Int. J. Nonlinear Anal. Appl. 15 (2024) 7, 325–336
ISSN: 2008-6822 (electronic)
http://dx.doi.org/10.22075/ijnaa.2023.30604.4447

A review of methods for estimating coefficients of objective functions and constraints in mathematical programming models

Ali Ramezani, Seyed Mohammad Ali Khatami Firouzabadi*, Maghsoud Amiri

Department of Industrial Management, Faculty of Management, University of Allameh Tabatabai, Tehran, Iran

(Communicated by Nallappan Gunasekaran)

Abstract

Considering the high importance of the optimization problem, this study evaluated mathematical programming models by considering various methods of estimating model coefficients. Correct and accurate data must be entered into the model to get accurate and robust result from the model. Most input data to the presented model are technical and objective function coefficients. Therefore, it is necessary to determine the information related to these coefficients with the utmost precision and, as much as possible, to develop a suitable scientific method to estimate the value of these coefficients [5]. Finding the best method for estimating the coefficients of mathematical programming models can significantly optimize the final values of the variables extracted from the mathematical programming model. For this reason, it is essential to study the methods used so far in this field and examine their advantages and disadvantages. This review study investigated various methods of estimating technical coefficients of mathematical planning models in the conditions of possible decision-making and uncertainty after reviewing 117 articles published between 1955 and 2022. These methods include fuzzy methods, statistical methods, and data analysis methods. Statistical methods such as regression methods, time series methods, exponential smoothing, and linear non-linear and non-parametric, machine learning and data mining methods were investigated in this article. The methods of data-driven analysis explained in this article can be referred to as decision trees, random forests and the Lasso methods. After evaluating and comparing these methods, suggestions for choosing the best method were provided.

Keywords: mathematical programming model, data-driven analysis methods, data mining methods, objective function, technical coefficients, time factor, machine learning
2020 MSC: 68T45, 90B50

1 Introduction

This study examined mathematical planning models by considering various methods of estimating the coefficients of the operation research model due to the high importance of the optimization problem. Mathematical programming models, regardless of whether they are linear or non-linear (depending on the power of the variables that are of the first degree or more), have parameters and variables that shape the problem of researchers or industrial owners.

*Corresponding author
Email addresses: creativity_r@yahoo.com (Ali Ramezani), a.khatami@atu.ac.ir (Seyed Mohammad Ali Khatami Firouzabadi), amiri@atu.ac.ir (Maghsoud Amiri)

Received: April 2023 *Accepted:* June 2023

The first and most crucial step is the correct understanding of the problem, which begins with determining the goal or goals of optimization after the problem is raised. The variables used in reaching the goals are determined after the goals are determined. Optimization aims to estimate these variables' values to achieve the highest accuracy in the objective functions and constraints. After determining the variables and objectives, it is time for the second stage, which is to write the mathematical planning model. The target function or functions with the coefficients of each variable are reached using the variables and their impact on the target value or targets. Then, the limits of the definition problem are determined based on the employer's experience and the historical data of each company or organization and by assigning coefficients to the variables in different intervals to practically complete the conditions of the model. The most crucial task in mathematical programming (linear or non-linear) is determining the coefficients of the variables in the objective functions and constraints after determining the goal or goals and the influential variables in achieving the goals (decision variables). Determining the coefficients of the variables in the objective function and limitations was, in most cases, based on the employer's decisions and experience, which may have many errors, and it is not possible to consider all the essential factors or even most of them in the estimation of the coefficients.

Paying attention to the stages of the modeling process can be a way to increase the reliability of mathematical programming models. This process includes seven steps: defining the problem, collecting data, building the model, checking the accuracy of the model's performance, solving the model, presenting the study results to the organization, and finally, implementing and using the model [60]. Each of these seven steps can be used to improve the efficiency of mathematical programming models in solving problems. However, uncertainty in models is often caused by problems and limitations in the stages of data collection, model construction, and model solving. The possibility of solving mathematical programming models has become easier with the increasing efficiency and quality of computers and various software. What seems more important than the past is modeling the problem and not solving the model [40]. This paper focused on reducing the error due to uncertainty by improving the modeling process. Most decision-making problems in the real world are solved based on calculations which depend on assumptions data, but as explained above, it is practically impossible to assume the definiteness of parameter values and coefficients. The review of past studies indicates that many optimization methods have been used to overcome the conditions of uncertainty, such as stochastic programming, fuzzy programming, and interval programming methods [1, 2, 37].

Accurate and actual data should be entered into the model to get robust results from the model. technical and objective function coefficients. Therefore, it is necessary to determine the information related to these coefficients are inputted into these mathematical programming models with the utmost precision and, as far as possible, to develop a suitable scientific method to determine the value of these coefficients [5]. In addition, finding the best method for estimating the coefficients of mathematical programming models can play a significant role in optimizing the final values of the variables extracted from the mathematical programming model as is the main problem in this research.

This study reviewed the available methods for estimating the coefficients of objective functions and constraints in mathematical programming models and explained the data-driven optimization and statistical methods used in mathematical programming models.

