ISSN 1392-3730/eISSN 1822-3605

VILNIUS TECH
Vilnius Gediminas Technical University

JOURNAL of CIVIL ENGINEERING and MANAGEMENT

2025
Volume 30
Issue 1
Pages 76–89
https://doi.org/10.3846/jcem.2024.22472

A REGRESSION-BASED MODEL FOR PARAMETRIC COST ESTIMATION OF INDUSTRIAL STEEL STRUCTURES

Adel ALSHIBANI[1,2], Osama ALMUHTASEB[1], Awsan MOHAMMED[1,3]✉, Ahmed M. GHAITHAN[1,3]

[1]*Architectural Engineering and Construction Management Department, King Fahd University of Petroleum and Minerals, Dhahran, Saudi Arabia*
[2]*Interdisciplinary Research Center of Construction and Building Materials, King Fahd University of Petroleum and Minerals, Dhahran, Saudi Arabia*
[3]*Interdisciplinary Research Center of Smart Mobility and Logistics, King Fahd University of Petroleum and Minerals, Dhahran, Saudi Arabia*

Article History:
- received 12 August 2023
- accepted 19 August 2024
- first published online 10 December 2024

Abstract. Construction industry is considered one of the most versatile industries characterized by uncertainties and risk. Estimating the steel structure cost of industrial buildings is a challenging task compared with traditional buildings due to the uniqueness of this class of projects. This paper aims to introduce an effective and accurate parametric model for construction cost estimation of industrial steel structures. The paper proposes a regression-based model for estimating the cost of a critical construction component: the industrial steel structure where the is not enough historical data is available. The factors that affect the construction cost of industrial steel structures are initially identified based on the literature and interviews with local experts. The correlation between input factors and model's output is then investigated. In addition, sensitivity analysis is performed to examine the relative importance of the regression model's inputs. The model is validated using actual data on industrial steel structure costs in Saudi Arabia. The model adequately predicted the construction costs of actual projects with an accuracy of more than 88%. This indicates that the model is capable of accurately predicting the cost of such structures. The proposed model can be of great assistance to investors and decision-makers looking to invest in the industrial sector.

Keywords: construction, industrial steel structures, parametric cost estimation, multiple linear regression.

✉Corresponding author. E-mail: *awsan.mohammed@kfupm.edu.sa*

1. Introduction

Construction industry is one of the most versatile industries characterized by uncertainties and risk. At the early stage, before the design stage starts, estimating construction cost to the owner is a crucial task as the required information is little or is not available at all. The steel structure is one of the major categories of industry construction buildings. The most common types of steel constructions are bridges, towers, high-rise buildings, and industrial steel structures. The cost estimation of these steel structures is a critical task that occurs at the early stage of project life and it has a significant impact on its success. Several techniques are used to estimate project costs according to the required level of accuracy. The main methods include conceptual cost estimation, parametric cost estimation, square-meter-of-floor-area, and detailed cost estimation methods. In detailed cost estimation, the scope of the work is usually clear and the project documents are complete, therefore, the estimated cost is usually accurate. However, this process is usually considered time-consuming and costly which is not always done by owners at the early stage of the project life. On the other hand, square-meter-of-floor-area estimates the cost of a steel structure using its area only. This technique is fast but, is also considered relatively inaccurate compared to other cost estimation approaches.

Parametric cost estimation is a fast technique and is performed at the early stages of a project's life, when less information about the project's design is available, at almost no cost. Parametric cost estimation uses certain parameters that have the greatest effect on the cost of a product to estimate its cost. The main use of the parametric cost estimation is to conduct economic/feasibility

analysis. After identifying these parameters which are called cost drivers, parametric cost estimation models can then be developed to estimate product's cost (Siqueira, 1999). These models provide quick estimates, but they are typically less accurate than detailed cost estimates. According to the Construction Industry Institute, or CII, and the Association for Advancement of Cost Engineering, or AACE, the error range for detailed cost estimates is usually around 10%, whereas it is between 30%–50% for parametric estimates. However, the margin of error of parametric cost estimation is significantly reduced when it is well-prepared. Using well-prepared parametric cost estimation models in the construction industry might save the entities working within the industry a significant amount of time and resources whenever a cost estimate of a project is to be prepared (U.S. Department of Defense, 1995; Siqueira, 1999).

The literature reveals that although several techniques such as square-floor-area are widely used in estimating the construction cost of the building project at an early stage, those techniques cannot provide a reasonable estimate of steel structure. The previous studies also indicated the capability of regression models in predicting the construction especially where is not enough data from already completed projects. It should be noted that if there is enough data available, neural networks can be better options. In addition, the majority of the sample size in construction is small to medium. This makes the regression models the most suitable techniques for prediction especially when the relationship between the factors and the output is linear. This phenomenon exists in most cases in construction. Furthermore, regression models enable the assessment of the importance of different predictors, aiding in variable selection and prioritization of factors influencing the outcome. Moreover, regression models assist in identifying and selecting relevant variables. This is crucial for understanding which factors contribute significantly to the variation in the dependent variable (James et al., 2013). Regression models, when properly developed and validated, can provide accurate predictions. They can be fine-tuned to fit the data well, and various diagnostic tools help in assessing the model's predictive performance. It can greatly improve the model's accuracy (Hastie et al., 2009; Chopra et al., 2014). The multiple linear regression models also offer a clear interpretation of the relationship between the dependent and independent variables through the coefficients in the model, aiding in the understanding of the underlying dynamics (Gujarati, 2003). Consequently, the objective of this paper is to develop a data-driven model to estimate the cost of steel structure. The developed model is regression-based model which is very well suitable as there is no enough data to develop parametric cost estimate. A model based on multiple linear regression is proposed to forecast the construction costs of industrial steel structures. Furthermore, the study seeks to determine the variables that contribute to these costs. The model is developed, examined, and verified to ensure its accuracy and efficiency. It also provides correlations between input variables and between inputs and outputs. Furthermore, sensitivity analysis is carried out to determine how changing the input variables affects the model's performance.

