; MENDEŞ, M. Which correlation coefficient should be used for ınvestigating relations between quantitative variables? *Am. Acad. Sci. Res. J. Eng. Technol. Sci*., v.85, p.265-277, 2022.

YAKUBU, A.; OLUREMİ, O.I.A.; EKPO, E.I. Predicting heat stress index in Sasso hens using automatic linear modeling and artificial neural network. *Int. J. Biometeorol*., v.62, p.1181-1186, 2018.

YAN, X.; SU, X.G. *Linear regression analysis*: theory and computing. Singapore: World Scientific Publishing Co. Pte., 3009. 315p.

YANG, H. The case for being automatic: ıntroducing the automatic linear modeling (LINEAR) procedure in SPSS Statistics. *Multiple Linear Regression Viewpoints*, v.39, p. 27-37, 2013.