Over the past years, many solutions have been presented to reduce the error of estimating the coefficients of variables in mathematical programming models. Using historical surveys to estimate coefficients through statistical distributions is one of these solutions that results in random programming and calculates each coefficient based on its hypothetical statistical distribution [15]. Stochastic or probabilistic programming deals with situations where some or all parameters of the optimization problem are expressed by random variables instead of definite quantities. The assumption that the matrix of coefficients is definite in real-life problems, will results in an estimate of the parameters on the right side of each function, and the coefficients of the objective function and constraints are mostly untrue. In general, these coefficients are often random in nature. Another method is to use the fuzzification of coefficients. The fuzzy set theory was first introduced by Zadeh [61] to overcome uncertainty in decision-making.

Instead of using numbers as definite coefficients, this approach defines fuzzy numbers as interval, triangular, or trapezoidal fuzzy. These fuzzy numbers reduce the estimation error somewhat by considering a range of numbers instead of a single number. The logic governing the type of coefficients of this method reduces its error compared to deterministic approaches, but the complete dependence of this method on the opinion of the employer or field expert and the determination of coefficients can still cause much error in the accurate estimation of coefficients. Factors affecting each coefficient are ignored, especially in the presence of the time factor. The combination of stochastic and fuzzy methods has been used in research to solve linear programming problems. Despite having more advantages than the two methods independently, this combined method still has disadvantages. Troutt, Pang, and Hou [54] proposed a method to estimate parameters based on minimum decision error. They focused on mathematical programming models with objective functions that depend linearly on other parameters. The proposed estimation method finds the parameter values in a way that the estimated values are as close as possible to the observed values. The researchers

called this behavioral technique estimation because their approach is based on the decision-making and behavior of company managers. As the authors acknowledge at the end of their article, this method has shortcomings in the presence of variables that change over time. The generalized estimating equations (GEE) method is one of the statistical methods in this article that is compatible with both types of variables (whether they change over time or not) and accurately models the changes and the impact of these changes on the final estimates. In random and statistical methods, there are always limitations and assumptions, such as having a specific distribution for the data (for example, the data must have a normal distribution), which severely limits and causes errors in calculations due to the impossibility of meeting these conditions in most cases. The GEE estimation method does not work based on the use of weight coefficients (or a range of weight coefficients) but uses covariance matrices to estimate the optimal coefficients, and as a result, it does not have the problems mentioned above. Another part discussed in this article is using data mining and machine learning methods to select the influential factors and estimate the coefficients of the objective functions and constraints in mathematical programming models.

Zhang [63], Lever et al. [30], and Rocks and Mehta [41] have shown that the use of too many variables will also lead to errors and lack of generalization of the results and that the use of too few variables in a model creates uncertainty in the estimates. Therefore, it was recommended not to use too few or too many variables in a model and to use the methods of choosing the optimal number of the most critical and influential variables, such as Random Forest and LASSO. These methods are also explained in detail in this article.

2 Types of decision-making conditions

These solutions can be categorized as follows with an overview of the types of decision-making conditions that are proposed in the decision theory [6]:

- Decision-making in conditions of certainty;
- Decision-making in possible situations;
- Decision-making in conditions of uncertainty.

Decision-making in confident or certain conditions is the traditional method of designing mathematical models, which this article does not need to explain again.

2.1 Decision-making in possible situations

Decision-making in probabilistic conditions uses historical surveys to estimate coefficients through statistical distributions that result in random programming and calculates each coefficient based on its hypothetical statistical distribution [15]. Stochastic or probabilistic programming deals with situations where some or all parameters of the optimization problem are expressed by random variables instead of definite quantities. In real-life problems, the assumption that the matrix of coefficients, the parameters on the right side of each model, and the coefficients of the objective function are determined is mostly not realistic. In general, these coefficients are often random in nature. Depending on the nature and type of problem, there may be several sources to explain random variables. The majority of past of research, the main idea in solving stochastic programming problems is to transform them into equivalent deterministic or certain problems. The resulting deterministic issues can be solved using well-known linear and nonlinear programming methods [1]. According to experts, this method has a high error due to the random behavior of the data, and it works even more when the data follows a normal distribution, which we know is an unrealistic assumption.

2.2 Decision-making in conditions of uncertainty

First, the fuzzy method, which has more history in mathematical modeling, is explained, and other methods are examined next.

2.2.1 Fuzzy method

The most prominent method in uncertain decision models is using fuzzy logic in estimating model coefficients. The decision maker can rarely express the exact weights or values while facing uncertainty or ambiguity. Sometimes it is impossible to determine the precise weights of the decision maker, and their estimation cannot be accurate because uncertainty and ambiguities around the problem will always accompany the decision problem. The expertise of several decision-makers makes the topic and model more efficient, but surely, one cannot be certain of the correctness of the judgment without involving their personal opinions in the issue [2].

In this approach, fuzzy numbers are used as coefficients instead of absolute numbers to give us a more accurate estimate and to be able to express the conditions more accurately [51]. Fuzzy logic is one of the most essential methods for explaining uncertainty. This logic originates from the inability of zero and one logic to include ambiguous language, common sense reasoning, and innovative solutions to problems used by ordinary people every day, and it can mathematically represent many imprecise and ambiguous concepts, variables, and systems [3]. The formulation of the fuzzy linear programming problem [65] is highly dependent on the opinion of the employer or technical experts, which is flawed and contains error in many ways and compensates for a limited amount of the error in the estimation [4]. A hybrid approach is also used in solving such problems in articles and research. For example, Naseri and Bavandi [37] solved linear programming problems with a stochastic and fuzzy approach, which used a combination approach of stochastic and fuzzy programming models. The most crucial uncertainty environments that are widely used today are fuzzy and random environments, but when faced with real-world situations, it is well understood that the complexity of the phenomena is such that it requires the simultaneous use of two random and fuzzy uncertainties [37].