The paper is organized as follows: Section 2 discusses the literature review, followed by the proposed model in Section 3. Section 4 discusses the findings, and Section 5 concludes the findings of the study.

2. Literature review

Compared with building construction, the literature indicated that a limited number of parametric cost estimate models of steel structure were reported. Hegazy and Ayed (1998) developed a model based on neural networks to perform parametric cost estimation for highway projects in Canada. Data from 18 highway projects were used in the development of the model. The model also considered 10 input parameters which are the project type, the project scope, the year of construction, the construction season, the location of the project, the duration of the project, the size of the project, the capacity of the project, the existence of a water body, and the soil condition.

Moselhi and Siqueira (1998) developed a model based on a neural network to predict the construction cost of low-rise structural steel buildings in Canada. The authors initially identified the factors that affect the construction cost of steel structures. El-Sawah and Moselhi (2014) presented a comparative study using regression analysis and neural networks to perform order of magnitude cost estimation of low-rise structural steel buildings and timber bridges. The models developed for timber bridges considered three input parameters which are "the web volume", "the deck volume" and "the weight of the steel used in the bridge". As for the low-rise steel structures models, four parameters were used in the models' development, namely, "the area of the structure", "the perimeter of the structure", "the height of the structure" and "the Joist span". The results showed the capability of both approaches for cost estimation.

Günaydın and Doğan (2004) presented a cost estimation model based on neural networks for structural systems of buildings. The model uses 30 data sets from different projects to train and test the model. Kim et al. (2004) presented three models to estimate the cost of building construction; case-based reasoning, neural networks, and multiple regression analysis. The parameters used to develop the models are the finishing quality, usage of basement, the foundation type, the roof type, the construction duration, the total unit, the number of stories, and the gross floor area. Sonmez (2004) presented a conceptual cost estimation model based on neural networks to estimate the cost of building construction and compared it to common regression for the closeness of fit and prediction performance. Data from 30 building projects in the USA were used in the models' development. The parameters

used in the development of the neural network models include "area of the building", "time of construction" and "the location of the building". Both techniques showed satisfactory results.

Lowe et al. (2006) presented several parametric models to estimate the construction cost of buildings in the United Kingdom. The dependent variables of the models were cost, log of cost, and log of cost/m^2. The most significant factors affecting the construction cost were found to be the piling, mechanical installation, duration, function, and gross internal floor area.

Petroutsatou et al. (2006) developed early cost estimation models for tunnel construction projects in Greece. Two steps of multiple regression analysis were used in models' development to first correlate the geotechnical properties with the structural properties and then, depending on those, estimate the construction costs. Data from 33 tunnels bored in Greece was utilized in the study. The accuracy of the developed cost estimation model was reported to be 88.46%, indicating very good performance. Aibinu and Pasco (2008) used multiple regression analysis to investigate the accuracies of pre-tender estimates for construction projects in Australia. Results of the study indicate that small-sized projects tend to be overestimated more than large-sized projects. Sonmez and Ontepeli (2009) developed parametric cost estimation models for light rail projects in Turkey using neural networks and regression analysis. Data from 13 projects executed over the period between 1986–2005 were collected and utilized. the findings showed that the regression model is reliable in estimating the cost of railway construction. Wang et al. (2010) developed a back-propagation neural networks-based model to estimate the construction cost of highway projects in China. Data from 28 projects were collected for nine input parameters which were considered in the development of the model. The results of the study showed a very small relative error that reached 5% in some cases.

Mahamid and Bruland (2010) presented several models to estimate the cost of road construction activities, i.e., earthworks, base course work, and asphalt work. Mahamid (2011) used multiple regression analysis to develop an early parametric cost estimation model for road construction in Palestine. The best-performing model resulted in a MAPE of 13%, indicating the high prediction capability of the model. Gunduz et al. (2011) used neural networks and multiple regression analysis to build parametric cost estimation of light rail transit and metro in Turkey that can be used in the early stages of projects' life cycles. Data from 16 projects in Turkey was used in the model development.

Arafah and Alqedra (2011) developed a model based on neural networks to estimate the construction cost of building projects at early stages in the Gaza strip. The developed model considered 7 parameters that can be obtained at the pre-design stage. The considered parameters are number of rooms, number of columns, typical floor area, number of elevators in the building, types of foundation, number of stories, and ground floor area. Latief et al. (2013) used regression analysis, Fuzzy logic, and neural network to develop a parametric cost estimation model for low-cost apartment projects in Indonesia. Regression analysis was used for the determination of key cost drivers to be used as inputs to the developed model. Four key cost drivers were identified as input parameters for the model. The developed model was compared to neural-based and regression-based models, and the MAPE were found to be 3.98%, 10.12%, and 6.92%, respectively, indicating that the proposed model outperforms the other two.

