http://dx.doi.org/10.1590/1678-4162-12413

Arq. Bras. Med. Vet. Zootec., **v.73, n.6, p.1391-1402, 2021**

Evaluating performance and determining optimum sample size for regression tree and automatic linear modeling

[*Avaliando o desempenho e determinando o tamanho ideal da amostra para árvore de regressão e modelagem linear automática*]

S. *Genç*[1], M. *Mendeş*[2]

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

ABSTRACT

This study was carried out for two purposes: comparing performances of Regression Tree and Automatic Linear Modeling and determining optimum sample size for these methods under different experimental conditions. A comprehensive Monte Carlo Simulation Study was designed for these purposes. Results of simulation study showed that percentage of explained variation estimates of both Regression Tree and Automatic Linear Modeling was influenced by sample size, number of variables, and structure of variance-covariance matrix. Automatic Linear Modeling had higher performance than Regression Tree under all experimental conditions. It was concluded that the Regression Tree required much larger samples to make stable estimates when comparing to Automatic Linear Modeling.

Keywords: Regression Tree, Automatic Linear Modeling, simulation, biased, data mining

RESUMO

Este estudo foi realizado com dois objetivos: comparar os desempenhos da Árvore de Regressão e da Modelagem Linear Automática e determinar o tamanho ideal da amostra para estes métodos sob diferentes condições experimentais. Um abrangente Estudo de Simulação de Monte Carlo foi projetado para estes propósitos. Os resultados do estudo de simulação mostraram que a porcentagem de estimativas de variação explicada tanto da Árvore de Regressão como da Modelagem Linear Automática foi influenciada pelo tamanho da amostra, número de variáveis e estrutura da matriz de variância-covariância. A Modelagem Linear Automática teve um desempenho superior ao da Árvore de Regressão em todas as condições experimentais. Concluiu-se que a Árvore de Regressão exigia amostras muito maiores para fazer estimativas estáveis quando comparada à Modelagem Linear Automática.

Palavras-chave: Árvore de Regressão, Modelagem Linear Automática, simulação, parcialidade, pesquisa de dados

INTRODUCTİON

Thanks to advances in science and technology, scientists and researchers can establish more complex experiments. As a result, they must work with large and complex data sets. Discovering or extracting hidden and interesting knowledge from large amounts of data sets in a correct and reliable way is extremely important. This is because this knowledge contributes a lot of benefits for the fields of medical, agricultural, environmental, genetics, biological, economic, social, and business strategies. Therefore, studying with large and complex data sets might enable us to find the answers to many different questions and lead to obtain more detail and reliable information about the phenomena on the condition that an appropriate statistical technique is used in data analysis. Although working with large and complex data sets provides us important

Corresponding author: serdargenc1983@gmail.com
Submitted: April 26, 2021. Accepted: June 18, 2021.

advantages in terms of obtaining more detailed and reliable information about a subject of interest, it also brings some difficulties (Hill *et al*., 2004; Larose, 2005; Cios *et al*., 2007; Mendeş and Akkartal, 2009; Witten *et al*., 2011; Ratner, 2012; Kolaczyk, 2013; Fan *et al*., 2013). An important part of these difficulties is related to determining the appropriate methods or approaches to be used in the statistical analysis of the data and determining optimum sample size. Since commonly used traditional statistical methods, tests, and approaches do not scale to massive and complex data sets, usage of traditional methods in analyzing these kinds of data sets will not be convenient. Therefore, to handle the challenges of large and complex data or big data, new statistical thinking and computational methods are needed. Since data mining and machine learning methods or algorithms have a great potential for analyzing large and complex data sets, they can be effectively used in all branches of sciences for this purpose. In practice, the Regression Tree (RT) method is the one which is widely used in order to predict a continuous dependent variable and to determine the important factors affecting it. Another method, which is not as commonly used as RT, but has started being used especially in recent years, is Automatic Linear Modeling (ALM) (IBM SPSS, 2012; Field, 2013; Yang, 2013; Rahnama, 2016). Automatic Linear Modeling refers to a data mining approaches like Regression Trees, which utilizes a machine learning approach to find the best predictive model using the available data. Therefore, RT and ALM can be considered as alternatives to each other. The important point here is determining the performances of these methods under different experimental conditions and then revealing how the differences in the experimental conditions affect the performances of these methods. Conversely, comparing the performances of different techniques, algorithms or approaches which can be used for the same purpose is an important issue. Since sample size is one of the important factors affecting reliability of the results and stability of estimates, it is extremely important to determine optimum sample size for these methods to get reliable results and make stable estimates. As a result, determining proper sample size is one of the other important issues to be considered. It is because that way, it will be possible both to determine the most suitable methods of algorithm in analyzing data set and required proper sample size. This is only possible with a comprehensive simulation study. When the literatures is examined, it is seen that researchers generally try to compare the performances of different data mining techniques or machine learning algorithms through only one data set. Although this is a widely used application, it is not sufficient for the reliability and stability of the results. Because there are many factors (such as p, n, correlation) that can affect the performances of these algorithms, and thus it will not be possible to investigate the effects of these factors when only a single data set is considered. In light of these points, this study has basically two goals a) To compare the performances of Regression Tree and Automatic Linear Modeling under different experimental conditions via a comprehensive Monte Carlo Simulation Study and thus to determine which method gives more reliable results under which experimental conditions, and b) To determine optimum sample size.