Now that Common methods for estimating coefficients of objective functions and limitations of mathematical programming models were explained, it is time to consider methods that account for more uncertainty in estimating coefficients of objective functions and constraints. These methods are described below in two statistical and data mining subsets. The data analysis methods will be discussed, further.

2.2.2 Statistical methods

Based on the research in estimating the coefficients of objective functions or constraints, the statistical methods that have been used so far are divided into three general categories:

1. Simple regression methods without considering time changes
2. Time series methods that, although they include time changes, do not consider many other influential variables (the usual method is Auto Regressive Integral Moving Average-ARIMA) [50].
3. Data mining methods that are limited to computer science methods.

The first regression method was introduced in 1805 by Legendre [29] and then in 1809 by Gauss [19]. Time series methods and their use in operations research date back to the 1980s. Box's research [9] on autoregressive methods and ARIMA models have been used in time series since the 1970s and extensively discussed in Tan et al. [49]. Predictions made by time series models are not accurate enough compared to their actual values, and the problem of uncertainty in predicted values reduces the effectiveness of these models [16]. Taylor [52] proposed the Multiplicative Damped Model for exponential smoothing. Non-linear regression and time series methods have also received much attention. There are two general areas in estimating the coefficients and output of the objective function in this field:

1. Non-linear statistical methods based on stochastic characteristics to optimize model estimates.
2. Computer science methods based on structured algorithms that minimize a loss function (typically squared error).

Over the years, research in the field of time series continues, and more advanced methods have been presented that are more complex to evaluate. These methods include:

- Exponential smoothing in time series
- Bayesian ARIMA in time series

These more advanced methods were not very promising. Simpler time series smoothing methods performed better than newer, more complex methods.

Nonlinear statistical methods are also applied to time series and regression model errors. Most of the nonlinear statistical models used in this field include an explicit statistical structure in the model, while some others are defined algorithmically. Methods based on fuzzy logic are also in this category. The limitation of these methods is that:

1. These methods are primarily used in cross-sectional data (not over time) and cross-sectional classification problems such as consumer credit risk.
2. These methods have limited application in time series with contradictory results.

The third type of statistical method used in estimating coefficients of objective functions and constraints is data mining, a combination of different methods that have been successful to some extent over time. Operations research successfully uses data mining methods in the following areas:

- Selection of variables and data
- Preprocessing variables through programming and simulation

Research has proven that fundamental problems in data preprocessing and model evaluation [14] have been neglected among researchers in the operation research community.

Data mining has been developed independently of operations research or statistical forecasting [50]. Two data mining methods used in operations research are:

1. Non-linear data mining methods
2. Semi-parametric methods of data mining

Non-linear methods in data mining have caused parameter estimation problems [46]. Linear regression methods in data mining methods are unsuitable for modeling coefficients without normal distribution. When the variables that affect the model's coefficients (the response variable of a regression model) do not have a linear relationship with the dependent variable, linear regression models are not considered appropriate methods. The method of Super Vector Machines (SVM), which are based on statistical theories, can partially solve this problem and solve nonlinear classification problems by using quadratic optimization [17, 55, 56]. Non-parametric methods of support vector regression are also presented as another solution in data mining methods, which certainly do not have the accuracy of parametric methods. Classification and Regression Decision Trees algorithms which use recursive division rules are part of the designed data mining and data-driven methods. These algorithms use a lot of computing power, group methods that combine individual classification and regression methods through Boosting, Bagging, or Random Forest, and because of the significant gains in prediction accuracy, they are of interest to a large part of the community. These methods develop multiple combinations of variables and predictions based on random or weighted subsamples of the data and then combine the results using averaging (regression) or voting (classification). Although this reflects findings related to combining methods in forecasting, there is no interaction between the two fields. All the mentioned methods have shortcomings and presuppositions, such as the assumption of normal distribution in regression or different hypotheses related to time series in predicting coefficients of the target model, and restrictions should be considered. When we work with actual data, assumptions such as normal distribution or more complex assumptions assumed in time series methods are not realistic. Smith, Agrawal, and Meintyre [45] pointed out the importance of regression models and stated the many problems these models have in estimating coefficients if defaults are not established.

Howitt [24] stated that the equations used in objective models and their coefficient estimates (or even constraint functions) should have a structure that follows the reality and behavior observed in actual factory conditions. As a result, using models whose presuppositions do not necessarily match reality lead to unrealistic and erroneous estimates and, finally, the construction of functions that are far from reality. Some researchers use non-parametric methods to estimate the coefficients of the objective function so that the assumptions related to the data distribution do not become problematic. Although non-parametric methods solve the problem of rejecting data distribution defaults, they have less power than parametric methods. In addition, the non-parametric methods that consider the time factor accurately in estimating the parameters are more limited. As a result, semi-parametric methods have been introduced and preferred in this article.

2.2.3 Data analysis methods

Recent advances in optimization methods based on data analysis or data-driven techniques have made it possible to estimate the coefficients of the objective functions and constraints with the highest accuracy and the lowest error without applying probability distributions and considering the time factor.