Kim et al. (2013) developed three models to estimate the construction cost of school buildings at early stages of the project's life. The models are neural-based, regression-based and SVM-based (Support Vector Machine). The MAPEs of the neural-based model, the regression model and the SVM-based model are 5.27%, 5.68% and 7.48%, respectively. The findings showed that both the neural-based and the regression-based models are very accurate but with the former being slightly superior to the others. Cho et al. (2013) compared a regression-based model to a neural-based model to determine which model is better for estimating the construction cost of elementary school buildings in Korea. Results demonstrated that neural-based model outperformed the regression-based model in terms of accuracy. Mahamid (2013) developed regression-based models to estimate the construction cost of road construction projects in Saudi Arabia. Data from 52 projects were utilized in the models' development. The best-performing model had a MAPE of 17%.

Roxas and Ongpeng (2014) used data from 30 projects in the Philippines and neural networks in the development of a model to estimate the construction cost of building projects. The developed model considered six parameters affecting the cost of building construction. The model showed a correlation coefficient corresponding to the testing data set of 0.95, which was considered satisfactory. However, there is no consideration in the study for the application of a sensitivity analysis on the parameters to determine the most impacting ones on the estimated cost. Shin (2015) applied Boosting Regression Tree (BRT) in the development of a parametric model to estimate the building construction cost. The performance of the developed model was compared to that of a similar neural-based model that was high in prediction capability. The developed model showed similar results to those of the neural-based model in terms of accuracy.

Ofori-Boadu (2015) used multiple regression analysis to develop parametric cost estimation models for high-rise buildings. The study initially considered 12 parameters in the models' development, however, only five were found to be significant, i.e., "the gross floor area of the building", "the location of the building", "height of the building", "the structural material used" and "the completion date". The results of this study indicated that the error rate of the best performing model was 9%, indicating adequate performance. Fragkakis et al. (2015) used regression analysis to build a cost estimation model that would gener-

ate predesign estimates for culverts along a motorway in Greece. The developed model shows a MAPE of less than 20% when it comes to the determination of concrete and reinforcement steel required, which was considered satisfactory for the purpose of the study. Dang and Le-Hoai (2018) developed a regression-based model for residential buildings cost estimation in Vietnam. The developed models used Storey Enclosure Method (SEM) to identify the parameters needed to build a parametric cost estimation model. Range estimation was also presented using nonparametric bootstrap method. The model with the best performance was found to have a MAPE of 6.63%, which indicates very good performance.

Alshamrani (2017) used multiple regression analysis to construct a parametric model for estimating the construction costs of college buildings in North America. To estimate the cost of construction, the model used four parameters: the area of the building, the number of floors, the floor height, and the type of construction material (concrete or steel). The findings showed that the developed model has a MAPE of 5.7%, demonstrating the high level of accuracy the developed model enjoys. Dharwadkar and Arage (2018) developed and compared the performance of a regression-based model and a neural-based model which estimate building construction cost in India. Results show that both models perform well and that the difference between them in terms of accuracy is insignificant. Badawy (2020) developed a model based on neural networks and multiple linear regression to estimate the construction cost of a residential building in Egypt. Results showed that the developed model has a MAPE of 10.64% and that the parameters most affecting the cost are the floor area and a number of floors.

Badra et al. (2020) developed several models to provide a conceptual cost estimation of building projects in Egypt. The models considered seven parameters on which data can be easily found in the early stages of a project, i.e., type of slab, floor area, number of floors, type of external finishing, type of internal finishing, number of elevators and type of electro-mechanical work. Xue et al. (2020) presented a cost estimation model for expressway projects. The model uses convolutional neural network algorithm in its development. Chro (2021) used multiple regression analysis to develop parametric models to estimate the construction cost of roads and tunnels. The developed models incorporated the tunnels' lengths, diameters, and the tunneling method used, i.e., mechanized or conventional. Data from 25 projects located in western Europe was used to develop the models. The results of the study demonstrate that the accuracy of the developed models is above 75%. Sanni-Anibire et al. (2021) developed several models to provide an early estimate of the construction cost of tall buildings using different techniques. These techniques are multi classifier system, support vector machine, neural networks, K-nearest neighbors, and multiple regression analyses. Results showed that a hybrid model developed using (KNN) and (MCS) outperformed the other models.

The existing body of literature highlights the prevalent reliance on established techniques, such as square-floor-area, for estimating construction costs in the initial phases of building projects. However, these techniques cannot provide an accurate estimation of steel structure. Furthermore, the literature demonstrates that no study used multiple linear regression to predict the cost of industrial steel structures in the construction industry. Furthermore, previous studies revealed a lack of studies that investigate the relationship between the cost estimation of the industrial steel structure and the factors influencing the cost estimation. Consequently, this paper provides a multiple regression model for forecasting the costs associated with industrial steel structures. It examines both existing research and practical examples to determine and select essential variables influencing the estimation of steel structure costs. It also establishes correlations between input variables and input factors, as well as the resulting output. Furthermore, sensitivity analysis is used to investigate how changing the model's inputs affects the results.