MATERIAL AND METHODS

The material of this study is the random numbers generated from multivariate normal distribution by Monte Carlo simulation technique. RNMVN function of IMSL library of Microsoft FORTRAN Developer Studio were used in generating random numbers. R^2, accuracy, and the rank of the place of importance of predictors were considered as performance criteria. In order to determine reference or actual R^2 and accuracy values, 1,000,000 random numbers were generated from multivariate normal distribution under three different variance-covariance matrix structures and these random numbers were transferred to SPSS. Then, Regression Tree and Automatic Linear Modeling were performed based on these random numbers and R^2 and accuracy values were computed. These values were accepted as the reference or actual values of the R^2 and accuracy. Then, for the number of variables of 5, 10, and 15, different samples based on sample sizes were sampled from 1,000,000 random numbers and the RT and ALM procedures were applied to those samples and the R2 and accuracy values were estimated. These processes were repeated 500 times. So, each estimation was made based on 500 trials. Then, in order to determine proper or optimum sample size for the RT and ALM, estimated values versus reference value was graphed and thus it was possible to evaluate the effect of sample size on reliability and stability of the results. Correlations between the predictors

ranged from -0.25 to 0.85. 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:

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

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

Since 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 below:

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

Table 1. Simulation Study Characteristics

Sample sizes	Number of Variables (P)	Performance Criteria	Correlation range for predictors	Simulation Number
500, 1000, 2000, 3000 4000, 5000, 10000 20000, 30000, 40000 50000 Reference: 1000,000	5, 10, 15	1. R^2 2. Accuracy (R^2_{adj}) 3. Rank of the place of importance of predictors	[-0.25, 0.85]	500

Automatic Linear Modeling (ALM), is considered a relatively a new method, introduced in SPSS software (version 19 and up), enabling researchers to select the best subset automatically especially when there are a large numbers of variables. In ALM, the predictor variables are automatically transformed in order to provide an improved data fit, and SPSS uses rescaling of time and other measurement values, outlier trimming, category merging and other methods for the purpose (IBM SPSS Inc., 2012; Yang, 2017).

Regression Tree Analysis (RTA) is a recursive partitioning method that helps researchers to predict continuous response, to determine the most important variables in data set, and can help researchers craft a potent explanatory model. Since it does not require any priori assumptions about the nature of the relationships among the dependent and independent variables, it allows for the possibility of interactions and nonlinearities among the variables. The RT has clear advantages over classical statistical methods (Breiman *et al*., 1984; Moore *et al*., 1991). Due to its advantages, it has become increasingly popular for all branches of sciences especially in the presence of large and complex data sets.

RESULTS

Results of Monte Carlo Simulation studies for P=5, 10, and 15 have given in Table 2 and Figure 1-6, respectively.

When the RT and ALM were compared in terms of their performances, it was observed that although both methods tend to give similar results as sample size increased, the ALM method generally showed a better performance. For example, R^2 values estimates for RT methods were varied between 60.72 and 69.75% for P=5, 57.71 and 73.69% for P=10, and 67.11 and 81.04% for P=15. Maximum deviations from the referenced value were observed for sample sizes of 5000 and less. For the ALM method, accuracy or adjusted R^2 estimates were varied between 73.25 and 78.92% for P=5, 84.53 and 86.60% for P=10, and 85.03 and 88.80% for P=15. As in the RT, the maximum deviations from the referenced value were observed for sample sizes of 2000 and less for ALM as well.

Table 2. Simulation Results for P=5, 10, and 15

Sample Size	P=5			P=10			P=15		
	Risk	R^2	Accuracy	Risk	R^2	Accuracy	Risk	R^2	Accuracy
500	31.53	60.72	73.25	31.99	59.09	84.53	685.50	67.11	85.03
1000	31.50	64.89	74.27	33.92	57.71	85.3	564.10	71.12	85.48
2000	27.89	66.31	75.57	27.87	66.33	86.34	561.10	75.24	86.8
3000	29.59	64.45	75.53	26.36	67.28	86.43	556.20	75.85	87.13
4000	29.35	64.62	75.83	28.71	65.28	86.5	539.20	76.42	87.44
5000	28.43	64.69	75.85	26.92	68.71	86.47	525.77	76.19	87.53
10000	26.56	68.40	76.7	25.42	71.15	86.43	505.05	78.17	87.47
20000	25.72	68.47	77.68	25.33	72.33	86.5	485.18	78.85	88.8
30000	25.57	68.09	77.5	23.23	72.97	86.42	466.95	79.22	88.93
40000	25.14	69.75	78.2	23.05	73.69	86.52	463.77	80.35	88.87
50000	24.91	69.35	78.92	25.12	73.60	86.54	464.03	81.04	88.77
Reference	25.99	68.36	75.60	23.15	72.26	86.40	465.42	79.82	87.56