Data-driven estimation methods use observations of random variables as the primary input to mathematical models [7]. Many real-world optimization problems can only be solved using data-driven estimation methods because there

is no objective analytical function to evaluate all available responses to a mathematical model [57]. For this reason, many researchers seek to solve this problem using data-oriented optimization methods [10].

Traditional decision-making models under uncertainty assumption consider access to complete information, which assumes that the exact values for system parameters and probability distributions for random variables are known. Such accurate knowledge is rarely possible in practice, and a strategy based on wrong inputs may be unworkable or perform poorly when implemented. Data-driven optimization methods consider the limited information in life problems well by providing such a mathematical framework and provide stable optimization based on data analysis by taking advantage of the details related to uncertainty in modeling. Optimization and operations research researchers tried to solve the existing limitations in the data collection stage or problems caused by violating model assumptions by expanding the use of data-oriented optimization methods and evolving data-driven models.

Most evolutionary optimization algorithms assume that evaluating objective and constraint functions is simple. In contrast, such objective functions may not exist in solving many real-world optimization problems. Instead, costly numerical and computational simulations or expensive experiments should be performed to evaluate the optimality. Many articles and books are available about the evolutionary methods of data-driven optimization, each of which presents a different method of optimization based on data analysis and solving the problem of the limitations of this method. For example, Sun et al. [48] propose combining deep learning methods and random forest methods to optimize and predict the performance of data-driven methods [26].

Sun et al. [48] were not the only ones who used data-driven methods of random forests to optimize estimates in mathematical models. Many others also used this method due to the advantages of random forest data-driven methods in optimization. Wang and Jin, a member of the Institute of Electrical and Electronics Engineers, are members of a group of researchers who have studied constrained multi-objective hybrid optimization problems to solve data-driven challenges using data-driven optimization methods. Solved problems can calculate model coefficients and limits of the objective function only based on a large amount of data. Random forests and neural networks were proposed as substitutes for estimating target function coefficients and constraints to solve this class of problems. Random forest models are the most effective and efficient methods available for the best estimation of coefficients of objective functions and constraints in mathematical models among optimization methods based on data analysis [57]. Biggs, Harris, and Perakis from the University of Virginia, McGill, and MIT published the results of their three-year research in data-driven methods. They proposed random forest methods with more random trees and proved their superiority by analytical methods. They also conducted case studies on real estate investment problems and showed that these models perform well against evaluation criteria and are very suitable for algorithm performance sensitivity for different random forest parameters [8]. Other researchers, including Wang, Liu, and Chan [59], emphasized random forest methods in solving problems and estimating model coefficients using optimization methods based on data-driven optimization. They showed that the power and accuracy of coefficient prediction increases by using random forest methods in optimization methods based on data analysis. Analytical methods show proof of this claim, and two series of data were used to indicate the practical efficiency of these models and their results were significantly improved. Many practical articles have used random forest methods based on data analysis in the optimization field, showing their high efficiency. Smarra et al. [44] referred to the performance of these methods and their prediction and estimation power in energy optimization and climate control. They state that the model predictive control is a model-based technique that has been widely and successfully used over the years to improve the performance of control systems. A key factor prohibiting using the Model Predictive Control-MPC model for complex systems, such as identifying a predictive model, is related to problems such as cost and time. A new idea was presented for the predictive control of coefficients based on Model Predictive Control-MPC algorithms, such as using regression and random forests using historical data.

Zheng, Fu, and Xuan [64] also emphasized the superiority of data-driven optimization methods and especially the use of random forest methods in this optimization. They state that only the data-driven method provides the ability to solve and estimate the coefficients for practical problems where there is no exact function and method to evaluate the proposed solutions. Random forests can be used as a substitute for traditional methods to estimate coefficients of objective functions and constraint functions of limited hybrid mathematical models to solve these problems.

Li, Liang, and Ma [31] pointed out the advantages of using this method in solving mathematical planning and optimization models in the stock market system using the LASSO machine learning method. Mazumder, Radchenko, and Dedieu [36] extensively investigated the use of the LASSO method in selecting final variables in mathematical planning and optimization models in operations research. Corsaro, De Simone, Marino [13], and Özmen [38] have also conducted extensive research on using LASSO in mathematical programming models.

In recent years, the GEE method has been used to solve many equations and models that occur over time.

Campanella, Serino, Crisci and Dambra [12], Lee, Choi [28], and Louis and Baesens [35] have used the GEE method to estimate organizational profit or system optimization. This method facilitates the evaluation of coefficients of mathematical programming models and allows for the consideration of time [58]. Table 1 shows the strengths and weaknesses of different methods used in the past and the combines data-driven and statistical methods.

A combination of data-driven optimization methods (random forests), LASSO method, and finally, GEE advanced statistical method is reviewed in more detail in the rest of the article.