3. The proposed model

Estimating the cost of industrial steel structures is a difficult task for the decision-making team. Challenges occur when there are unidentified factors that have a major effect on cost estimation, such as the insulation material used or the type of structure itself. The difficulty lies in identifying appropriate patterns that link cost estimation for industrial steel structures to input factors, which in turn affects the accuracy of these costs. In this regard, regression-based model is developed to tackle such issues. A multiple regression model for estimating the cost of industrial steel structures is proposed in this study. Minitab software is utilized to develop the estimation model. After training and testing several models, the best regression model was chosen. Data for the model were gathered from various industrial steel structures throughout Saudi Arabia. Initially, the proposed model was developed by identifying factors that influence the cost estimation of these structures. Correlations between explanatory variables were then examined, resulting in a more complete model explanation and improved analytical abilities. The most significant parameters were then selected using a stepwise regression approach. Stepwise variable selection is a systematic approach aimed at identifying the most relevant independent variables to include in the regression model. This method iterates through a series of steps to either add or remove variables based on their contribution to the model's explanatory power. Typically, it begins with an empty model and proceeds with either forward selection, backward elimination, or a combination of both. Forward selection starts by including one variable at a time, assessing its significance using a predetermined criterion such as p-values or information criteria like adjusted R-squared. Variables that meet the significance threshold are retained, and the process continues iteratively by adding the next most significant variable until no further improvement in model fit is observed.

Conversely, backward elimination begins with all potential independent variables included in the model. At each step, the least significant variable is removed, and the model's fit is evaluated. This process continues until removing additional variables no longer improves the model's performance based on the chosen criterion. Stepwise regression alternates between forward selection and backward elimination until reaching a stopping criterion, which may include factors such as no further variables meeting the significance threshold, a predefined maximum number of variables, or diminishing returns in model improvement. Throughout this iterative process, variables are continuously assessed and compared based on their impact on the model's goodness of fit and their statistical significance. Finally, actual external points were used to develop and validate the developed model. Figure 1 illustrates the steps of model development. These steps are explained in subsequent sections in detail.

3.1. Factor identification

The identification of the factors affecting the cost estimate is the first step in this study. The factors are identified from two main sources: the literature and experts' interviews. Eight factors are identified from the literature. A number of local experts from the Saudi Industrial Development Fund (SIDF) acknowledged 7 of the 8 factors and eliminated one, i.e., The type of foundation system, on the basis that the type of foundation system used in the structures considered by the study is the same, and that other types of foundation systems are rarely used. Furthermore, the experts added 3 other factors. Table 1 illustrates the factors identified.

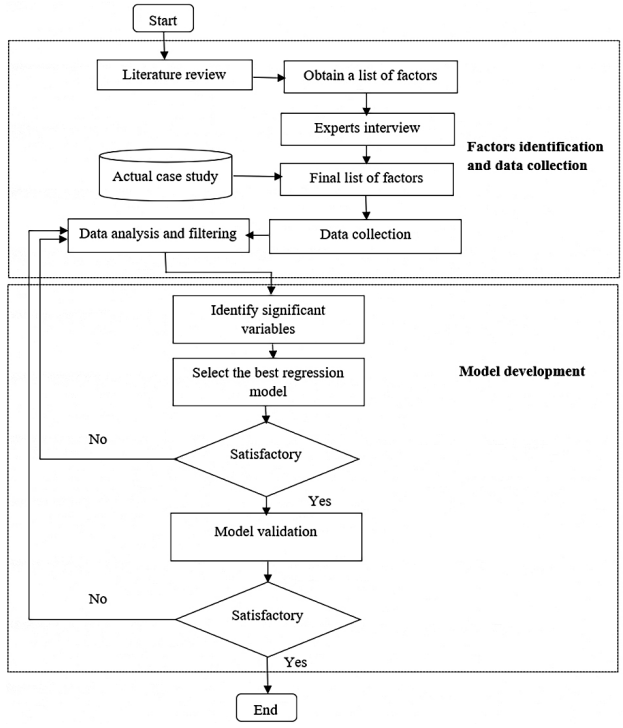

Figure 1. Flowchart of research methodology

3.2. Data collection

The first step of the model development phase was data collection. Data from real construction projects were collected with the help of SIDF. Data collection targeted several factors influencing the construction cost of industrial steel structures in Saudi Arabia which were incorporated

Table 1. Identified factors affecting the cost estimation of industrial steel structures

Identified factor	Description	Source of identification
Area of the structure	The total ground area enclosed by the walls of the structure (including the walls) in square meter.	Moselhi and Siqueira (1998), Kim et al. (2004), Sonmez (2004), Arafah and Alqedra (2011), Roxas and Ongpeng (2014)
Eave height of the structure	The height from the base plate to the point where the sidewall and roof intersect measured in meters.	Kamarthi et al. (1992), Moselhi and Siqueira (1998), El-Sawah and Moselhi (2014)
Joist span	Refers to the longest distance, in meters, covered by the Joist between to supporting members such as beams, columns, etc.	Moselhi and Siqueira (1998), El-Sawah and Moselhi (2014)
Vertical load on the structure	The maximum load value acting downwards on the structure in Kilonewtons (KN).	Moselhi and Siqueira (1998)
The existence of mezzanine in the structure	Whether the structure contains a mezzanine or not.	Added by Experts
Type of insulation material	The type of insulation material used in the construction.	Added by Experts
Type of the structure	A warehouse, A shed, A shop, etc.	Hegazy and Ayed (1998)
Type of steel used in the construction	Hot rolled steel or pre-fabricated steel.	Added by Experts
Date of construction	The year at which the construction of the structure started.	Hegazy and Ayed (1998), Lu et al. (2000), Sonmez (2004)
Location of the structure	Exact location in KSA	Hegazy and Ayed (1998), Lu et al. (2000), Sonmez (2004), Alshibani and Alshamrani (2017)