Note: Reference values were obtained based on 1,000,000 simulation runs

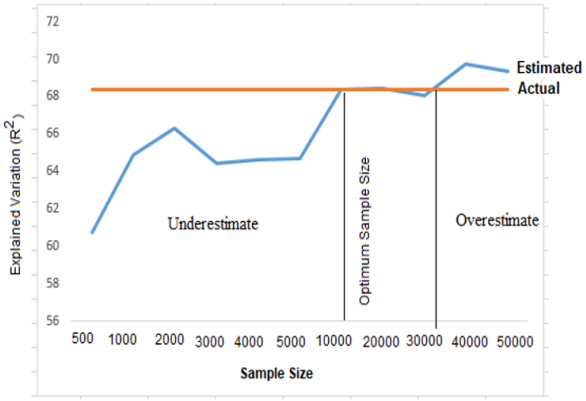

Figure 1. Estimates versus actual values for RTM when p=5.

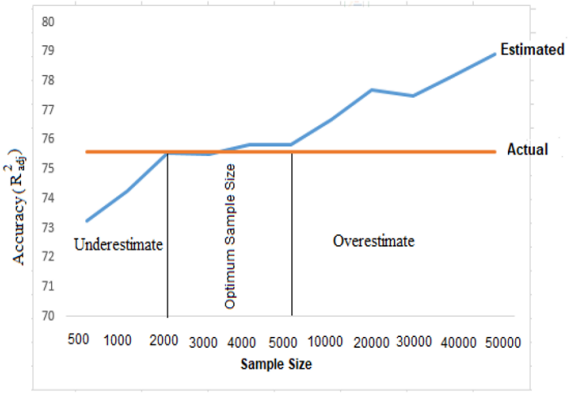

Figure 2. Estimates versus actual values for ALM when p=5.

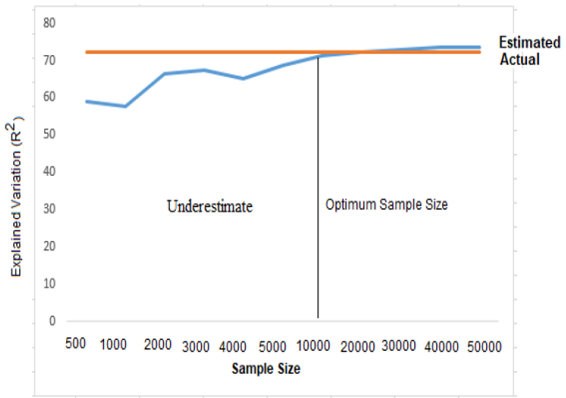

Figure 3. Estimates versus actual values for RTM when p=10.

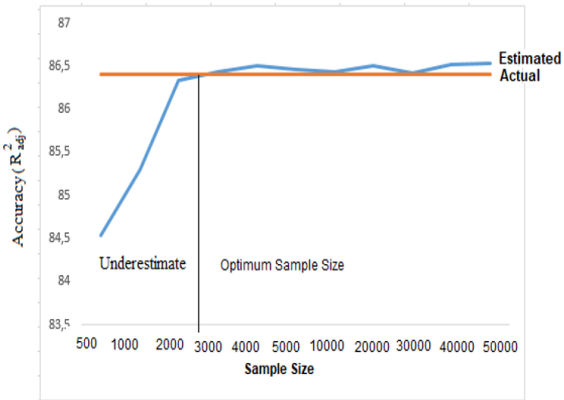

Figure 4. Estimates versus actual values for ALM when p=10.

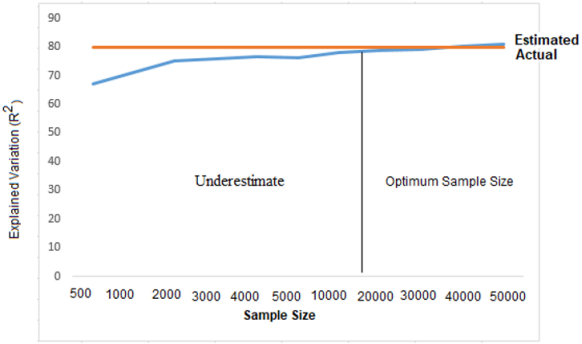

Figure 5. Estimates versus actual values for RTM when p=15.

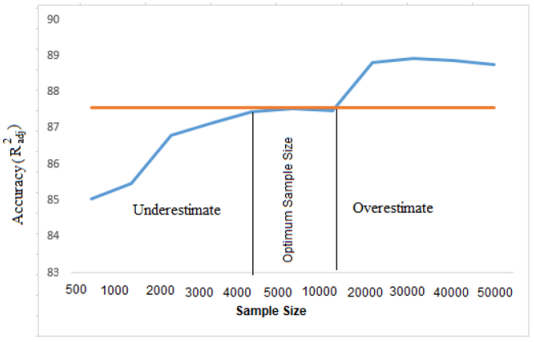

Figure 6. Estimates versus actual values for ALM when p=15.