Table 1: Comparison of different coefficient estimation methods

No.	The desired method for estimating coefficients of objective functions and constraints	Considered and evaluable items							
		Considering the time factor	Not forcing statistical assumptions	Summarizing factors in a smaller number (identifying important factors)	Considering (choosing) influential factors	Ability to estimate the nonlinear relationship	Higher power than non-parametric methods	Limiting the number of influential factors	High confidence in predicted values
1	Fuzzy method	-	-	-	-	-	+	-	-
2	Random programming	-	-	-	-	-	+	-	-
3	ARIMA and Bayesian time series	+	-	-	-	+	+	-	-
4	Nonparametric methods in time series	+	+	-	-	+	-	-	-
5	Data mining method	-	+	+	+	+	+	+	-
6	Linear regression method	-	-	-	-	-	+	-	+
7	Exponential smoothing method	-	-	-	-	+	+	-	+
8	A data-driven optimization method of random forests combined with LASSO and GEE	+	+	+	+	+	+	+	+

3 An overview of the three best data-driven, data mining, and statistical methods in estimating the coefficients of mathematical programming models

3.1 Random forest method

The random forest method helps to optimize the estimations, and it can choose the most important factors among many determining factors in the order of importance of these factors in the analysis of model coefficients. Therefore, the random forest method makes it possible to select the most critical factors in the order of priority and importance in estimating the coefficients of the objective function and constraints in the presence of many factors that affect the coefficients of the objective function and constraints.

Random forests or random decision forests provide the best combination of factors in predicting and estimating model coefficients by building many decision trees when determining and extracting the model [22]. The first random decision forest algorithm was created in 1995 by Ho using the stochastic subspace method. Hu's formula, as a method for implementing the "random discrimination" approach, uses the method provided by Kleinberg [27] to classify factors and variables and optimize estimates. This algorithm was later called "random forests" by Breiman [11] and Liaw [34]. As explained, random forests can be used to naturally rank variables' importance in a regression or classification problem. The following method is described in Breiman's original article [11].

The first step to measure the importance of a variable or factor in a data set is to run a random forest. The out-of-bag (sample) error is recorded for each dataset during the run and averaged over the forest (if the sample is not

used during model training, independent test set errors can be substituted). The meaning of model training is that the model is based on a part of the data, which is called training data and learns the performance of the model data. Then the trained model is extracted from the data and tested on data outside the bag or sample (or a part of the data that is part of the sample that did not train the model). This process continues in random forests until a model with the highest optimization is extracted to minimize the error of the tested model. The least error estimates are found by this method. The values of that factor are interpolated among the training data to measure the importance of a factor after training the model. The out-of-sample error is recalculated on this permuted data set. The mean difference in out-of-sample error calculates the significance score for this feature before and after permutation over all trees. Then, the score computed for that factor is normalized by considering the standard deviation of these differences. Features that generate large values for this score are more important than features that generate smaller values. An exception that may make it difficult to rank the importance of variables occurs for data with categorical variables with different levels or categories. In such a case, random forests may be biased in favor of the factors that have a higher number of categories and select those class variables with more classes as the most critical variable or factor [39, 47]. The simultaneous use of the selection operator method and the smallest absolute contraction or LASSO to select the influencing variables from a group of variables is recommended to solve this problem and ensure that the influential factors on the coefficients of the objective function and constraints are correctly selected. Although the random forest method also estimates the coefficients, these coefficients are biased.

3.2 LASSO method

LASSO is an analysis and prediction method used to increase the accuracy of prediction and interpretation of statistical and mathematical models, select variables, and adjust the model. This method was first introduced in geophysics by Santasa, Williams [42] and later by Tibshirani [53], who used the term LASSO.

Lasso was initially formulated for linear regression models to provide estimators for estimating coefficients and selecting variables that contain the least error. Lasso is created by combining its relationship with ridge regression, the method of Best Subset Selection, and connections between Lasso coefficient estimation and threshold smoothing. Lasso's ability to select a subset of factors depends on the shape of the constraints and has various interpretations, including in geometry, Bayesian statistics, and convex analysis.

Before Lasso, the most widely used method for selecting variables was the stepwise selection method. This approach improves prediction accuracy only in exceptional cases, for example, when only a few predictor variables have a strong relationship with the response variable. In other cases, the method of selecting variables (factors affecting model coefficients) step by step can increase the prediction error.

An alternative to this variable selection method, a model containing all p predictors, can be fitted using the restricted or regularized coefficient estimation technique.

LASSO reduces the estimate of non-significant coefficients to zero. Such a restriction improves the fit of a model, but shrinking the coefficient estimates can significantly reduce their variance. LASSO is one of the most well-known techniques for shrinking regression coefficients to zero.

According to Zou et al. [66], LASSO achieves both goals by forcing the sum of the absolute values of the regression coefficients to be smaller than a fixed numerical value. LASSO forces some small model coefficients to shrink to zero for their effective elimination. LASSO is similar to ridge regression, but the shrinkage size is larger, and unlike ridge regression, which did not make insignificant factors zero, LASSO makes them zero. Jiang [25] found that the LASSO method can be implemented before the optimization and estimation of coefficients to select essential factors that can be entered into a statistical model as determining factors. As with the random forest method, the estimates estimated by the LASSO method also include bias. Therefore, when unbiased estimates are essential, statistical methods should be used for estimation.

When the ordinary regression method estimates the coefficients in the presence of the time factor, time changes are not accounted for. When time series is used, the model does not include many influential factors. In the proposed method, semi-parametric methods are used to estimate the coefficients, which have a higher power than non-parametric methods and do not force the probability distribution assumptions that parametric regression or time series methods have [23]. Higher power refers to the overall higher ability of studies based on historical (longitudinal) data that provide more observations and data than non-longitudinal data [62]. As Liang, Zeiger, and Qaqish [33] have shown, more stable and reliable models can be built through longitudinal model because some incorrect assumptions when defining of the model algorithm can be avoided in longitudinal studies. This advantage is because data analysis and longitudinal models are not sensitive to ignoring factors that change over time. For this reason, the semi-parametric GEE method, which can consider historical (longitudinal) data in estimating the coefficients of mathematical models,

has higher reliability in estimating model parameters than non-parametric methods [21]. The superiority of parametric and semi-parametric methods has been discussed in Sheskin [43], Hardle and Mammen [21], and Zeiger and Liang [62].