in the proposed model as input factors. Data collection also targeted the projects' actual construction costs, the parameter which was incorporated as the output of the developed model. Data were then sorted and filtered. Incomplete data sets were removed since they can cause disturbance to the model. Actual data from a total of 31 projects were left to undergo the Grubb's test. Grubb's test is used to detect outliers in data sets. Outliers decrease the accuracy of a regression model significantly and therefore, should be eliminated. Outliers can have a substantial influence on the accuracy and reliability of regression models. When outliers exist, they can distort the estimated parameters of the model, potentially leading to biased coefficient estimates. Since regression models aim to minimize the sum of squared residuals, outliers with large deviations can exert disproportionate influence during the fitting process, potentially skewing the regression line. This, in turn, can lead to inaccuracies in predictions and undermine the model's performance. Additionally, outliers may violate key assumptions of regression analysis, such as normality, constant variance, and linearity, further complicating model interpretation and reliability. The rationale for removing outliers lies in the pursuit of improving model accuracy and robustness. By excluding extreme observations, the model can better capture the underlying relationship between variables without undue influence from atypical data points. This can result in more reliable parameter estimates, better model fit, and more accurate predictions, particularly when outliers stem from measurement errors or represent anomalies not reflective of the population being studied. However, it's essential to exercise caution in outlier removal, as outliers may sometimes carry valuable information or represent genuine observations of interest. Thus, the decision to remove outliers should be made judiciously, considering statistical techniques, domain knowledge, and the potential implications for model interpretation and generalizability.

The null hypothesis of Grubb's test assumes there are no outliers in the data while the alternative hypothesis assumes there is. If the test characteristic, G, corresponding to a certain value of the tested parameter is greater than a specific critical value, G_{Crit}, then this certain value is considered an outlier. The parameter considered in Grubb's test is the total construction cost, the output parameter of the model. Table 2 shows a summary of Grubb's test results.

Table 2 demonstrates that the test characteristic corresponding to the maximum value of the tested parameter, G_{max}, is greater than the critical value of the test characteristic at the 95% level of confidence and that the p-value for the maximum, P_{max}, is less than 0.025, the $\alpha/2$ value. This leads to rejecting the null hypothesis that assumes there are no outliers present in data sets. Therefore, the maximum value is considered an outlier and the data set to which the aforementioned maximum value corresponds to is eliminated from the study. Table 3 shows a summary of Grubb's test results after eliminating the outlier data set.

As shown in Table 3, the test characteristics corresponding to the minimum and the new maximum values, G_{min} and $G_{max,\ new}$, respectively, are both less than G_{Crit} and that their corresponding p-values are both greater than 0.025. This means that the null hypothesis, now holds and there are no more outliers to be eliminated. Therefore, the total number of projects included in the study is now 30 projects. In addition, residual errors for the output have been constructed. Figure 2 shows the probability distribution of the error of the model output. It demonstrates that the residuals have a normal distribution.

3.3. Defining model variables

During the factor identification phase, eight factors affecting the cost were identified from the literature, one of them was removed and other three were added by experts, making a total of 10 factors affecting the construction cost. Upon the end of the data collection phase, two more variables, namely "Joist span" and "Vertical load on the structure", were eliminated from the study due to the lack of data gathered on these two factors. Therefore, the developed model incorporated eight input variables, three of which are numeric and the other five are categorical variables. The regression model was developed using Minitab, a spreadsheet-based software. The scaling process for numeric variables is automatic in Minitab. Table 4 demonstrates the basic statistics of the numeric variables based on which the scaling process for the numeric input variables and the output is performed. It is noted that the

Table 2. Grubb's test summary

Parameter	Values numbers	Mean	Standard deviation	Maximum Value	Minimum Value	G_{Crit}	G_{max}	G_{min}	P_{max}	P_{min}
Total building and civil work cost (SAR)	31	10,074,377	6182452.6	28,505,271	2,379,500	2.924	2.981	1.245	0.007	0.181

Table 3. Grubb's test results after eliminating outliers

Parameter	Values number	Mean	Standard deviation	Maximum Value	Minimum Value	G_{Crit}	$G_{max,\ new}$	G_{min}	$P_{max,\ new}$	P_{min}
Total building and civil work cost (SAR)	30	9,460,014	5237992.8	20,837,780	2,379,500	2.908	2.1722	1.3518	0.041	0.181

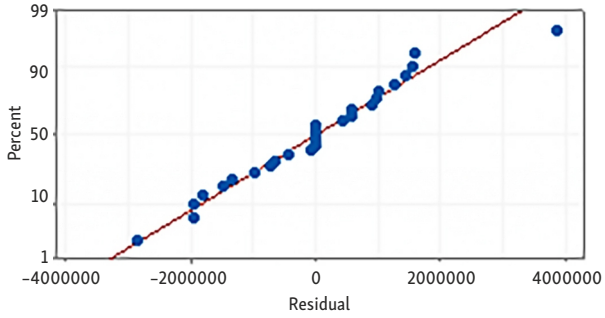

Figure 2. Probability plot of the error of model's output

data is uniformly distributed, indicating that the model has a good level of accuracy. As for the categorical variables, the "Existence of mezzanine" variable included 2 categories labelled in Minitab with numbers 1&2, the "Type of structure" variable included 8 categories labelled with numbers from 1 to 8, the "Type of steel used" variable included 2 categories labelled with numbers 1&2, the "Type of insulation material" variable included 3 categories labelled with numbers from 1 to 3 and the "Location" variable included 3 categories labelled with numbers from 1 to 3.