Since both RT and ALM required large data sets, it is extremely important to determine proper or optimum sample size based on number of variables and correlations between the variables. Therefore, it is necessary to study with sufficient sample size to get the expected benefits from these methods. Results of this simulation study showed that although optimum sample size for both methods was influenced by number of variables and structure of the variance-covariance matrix, optimum sample sizes were not the same for RT and ALM. For the RT, optimum sample size was determined as 10000 for P=5, 20000 for P=10 and more predictors. As it can be noticed, as the number of variables increased, required sample size was also increased. Therefore, it is possible to suggest researchers' study with at least sample sizes of 10000 when they have 5 predictors, 20000 when they have 10 and more predictors in order to get more reliable and stable estimates. As in the RT, the required sample size increases as the number of variables increased in the ALM. However, the ALM method required much less samples to give reliable results and make stable estimates when compared to the RT. For example, for P=5 at least 2000 observations were needed while for P=10 and 15 at least 10000 observations were needed.

For both methods, as the sample size increased the rank of the place of importance of the predictors tended to be remained constant. It was observed that the rank of the places of the importance of the predictors, whose importance value was above 0.45, remained constant and unchanged. However, as a result of the RT analysis, the place of 20% of the variables whose importance values ranges between 0.30 and 0.45 were changed, while this ratio was maximum 10% for ALM. In general, variables with importance value below 0.15 were displaced and this situation became more evident especially when n<3000. However, it should not be ignored that this situation may vary based on experimental conditions like structure of variance-covariance matrix, multicollinearity problem, sample size, number of variables, and type of variables.

A part of questionnaire study which aimed at determining factors that affect the grade success of university students was used. In this questionnaire study, 15 of the 30 questions asked were taken into consideration (Keskin and Mendeş, 2019). Regression Tree and Automatic Linear Modeling Techniques were applied this data set to determine the factors affecting grade success of university students and estimate the grade success. SPSS (ver 22.0) was used in performing Regression Tree and Automatic Linear Modeling.

Results of RT have been presented in Figure 7. Figure 7 was determined as an optimal tree based on risk value, its standard error, and explained variation in the grade success. Explained variation percentage (R^2) of the optimal tree is found to be 67.2%, meaning that the optimal tree can explain 67.2% of variation in the grade success (dependent variable).

When optimal tree is examined, it is seen that this tree has been formed by using five factors namely if the students had a graduation plan, whether they read book or not, their place of residence, type of the study for preparing exam, and how they define themselves. Using these factors, seven terminal nodes were formed and the students in these nodes were accepted as homogenous in terms of their

grade average. In keeping with the tree analogy, the regions R_1, R_2, …, Rj are known as terminal nodes (homogenous subsets) or leaves of the tree. The purpose is to find terminal nodes that minimize the Residual Sum of Square (RSS):

$\sum_{j=1}^{J}\sum_{i\in R_j}(y_i - \hat{y}_{R_j})^2$,

where $\hat{y}_{R_j}$ is the mean of response for the training observations within the j[th] terminal nodes.

As a result, it is possible to conclude that in order to increase the grade, success in the student's five factors and their interactions should be taken into account and they must be improved. As it can be seen from the optimal tree, the students who wanted to be academician had the highest grade average. However, the proportion of the students who wanted to be academician is only 5.4% and these results are very realistic. As a result, by using Regression Tree Analysis Technique in data analysis, it was possible to determine the factors that affect student success the most and to investigate the effects of higher order interaction.

Figure 7. Optimal tree for predicting grade average estimates of students.

Results of the Automatic Linear Model (ALM) are presented in Figure 8, 9, and 10. In determining an appropriate model for fit our data set many models have been run (not discussed here) and it has been observed that except for the model that has been used in this study (-7.748), the other models have large information criterion values and above). The accuracy level of the model which is equivalent to the Adjusted R-squared value used to fit data and estimate the changes in Grade Success is 73.1%, meaning that this model can be used in fitting and estimating processes (Figure 8).

The lower the information criterion (AIC) is, the better the model is compared to models with a higher information criterion. Since the model used here has the lowest information criterion value compared to many other models (not discussed in this document), this model has been preferred in investigating the relations between Grade Success and predictors.

Importance levels of the predictors have been presented in Figure 9. Figure 9 shows the predictors in the final model in rank order of importance. For linear models, the importance of a predictor is the residual sum of squares with the predictor removed from the model, normalized so that the importance values sum to 1. When Figure 9 is examined, it is seen that the most importance variables or factors that affect the Grade Success of the students are Graduation Plan, Place to Residence, Exam, Self-identification, Appreciated, and Reading Book. Therefore, it is possible to conclude that the factors related to these factors should be taken into consideration in order to get reliable and stable estimates.

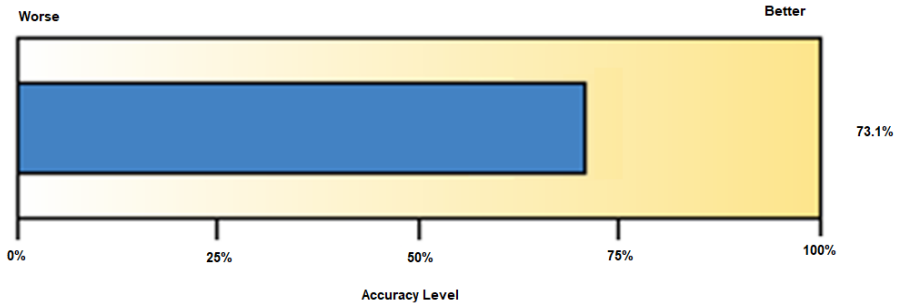

Figure 8. Accuracy level of the model.