The proposed statistical modeling and parametric estimation method is the GEE method, which does not have the shortcomings of linear regression models, time series, or non-parametric methods and predicts the coefficients with high confidence. Thus, various factors with the source of changes and uncertainty in the estimation of coefficients are entered into the model in each period and over time, and coefficients with the highest confidence are estimated. The estimates with the highest accuracy which considered the maximum uncertainties and dependencies between data and variables increase the quality of the objective function and constraints and give us a more accurate estimate for the coefficients. The details of the GEE method are explained below.

3.3 Generalized estimated equation (GEE)

GEE belongs to a class of advanced and generalized regression techniques called a semi-parametric estimation and optimization method. This estimation method relies only on specifying the mean and variance of the data and does not require more specifications for modeling [32]. Suppose there are data in the form of repeated measures of the response variable and the correlated variables in a group of subjects over time. A suitable model is created for individual observations and correlated variables to estimate the mean of the response variable (here are the coefficients of the objective functions or constraints). The meaning of repeated measures is using the same subject several times in the experiment, which was measured in different conditions at different places or times. This mode has created a series of data called Panel Data. GEE is one of the techniques for examining such data using "generalized estimating equations" [20]. GEE is commonly used in extensive studies, especially multivariable studies, because it can control many types of indirect dependence between factors and outcomes. This method allows for combining the advantages of each of the statistical techniques that have been used so far in operations research.

Zeger, Liang, and Diggle [18] explained that correlated data are modeled using a link function and linear predictor adjustment (systematic component) in the GEE method. The random component is described by the same variance function described in the case of data independence, but the covariance of correlated responses in the presence of correlated data should also be specified and modeled in GEE models. The GEE method provides more accurate estimates than regression without the limitations of distributional assumptions and the need for traditional (transformation) of parameters. In addition, the GEE method considers time changes and different relationships between data from other locations and factories, or growth or decline or changes in relationships between data over time. This model can capture unpredictable changes in the supply chain, production, distribution, or any other chain over time by adding different random variables. As explained earlier, GEE is a longitudinal modeling method over time. GEE methods can estimate the parameters of both linear and non-linear models over time. As mentioned, the parameter estimates in GEE still provide estimates close to reality even when the correlation structure between the data is not known or is incorrectly specified.

4 Conclusion

This study reviewed data-driven optimization, data mining, machine learning and statistical methods to estimate coefficients of objective functions and restrictions in the presence or absence of the time factor in mathematical programming models. Semi-parametric statistical methods are suggested for extracting unbiased estimates because it greatly benefits from the accuracy of estimating the predicted values of parametric statistical methods. Further, semi-parametric statistical methods do not have the inaccuracy of the estimated values of non-parametric methods and solve the major problem of most time series and regression methods by not depending on statistical distribution assumptions (which is the main problem of parametric methods). On the other hand, this approach allows us to estimate nonlinear relationships without the need for smoothing methods, which are known to negatively affect the estimation accuracy. Using the GEE method to consider the linear or non-linear relationships between the variables allows for viewing the time factor and the change of each of the variables and location (or factory) conditions over time and the effect of the coefficients of the objective function. Statistical methods alone do not allow us to consider many variables and factors in the model.

Considering a large number of influential factors theoretically or according to the opinion of experts or the owners of candidate industries, the random forest data-driven optimization method and the LASSO data mining method can evaluate hundreds of influential factors, but they give biased estimates. This is why their results are next entered into a GEE model for unbiased estimates.

Therefore, operational research science needs a model that does not force the use of a small number of factors in the estimation of coefficients and also gives unbiased estimates of coefficients.

References

[1] R. Alikhani, A. Azar, and A. Rashidi Kamijan, *Stochastic planning model for allocation of gas resources in Iran with energy security cost approach*, Future Stud. Manag. **23** (2012), no. 96, 25–36.

[2] R. Alikhani and M. Sadegh Amal Nik, *The fuzzy-stochastic multi-objective programming model for supplier selection problem*, Standard Qual. Manag. **4** (2013), no. 12, 96–101.

[3] M. Amiri and S.A. Ayazi, *Decision Making in Conditions of Uncertainty*, Allameh Tabatabayi University, 2017.

[4] M. Amiri, A. Darestani Farahani, and M. Mehboob Ghodsi, *Multi-Criteria Decision Making*, Kian University Press, 2015.

[5] A. Azar, R. Farhi Bailoyi, and A. Rajabzadeh, *Comparative comparison of deterministic and fuzzy mathematical models in production planning (Case: Shiraz Oil Refining Company)*, Modares Human Sci. J. **12** (2017), no. 1.

[6] A. Azar and M. Momeni, *Statistics and its Application in Management*, In Persian, Samt publication, 2006.

[7] D. Bertsimas and A. Thiele, *Robust and data-driven optimization: modern decision making under uncertainty*, Models, methods, and applications for innovative decision making, INFORMS, 2006, pp. 95–122.