3.4. Input variables correlations

Identifying the correlations between the input variables is an important step in developing a regression model. Correlation can be defined as a measurement of the degree with which the variables are related. The values of correlation coefficient, R, ranges between 0 and 1, and can be either positive or negative. When the value of a variable changes and another variable behaves the same, then, the two variables are said to be positively correlated. On the other hand, when two variables behave opposite to one another in terms of increasing or decreasing then, the two variables are said to be negatively correlated. To solve a regression problem with the greatest number of independent variables as possible is the purpose of multicollinearity. For that purpose, correlation coefficients between the input variables were determined as shown in Table 5. A very high value of a correlation coefficient suggests that the two variables are highly dependent and therefore, there is no need to use one of them in the model development.

Table 5 shows that the highest value of a correlation coefficient is 0.499 between the area of the structure and the date of construction. However, this value of correlation and the other values shown in Table 5 are all not high enough (all are less than 0.75) to indicate dependency among the variables. As a result, all the variables listed in the table were considered to be independent and therefore, all the variables are used in the model development. Uncorrelated variables in the correlation test can be attributed to a variety of factors inherent in the data or to the nature of the relationships between the variables. One prominent reason for uncorrelated variables is independence, which occurs when changes in one variable do not have a systematic influence on changes in another. In addition, variables may be uncorrelated if there is no discernible linear relationship among them. on the other hand, when variables are correlated, it suggests a pattern where changes in one variable tend to correspond with changes in another. This correlation can arise from various sources, including causal relationships, common underlying factors, time lags, interaction effects, and measurement error. Thus, understanding the correlation structure

Table 4. Numeric variables' statistics

Variable	Maximum	Minimum	Mean	Standard deviation
Numeric input				
Area of structure	18,945	617	5,877	4345.3081
Eave height of structure	19	6	11	2.7759
Date of construction	2020	2016	2018	1.6189
Output				
Total construction cost (SAR)	20,837,780	2,379,500	9,460,014	5237992.8

Table 5. Correlations of input variables

	A	EH	EoM	ToSr	ToSt	ToI	D	L
A	1							
EH	–0.135	1						
EoM	–0.125	0.171	1					
ToSr	–0.251	0.277	0.122	1				
ToSt	0.069	0.236	0.149	0.399	1			
ToI	–0.187	–0.202	–0.270	–0.070	–0.403	1		
D	**0.499**	–0.128	–0.016	–0.372	–0.034	0.066	1	
L	–0.186	0.208	0.104	0.175	0.132	–0.152	–0.426	1

Note: A – Area of the structure; EH – Eave height of the structure; EoM – Existence of mezzanine in the structure; ToSr – Type of the structure; ToSt – Type of steel used in the construction; ToI – Type of Insulation; D – Date of construction; L – Location of construction.

of the data is essential for selecting relevant variables and building regression models that effectively capture the underlying relationships, ultimately leading to more reliable predictions.

3.5. Box plots

Box plots were used on the continuous numeric variables of the model. A box plot is a useful tool that gives an overview of the distribution of data for a certain variable. The graph of a box plot includes a box with its upper and lower boundaries representing the third and first quartiles, respectively, of the considered variable. It also includes a line drawn in the middle of the box representing the median value. The graph also shows the minimum and maximum values of the considered variable. If the data for a given variable is primarily between the median and maximum value then, the distribution shape for that certain variable would be skewed to the right. On the other hand, if the data for a given variable is primarily between the median and maximum value then, the distribution shape for that certain variable would be skewed to the left. Finally, if the data of a certain variable is equally distributed above and below the median then, the distribution shape of that certain variable would be symmetric. Furthermore, box plots offer a comprehensive visualization of the distribution of a dataset, enabling a nuanced interpretation of distribution shapes and the identification of potential outliers, both of which are crucial in understanding the underlying data and its implications for regression analysis. The central line within the box represents the median, providing insight into the central tendency of the data. The box itself illustrates the interquartile range (IQR), encapsulating the middle 50% of the data and providing a measure of variability. The whiskers extending from the box depict the range of the data, typically reaching 1.5 times the IQR or the minimum and maximum values within this range, whichever is shorter. Potential outliers, depicted as individual points beyond the whiskers or as distinct data points, signify observations that deviate significantly from the bulk of the data and may warrant further investigation. These outliers could indicate unusual or erroneous values, potentially influencing the estimation and interpretation of regression models. Therefore, a detailed examination of box plots, considering distribution shapes and outlier presence, is essential for informed decision-making in regression analysis, facilitating the identification and management of influential data points to ensure the accuracy and reliability of the model results. Figure 3 displays the box plot of the 'Total construction cost'. The first quartile value is 4,842,630, the median value is 9,120,000 and the third quartile value is 14,019,897.