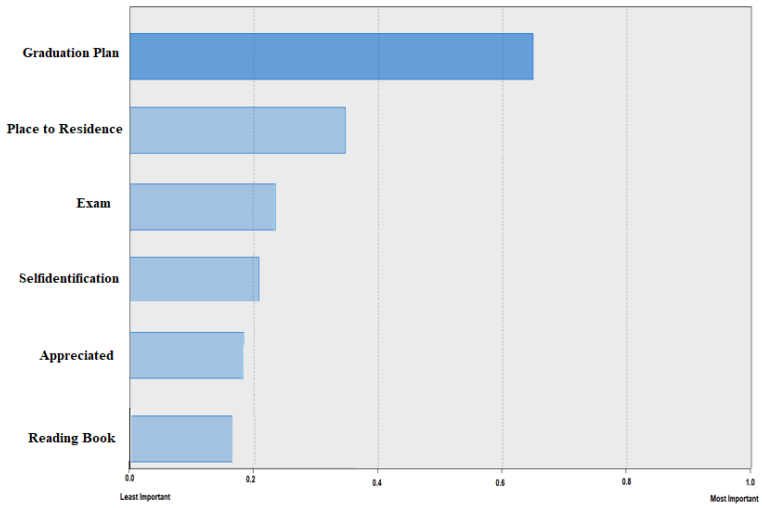

Figure 9. Importance levels of the variables or predictors.

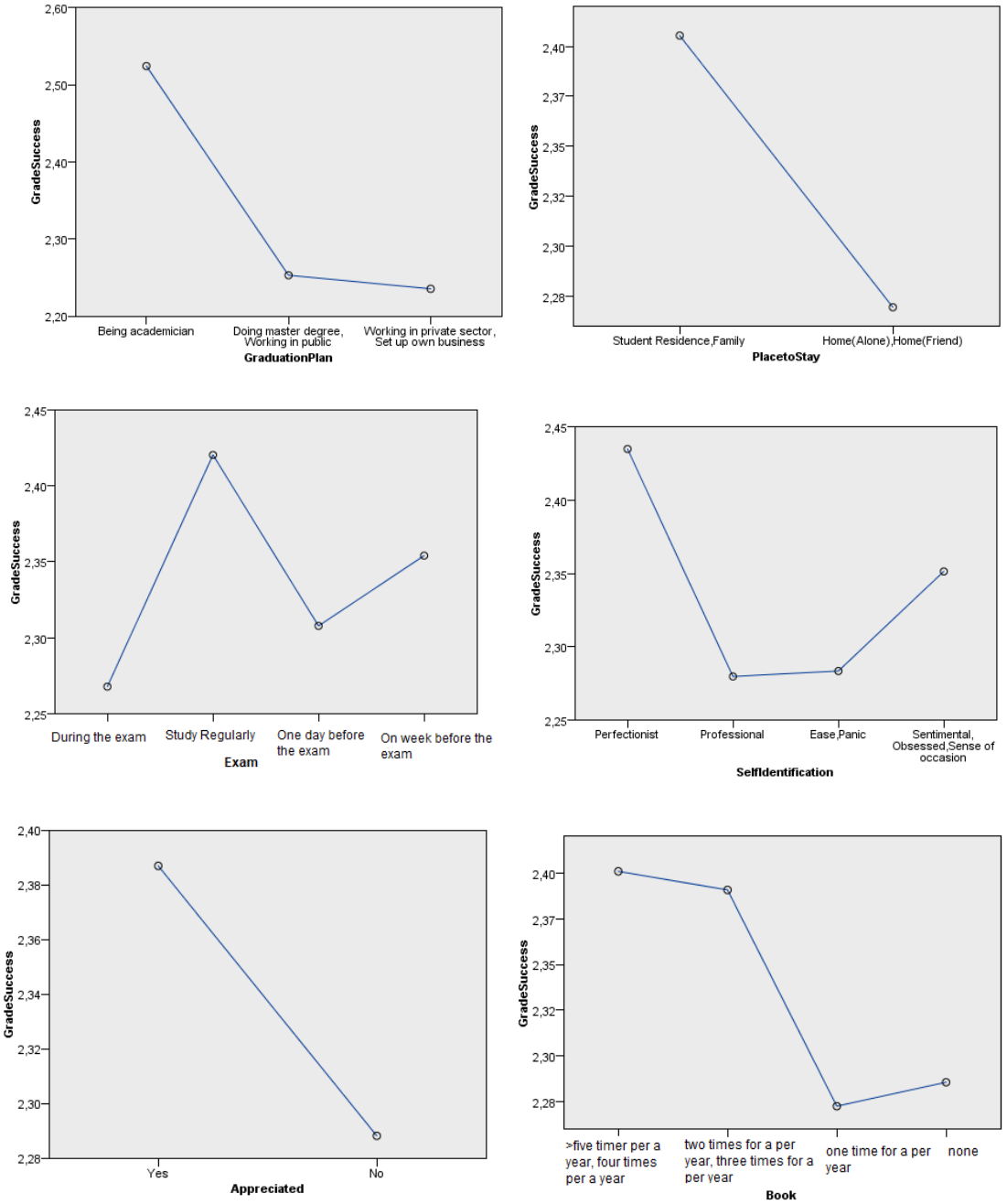

Figure 10. Individual effects of importance predictors.