[8] M. Biggs, R. Hariss, and G. Perakis, *Optimizing objective functions determined from random forests*, Available at SSRN: https://ssrn.com/abstract=2986630, (2018).

[9] G.E. Box and D.A. Pierce, *Distribution of residual autocorrelations in autoregressive-integrated moving average time series models*, J. Amer. Stat. Assoc. **65** (1970), no. 332, 1509–1526.

[10] S.P. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004.

[11] L. Breiman, *Random forests*, Machine Learn. **45** (2001), no. 1, 5–32.

[12] F. Campanella, L. Serino, A. Crisci, and A. D'Ambra, *The role of corporate governance in environmental policy disclosure and sustainable development. Generalized estimating equations in longitudinal count data analysis*, Corporate Soc. Responsibility Environ. Manag. **28** (2021), no. 1, 474–484.

[13] S. Corsaro, V. De Simone and Z. Marino, *Fused lasso approach in portfolio selection*, Ann. Oper. Res. **299** (2021), no. 1–2, 47–59.

[14] S.F. Crone, S. Lessmann and R. Stahlbock, *The impact of preprocessing on data mining: An evaluation of classifier sensitivity in direct marketing*, Eur. J. Operat. Res. **173** (2006), no. 3, 781–800.

[15] G.B. Dantzig, *Linear programming under uncertainty*, Manag. Sci. **1** (1955), no. 3–4, 197–206.

[16] J.G. De Gooijer and R.J. Hyndman, *25 years of time series forecasting*, Int. J. Forecast. **22** (2006), no. 3, 443–473.

[17] N. Deng, Y. Tian and C. Zhang, *Support Vector Machines: Optimization Based Theory, Algorithms, and Extensions*, CRC Press, 1979.

[18] P. Diggle, K.Y. Liang and S.L. Zeger, *Longitudinal Data Analysis*, New York: Oxford University Press, 1994.

[19] C.F. Gauss, *Theoria Motus Corporum Coelestum*, Werke, 1809.

[20] J.W. Hardin and J.M. Hilbe, *Generalized Estimating Equations*, CRC Press, 2012.

[21] W. Hardle and E. Mammen, *Comparing nonparametric versus parametric regression fits*, Ann. Statist. **21** (1993), no. 4, 1926–1947.

[22] T.K. Ho, *The random subspace method for constructing decision forests*, IEEE Trans. Pattern Anal. Machine Intell. **20** (1998), no. 8, 832–844.

[23] M. Hollander, D.A. Wolfe, and E. Chicken, *Nonparametric Statistical Methods*, John Wiley & Sons, 2013.

[24] P. Howitt and D. Mayer-Foulkes, *R&D, implementation and stagnation: A Schumpeterian theory of convergence clubs*, Brown University, 2002.

[25] Y. Jiang, *Variable selection with prior information for generalized linear models via the prior lasso method*, J. Amer. Stat. Assoc. **111** (2016), no. 513, 355–376.

[26] Y. Jin, H. Wang, T. Chugh, D. Guo and K. Miettinen, *Data-driven evolutionary optimization: An overview and*

case studies, IEEE Trans. Evol. Comput. **23** (2018), no. 3, 442–458.

[27] E.M. Kleinberg, *On the algorithmic implementation of stochastic discrimination*, IEEE Trans. Pattern Anal. Machine Intell. **22** (2000), no. 5, 473–490.

[28] J. Lee and J.Y. Choi, *Texas hospitals with higher health information technology expenditures have higher revenue: Longitudinal data analysis using a generalized estimating equation model*, BMC Health Serv. Res. **16** (2016), no. 1, 1–8.

[29] A.M. Legendre, *Mémoire Sur Les Opérations Trigonométriques: Dont les Résultats Dépendent de la Figure de la Terre*, F. Didot, 1805.

[30] J. Lever, M. Krzywinski, and N. Altman, *Points of significance: model selection and overfitting*, Nature Meth. **13** (2016), no. 9, 703–705.

[31] X. Li, C. Liang and F. Ma, *Forecasting stock market volatility with many predictors: New evidence from the MS-MIDAS-LASSO model*, Ann. Oper. Res. (2022). https://doi.org/10.1007/s10479-022-04716-1

[32] K.Y. Liang and S.L. Zeger, *Longitudinal data analysis using generalized linear models*, Biometrika **73** (1986), no. 1, 13–22.

[33] K.Y. Liang, S.L. Zeger and B. Qaqish, *Multivariate regression analyses for categorical data*, J. Royal Stat. Soc. Ser. B (Method.) **54** (1992), no. 1, 3–24.

[34] A. Liaw and M. Wiener, *Documentation for R package randomForest*, https://www.rdocumentation.org/packages/randomForest/versions/4.6-12, 2013.

[35] P. Louis and B. Baesens, *Do for-profit microfinance institutions achieve better financial efficiency and social impact? A generalized estimating equations panel data approach*, J. Dev. Effect. **5** (2013), no. 3, 359–380.

[36] R. Mazumder, P. Radchenko and A. Dedieu, *Subset selection with shrinkage: Sparse linear modeling when the SNR is low*, Oper. Res. **71** (2022), no. 1, 129–147.

[37] S.H. Naseri and S. Bavandi, *A proposed approach for solving multi-objective fuzzy stochastic linear programming problems with fuzzy probability*, Fuzzy Syst. Appl. **1** (2018), no. 2, 133–119.