3.6. Scatter plots

A scatter plot is useful tool that demonstrates the relationship between two variables. It is easy to plot and to extract the information from. As in box plots, a scatter plot shows

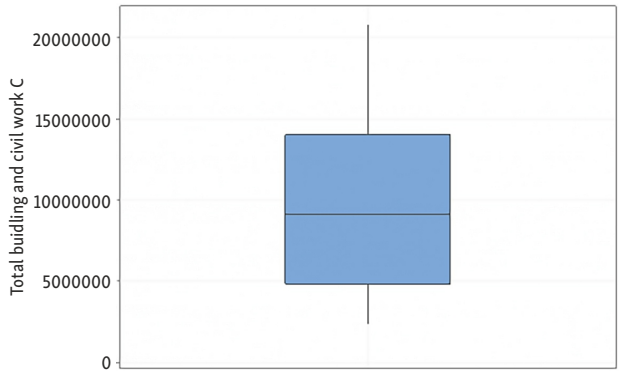

Figure 3. Box plot of total construction cost

the range of data flow, meaning that the minimum and the maximum are determined easily. Scatter plots were utilized here to demonstrate the relationship between the three continuous numeric independent variables with the dependent (output) variable of the developed model. Figure 4 shows the scatter plots of the 'Area of the structure', 'Eave height of the structure' and 'Date of construction', respectively, against the 'Total cost of construction'. By comparing the slopes of the three regression lines in the figures, it is clear that the slope in Figure 4a is the greatest. This indicates that the 'Area of the structure' has the greatest impact on the model output.

3.7. Setting the model's parameters

In this paper, the proposed model is based on a multiple linear regression model. The data points utilized in model development include information about the total construction cost of industrial steel structures, which is incorporated in the model as the dependent variable (i.e., model output). It also includes information about the variables influencing the construction cost which are considered as the model-independent variables or the input parameters. Therefore, the developed model has eight input parameters and one output.

Afterwards, the data sets from the 30 projects were divided into two categories to be utilized in the model training and validation. Data from 27 projects were used in the model training and data from 3 projects were set aside and used for model validation.

3.8. The proposed model development

The multiple-linear regression model was developed after indentifying model paperameter and dividing the data sets. Minitab software was used to generate the model. The method of least square error was used in the model generation. In this method, the difference between the output of the model and the target obtained from the dependent variable for each data set is calculated. Theses valuse, or errors, are then squared and summed together. The methed followed aims to minimize the aformentioned summation. The model with the least summation is considered the output of this process.

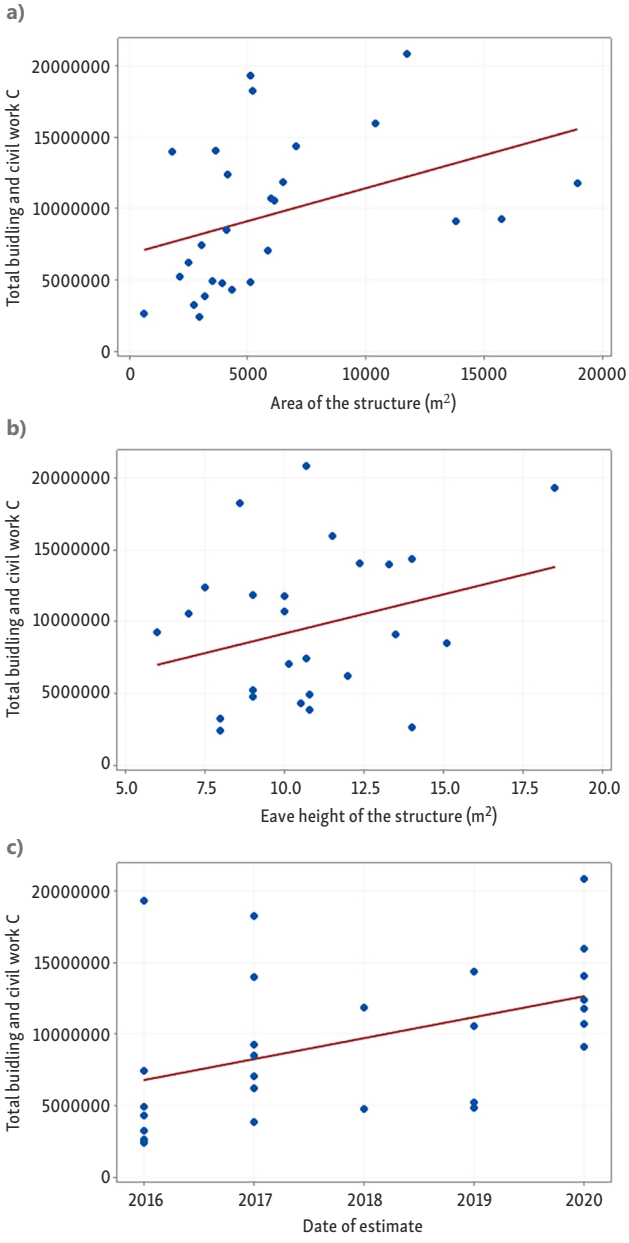

Figure 4. Scatter plot of input parameters against model output: a – Area of the structure; b – Eave height of the structure; c – Date of construction

The summation of the squared errors can be calculated by the following eqation:

$$RSS = \sum_{i=1}^{n} \left(y_i - b_0 - b_1 x_{i1} - b_1 x_{i2} - \ldots - b_p x_{ip} \right)^2, \quad (1)$$

where y_i refers to the target from the i^{th} data set; b_0, b_1, b_2, ... and b_p represent the coefficients from the regression model and RSS is the summation of the squared errors. The purpose of the method is to minimize the RSS. Applying the procedure described above and by using Minitab software, the following multiple linear regression model is generated:

Total building and civil work cost = 1135979463 + 804 Area of the structure (m^2)
+ 576794 Eave Height of the structure (m)
+ β_{EOMi} The existence of mezzanine (i)
+ β_{ToSrj} Type of the structure (j)
+ β_{ToStk} Type of steel (k) used in the construction
+ β_{Toll} Type of insulation material (l)
– 562213 Date of estimate
+ β_{Lm} Location (m) of construction,

where β_{EOMi} is the coefficient of Existence of mezzanine i, β_{ToSrj} is the coefficient of Type of the structure j, β_{ToStk} is the coefficient of Type of steel used k in the construction, β_{Toll} is the coefficient of Type of insulation l, and β_{Lm} is the coefficient of Location m of construction.