When the Figure 10 is examined, it is seen that students who want to be an academician, living with their family, who are appreciated, study regularly and read more books are more successful.

When results of RT and ALM are evaluated together, although it is observed that the explained variation as a result of ALM is slightly higher and there are some differences in the number of important independent variables and their

importance rankings, it is possible to say that both methods generally yield similar results.

DISCUSSION

Although studying with large and complex data set brings new opportunities to modern society, it also brings some challenges to data scientists or statisticians. One of these challenges is how to extract beneficial and useful information from the big data set. Regression based model, algorithms, or methods are commonly used for this purpose.

Thanks to advances in science and technology, different regression methods that can be used in the analysis of large and complex data sets have been developed. The Classification and Regression Tree (CART) is probably the most well-known decision tree learning algorithm in the literature (Breiman *et al*., 1984; Miloslava *et al*., 2008; Mendeş and Akkartal, 2009). Since the CART can statistically show which factors are important in a model or relationship in terms of explanatory power and variance, it has become more popular and especially it has been commonly used in multidisciplinary fields (Lin *et al*., 2008; Kaur and Pulugurta, 2008). One of the other reasons why this method became increasingly used in practice is that since it is a graphical technique, understanding and interpreting the results of the CART is very easy and enables the research especially for non-statisticians to evaluate higher order interactions more easily (Mendeş and Akkartal, 2009; Morgan, 2014; Yang *et al*., 2017). When advantages and characteristics of the CART are evaluated, it is possible to say that the CART is a simple but a powerful analytic tool that helps the researchers to determine the most important factors based on explanatory power in data sets. As a result, researchers may have the opportunity to create a powerful explanatory model.

It is noteworthy that especially in recent years a relatively new method especially for the non-statisticians has been developed for this purpose, but not as popular as Regression Trees, is Automatic Linear Modeling (ALM). Therefore, both RT and ALM methods are kinds of data mining, and they might be evaluated alternatives to each other. At this point, it is extremely important to determine the most appropriate statistical tests, methods or approaches in analyzing a data set. Since the performance of these tests can be affected by the sample size, determining the appropriate sample size is another important issue to consider.

Although data mining techniques require large and complex data sets, on the other hand, studying with very large samples cannot be appropriate. It is because it will be difficult to understand and interpret the results of large decision trees and as can be seen from the results of this study an overfitting problem may occur (Figure 1-6). However, it is not easy to determine proper sample size for decision tree algorithms, because there are many different factors that should be considered, and each decision tree algorithm has its own property in generating trees (Sug, 2009). Although a proper sample size for a variable or feature is generally accepted as 30 or so in statistics, this rule may not be valid for data mining algorithms since the target databases of data mining contain a lot of features. If the sampling is done based on this rule, the sample size can become enormous. Therefore, an alternative strategy or a new sampling thinking is needed for sampling. Based on this reality, in this study, sample size that correspond to the estimated values which the least deviated from the actual value that computed on 1000,000 observations was accepted as optimum or proper sample size.

In this study, two performance criteria namely R^2 and accuracy value or adjusted R^2 were used in determining optimum sample size. Considering these criteria, it was seen that the performances of the ALM test was better than that of the RT under all experimental conditions. Minimum and maximum R^2 estimates for the RT were 57.71% and 81.04% while the minimum and maximum accuracy estimates of the ALM were 73.25% and 88.93%. Results of this study suggested that optimum sample size for Regression Tree is obviously larger than that of the Automatic Linear Modeling. While required minimum sample sizes for the ALM in order to make stable estimations varied between 2000 and 10000 based on number of variables or dimensionality and structure of variance-covariance matrix, they were varied between 10000 and 30000 for the RT. This may be due to differences in step of procedures used in these methods. Sug (2009) in his study, suggested a progressive approach in determining a proper sample size to generate good decision trees with respect to generated tree size and accuracy and he

reported that experiments with two representative decision tree algorithms, CART and C4.5 showed very promising results.