[38] A. Özmen, *Sparse regression modeling for short-and long-term natural gas demand prediction*, Ann. Operat. Res. 322 (2021), no. 2, 1–26.

[39] A. Painsky and S. Rosset, *Cross-validated variable selection in tree-based methods improves predictive performance*, IEEE Trans. Pattern Anal. Machine Intell. **39** (2017), no. 11, 2142–2153.

[40] H. Rasouli, M. Imanipour and A. Khatami Firouzabadi, *A comprehensive guide to linear programming modeling*, Marandiz, Todays Managers, 2017.

[41] J.W. Rocks and P. Mehta, *Memorizing without overfitting: Bias, variance, and interpolation in overparameterized models*, Phys. Rev. Res. **4** (2022), no. 1, 013201.

[42] F. Santosa and W.W. Symes, *Linear inversion of band-limited reflection seismograms*, SIAM J. Sci. Stat. Comput. **7** (1986), no. 4, 1307–1330.

[43] D.J. Sheskin, *Handbook of Parametric and Nonparametric Statistical Procedures*, Chapman and Hall/CRC, 2003.

[44] F. Smarra, A. Jain, T. De Rubeis, D. Ambrosini, A. D'Innocenzo and R. Mangharam, *Data-driven model predictive control using random forests for building energy optimization and climate control*, Appl. Energy **226** (2018), 1252–1272.

[45] S.A. Smith, N. Agrawal and S.H. McIntyre, *A discrete optimization model for seasonal merchandise planning*, J. Retail. **74** (1998), no. 2, 193–221.

[46] K.A. Smith and J.N. Gupta, *Neural networks in business: techniques and applications for the operations researcher*, Comput. Oper. Res. **27** (2000), no. 11–12, 1023–1044.

[47] C. Strobl, A. Boulesteix and T. Augustin, *Unbiased split selection for classification trees based on the Gini index*, Comput. Stat. Data Anal. **52** (2007), 483–501.

[48] Y. Sun, H. Wang, B. Xue, Y. Jin, G.G. Yen and M. Zhang, *Surrogate-assisted evolutionary deep learning using an*

end-to-end random forest-based performance predictor, IEEE Trans. Evolut. Comput. **24** (2019), no. 2, 350–364.

[49] C.W. Tan, C. Bergmeir, F. Petitjean and G.I. Webb, *Time series extrinsic regression: Predicting numeric values from time series data*, Data Min. Knowledge Discov. **35** (1994), 1032–1060.

[50] P.N. Tan, M. Steinbach and V. Kumar, *Introduction to Data Mining*, Pearson Education India, 2016.

[51] H. Tanaka, T. Okuda and K. Asai, *Fuzzy mathematical programming*, Trans. Soc. Instrument control Engin. **9** (1973), no. 5, 607–613.

[52] J.W. Taylor, *Short-term electricity demand forecasting using double seasonal exponential smoothing*, J. Oper. Res. Soc. **54** (2003), no. 8, 799–805.

[53] R. Tibshirani, *Regression Shrinkage and Selection via the Lasso*, J. Royal Stat. Soc. Ser. B (Method.) **58** (1996), no. 1, 267–288.

[54] M.D. Troutt, W.-K. Pang and S.-H. Hou, *Behavioral estimation of mathematical programming objective function coefficients*, Manag. Sci. **52** (2006), no. 3, 422–434.

[55] V. Vapnik and O. Chapelle, *Bounds on error expectation for support vector machines*, Neural Comput. **12** (2000), no. 9, 2013–2036.

[56] V.N. Vapnik and A.Y. Chervonenkis, *Recovery of dependencies by empirical data*, M.: Nauka, 1979.

[57] H. Wang and Y. Jin, *A random forest-assisted evolutionary algorithm for data-driven constrained multiobjective combinatorial optimization of trauma systems*, IEEE Trans. Cybernet. **50** (2018), no. 2, 536–549.

[58] M. Wang, L. Kong, Z. Li, and L. Zhang, *Covariance estimators for generalized estimating equations (GEE) in longitudinal analysis with small samples*, Stat. Med. **35** (2016), no. 10, 1706–1721.

[59] W. Wang, X. Liu, and W.K.V. Chan, *Imbalanced classification problem using data-driven and random forest method*, Proc. 3rd Int. Conf. Data Sci. Inf. Technol., 2020, pp. 26–30.

[60] W.L. Winston, *Operations Research: Applications and Algorithms*, Cengage Learning, 1997.

[61] L.A. Zadeh, *Information and control*, Fuzzy Sets Syst. **8** (1965), 338–353.

[62] S.L. Zeger and K.Y. Liang, *Feedback models for discrete and continuous time series*, Stat. Sin. **1** (1991), 51–64.

[63] Z. Zhang, *Too many covariates in a multivariable model may cause the problem of overfitting*, J. Thoracic Disease **6** (2014), no. 9.

[64] Y. Zheng, X. Fu and Y. Xuan, *Data-driven optimization based on random forest surrogate*, 6th Int. Conf. Syst. Inf., IEEE, 2019, pp. 487–491.

[65] H.-J. Zimmermann, *Fuzzy programming and linear programming with several objective functions*, Fuzzy Sets Syst. **1** (1978), no. 1, 45–55.

[66] H. Zou, T. Hastie, and R. Tibshirani, *On the "degrees of freedom" of the lasso*, Annal. Statist. **35** (2007), no. 5, 2173–2192.