In this study, we tried to determine proper sample size by using explained variation and accuracy values under different variable numbers and variance-covariance matrix structures. The minimum sample size, which gave the closest estimate to the reference value was considered the optimal sample size. As a result, it was observed that sample sizes between 10000 and 30000 might be able to accept as optimum sample size for the RT when P=5 and 10. As it can be seen from Figure 1, studying with the sample size smaller than 10000 caused underestimation problem while studying the sample sizes of larger than 30000 caused overfitting problem. Therefore, it can be concluded that in cases where the number of variables is between 5 and 10, it is necessary to work with samples of at least 10000 sizes in order to achieve reliable results or to make unbiased estimates. As it can be seen from Figure 1, studying with the sample size smaller than 10000 caused underestimation problems while studying the sample sizes larger than 30000 caused overfitting problems. Therefore, the sample sizes between 10000 and 30000 might be able to accept as optimum sample size for the RT when P=5 and 10. Based on these findings, it can be concluded that in cases where the number of variables is between 5 and 10, it is necessary to work with samples of at least 10000 size in order to achieve reliable results or to make unbiased estimates under these conditions. For P=15 or more, it is possible to suggest that required minimum sample size should be at least 20000 in order to get reliable results or to make unbiased estimates. ALM method, on the other hand, has provided reliable results and stable estimates even when studying with much smaller samples when compared to RT. As it can be seen from Figure 2, the sample sizes between 2000 and 5000 might be accepted as optimum or sufficient sample sizes for the ALM method. As it can be noticed that the ALM method only produced unreliable results when sample size was smaller than 2000 (especially if sample size was smaller than 1000). The ALM tended to overestimate the actual value when sample size was larger than 5000. Domingos (1998) and Oates and Jensen (1997 and 1998) reported that decision tree based data mining tools were subject to over-fitting as the size of the data set increased. As it can be noticed from the results of this study (Figure 1-6), as the sample size increased overfitting problem occurred while underestimate problem occurred as the size of the data decreased. Therefore, the results of this study are compatible with the findings declared by Domingos (1998) and Oates and Jensen (1997). Morgan *et al*. (2003) reported that model accuracy improves at a decreasing rate with increasing sample size. When a power curve was fitted to accuracy estimates across various sample sizes, more than 80 percent of the time accuracy within 0.5 percent of the expected terminal (accuracy of a theoretical infinite sample) was achieved by the time the sample size reached 10,000 records. Although the idea that "the larger the sample size is increased, the more reliable results are obtained" in general, this idea is not particularly valid in practice. Because the time and cost allocated for a study must be taken into consideration. Therefore, studying with appropriate sample size will enable us to get reliable results and to make stable estimates by considering both factors of cost and time. Morgan *et al*. (2003) in their study reported that the relationship between sample size and model accuracy is an important issue for data mining and this relation should not be ignored since model accuracy improves at a decreasing rate with increasing sample size. Despite the increases in processing speeds and reductions in processing cost, applying data mining tools to analyze all available data is costly in terms of both economy and time required to generate and implement models. As it can be seen from the results of this study, R^2 and accuracy values increased with sample size. Similar results were also reported by Morgan *et al*. (2003).

Oates and Jensen (1998) reported that increasing the amount of data used to build a model often results in a linear increase in model size, even when that additional complexity results in no significant increase in model accuracy. Frey and Fisher (1999) in their study systematically examined the response of modeling accuracy to changes in sample size using the C4.5 decision tree algorithm applied to 14 datasets from the UCI repository. Although they did not focus on determining an optimal sample size, they found that the response of predictive accuracy to sample size was more accurately predicted by a regression based on a power law function than by regressions using linear, logarithmic, or exponential functions. As a result of this

simulation study, it was found that the performance of the ALM method was higher, and it could enable reliable estimates with samples with a smaller size compared to the RT method. The results of this study support the results of studies by Morgan (2003), Oates and Jensen (1998), Frey and Fisher (1999). Mannila (2000) suggested that the volume of the data was probably not a very important difference. He reported that the number of variables often had a much more profound impact on the applicable analysis methods in discussing differences between statistical and data mining approaches. However, the best way for getting reliable results and making stable estimates is considering both effect of sample size and number of variables simultaneously along with other factors like correlations between predictors.

When a general evaluation is made based on simulation results, it is possible to conclude that although decision trees are generally known as one of the most successful data mining tools, they may not always be the best or not produce reliable results due to being built based on some insatiable algorithms for limited or small data set. Therefore, comparison of the methods, algorithms or tools which may be used for the same purpose via a comprehensive simulation study will be beneficial for both evaluating their performance and to determine optimum sample size to get reliable results and stable estimates under different experimental conditions. As a result, it is possible to reach the following conclusions without ignoring the fact that differences in experimental conditions (i.e. number of variables, measurement types of variables, correlations between predictors etc.) may affect the reliability of the results to be obtained.

a) Based on our observations, it is possible to say that although there is not an exact consensus among statisticians yet about analyzing large and complex data sets and what statistics should be used to extract useful information from these sets in correct and reliable way. However, we believe that data mining and machine learning algorithms can be efficiently used for this purpose. As a matter of fact, both my previous studies on data mining methods and the findings of this study support this belief.

b) The ALM has The RT and ALM might be considered as alternatives to each other especially for predicting response, determining the important factors affecting response, and evaluate higher order interactions.

c) The RT requires much more samples when comparing to ALM. The minimum required sample size for the RT test should be approximately twice that of the ALM test.

d) Although it was observed that the ALM was much more powerful than RT for estimating continuous response and determining important factors, a potential threat of misuse of the ALM due to its simplicity should not be ignored.

REFERENCES

BREIMAN, L.; FRIEDMAN, J.; OLSHEN, R. *et al*. *Classification and regression trees.* [s.l.]: Wadsworth International Group, 1984.

CIOS, K.; PEDRYCZ SWINIARSKI, W.R.W.; KURGAN, L.A. *Data mining*: a knowledge discovery approach. New York: Springer, 2007.

DOMINGOS, P. *Occam's Two Razors: the Sharp and the Blunt*, Poceedings of the Fourth International Conference on Knowledge Discovery and Data Mining, AAAI Press, p.37-43, 1998.

FAN, J.; HAN, F.; LIU, H. Challenges of big data analysis. *Natl. Sci. Rev.*, v.1, p.293-324, 2014.

FREY, L.; FISHER, D. Modeling decision tree performance with the power law. In: INTERNATIONAL WORKSHOP ON ARTIFICIAL INTELLIGENCE AND STATISTICS, 7., 1999, San Francisco. *Proceedings…* San Francisco: Morgan-Kaufmann, 1999. p59-65.

HILL, C.M.; MALONE, L.C.; TROCINE, L. Data mining and traditional regression. In: BOZDOGAN, H. (Ed.). *Statistical data mining and knowledge discovery.* Boca Raton: CRC Press LLC., 2004. p.233-249

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

KAUR, D.; PULUGURTA, H. comparative analysis of fuzzy decision tree and logistic regression methods for pavement treatment prediction. *WSEAS Trans. Comp.*, v.5, p.979-988, 2008.

KESKIN, E.; MENDEŞ, M. A Graphical perspective for determining factors affecting grade success of students. *III.* In: INTERNATIONAL CONFERENCE ON AWARENESS, 3., 2019, Çanakkale. *Proceedings*… Çanakkale/Turkey: [s.n.], 2019. p.31-35.

KOLACZYK, E. Statistical analysis of Network data. In: YES STATISTICS FOR COMPLEX AND HIGH DIMENSIONAL SYSTEMS, 6., 2013, Eindhoven. Proceedings… Eindhoven: [Nirvana & WordPress],2013.

LAROSE, D.T. *Discovering knowledge in data*: an introduction to data mining. Hoboken, NJ: John Wiley & Sons, Inc., 2005.

LIN, Z.; CHEN, Y.; LIANG, Y. *Application of data mining classification algorithm for customer membership card model*. China: College of Transportation and Management, 2008.

MANNILA, H. Theoretical frameworks for data mining. *Sigkdd Expl*., v.1, p.30-32, 2000.

MENDEŞ, M.; AKKARTAL, E. Regression tree analysis for predicting slaughter weight in broilers. *Ital. J. Anim. Sci.*, v.8, p.615-624, 2009.

MILOSLAVA, K.; JIR, I.K.; PAVEL, J. Application of decision trees in problem of air quality control in the Czech Republic locality. *WSEAS Trans.Comp.*, v.7, p.1166-1175, 2008.

MOORE, D.M.; LEES, B.G.; DAVEY, S.M. A new method for predicting vegetation distributions using decision tree analysis in a geographic information system. *Environ. Manag.*, v.15, p.59-71, 1991.

MORGAN, J. Classification and regression tree analysis. 2014. Technical Report No. 1, Available in: https://www.bu.edu/sph/files/2014/05/MorganCART.pdf, 2014. May 8. Accessed in: 8 May 2019.

MORGAN, J.; DAUGHERTY, R.; HILCHIE, A. *et al*. Sample Size and Modeling Accuracy of Decision Tree based Data Mining Tools. *Acad. Inf. Manag. Sci. J.*, v.6, p.77-92, 2003.

RAHNAMA, M.R. Estimating housing prices using automatic linear modeling in the metropolis of Mashhad, Iran. *Int. J. Econ. Manag. Eng.*, v.10, p.2242-2253, 2016.

RATNER, B. *Statistical and machine-learning data mining*: techniques for better predictive modeling and analysis of big data. 2.ed. Boca Raton, FL: Taylor & Francis Group, 2012.

SUG, H. An effective sampling method for decision trees considering comprehensibility and accuracy. *Wseas Trans. Comp.*, v.4, p.631-640, 2009.

OATES, T.; JENSEN, D. The effects of training set size on decision tree complexity. MACHİNE LEARNİNG: INTERNATİONAL CONFERENCE, 14., 1997. *Proceedings*... [s.l.]: Morgan Kaufmann, 1997. p.254-262.

OATES, T.; JENSEN, D. Large Data sets lead to overly complex models: an explanation and a solution. INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, 4., 1998, Menlo Park. *Proceedings*… Menlo Park, CA: AAAI Press, 1998. p.294-298.

WITTEN, I.H.; FRANK, E.; HALL, M.A. *Data mining*: practical machine learning tools and techniques. 3.ed. Burlington, MA: Elsevier Inc., 2011.

YANG, L.; LIU, S.; TSOKA, S.; PAPAGEORGIOU, L.G. A regression tree approach using mathematical programming. *Exp. Syst. Appl.*, v.78, p.347-357, 2017.