Table 6 shows the developed model's analysis of variance, or ANOVA. It shows a value of R^2 of 92.82%. This implies that the developed model can explain 92.82% of the error, which reveals there is a strong correlation between the model output and the target. Figure 5 also support this indication.

3.9. Model validation

Figure 6 demonstrates the training process of the regression model. The x-axis shows the projects involved in the model training. The y-axis shows the projects' construction cost in SAR. The graph shows two sets of data points corresponding to each project on the x-axis. The points in blue represent the actual costs of the projects involved in the training of the model, which represent the targets the model is trying to achieve. The points in orange are the

Table 6. The developed model's analysis of variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	16	6.79170E+14	4.24481E+13	8.08	0.001
Area of the structure (m^2)	1	5.97065E+13	5.97065E+13	11.37	0.007
Eave height of the structure (m	1	4.08828E+13	4.08828E+13	7.78	0.019
The existence of mezzanine in	1	1.19951E+13	1.19951E+13	2.28	0.162
Type of the structure	7	2.67343E+14	3.81919E+13	7.27	0.003
Type of steel used in the construction	1	1.04750E+14	1.04750E+14	19.95	0.001
Type of insulation material	2	2.09745E+14	1.04872E+14	19.97	0.000
Date of estimate	1	5.85937E+12	5.85937E+12	1.12	0.316
Location of construction	2	4.11663E+13	2.05831E+13	3.92	0.055
Error	10	5.25158E+13	5.25158E+12		
Total	26	7.31686E+14			
R^2 = 92.82%					

model outputs representing the estimated cost for each project involved in the model training. The coefficient of determination, R^2, is an indication of how well the developed model represents the situation understudy or, how well data points fit a regression curve or line. The value of R^2 ranges between 0, which corresponds to a no fit at all with the situation under study, and 1, which corresponds to a perfect fit with the situation under study. Another term associated with regression models is the adjusted R^2. Adjusted R^2 is similar to R^2 in purpose with the difference between them that the adjusted R^2 accounts and adjusts for the number of data points and independent variables used in the regression model. If useful variables are added to model, adjusted R^2 increases. However, adjusted R^2 is always less than or equal to R^2. From Table 6, it is noticed that the value of R^2 is 92.82% which indicates good performance of the developed regression model.

Once the regression model was developed and trained, the model was validated. Data from 3 actual projects obtained from the SIDF, which the model was not exposed at any point during the development phase, were used to validate the model. Table 7 shows the results of the validation process. It shows the actual costs of the projects used to validate the model in comparison with their corresponding estimated costs obtained from the model. For all three projects, the developed model accurately estimated costs by more than 88%. Therefore, the performance of the developed model was considered more than satisfactory.

Table 7. Model validation

Project #	Actual cost (SAR)	Estimated cost	% Accuracy
1	13,371,461	12,720,418	95.13%
2	9,890,290	10,998,967	88.79%
3	2,514,589	2,332,657	92.76%

3.10. Sensitivity analysis

After the model has been developed, tested, and validated, sensitivity analysis was performed. Sensitivity analysis investigates the effect of change in the values of the model's input parameters on the output, i.e., the construction cost. The analysis was carried out by changing the values

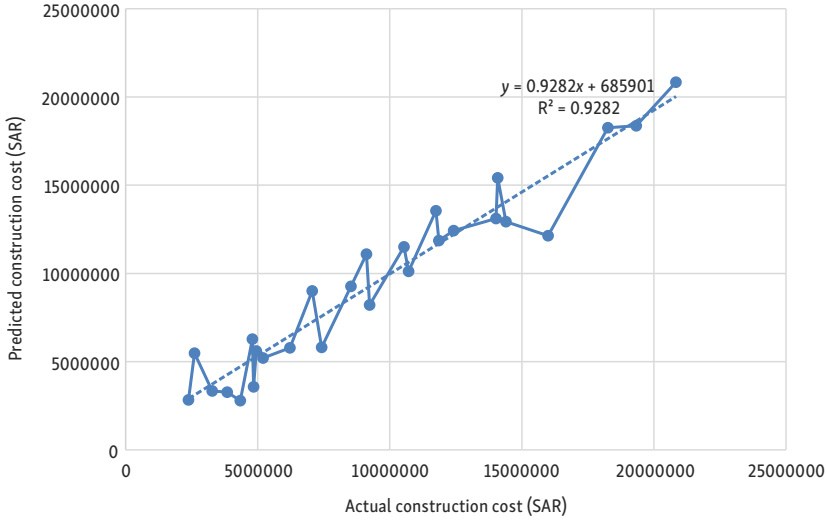

Figure 5. Actual versus predicted of construction cost for the training data

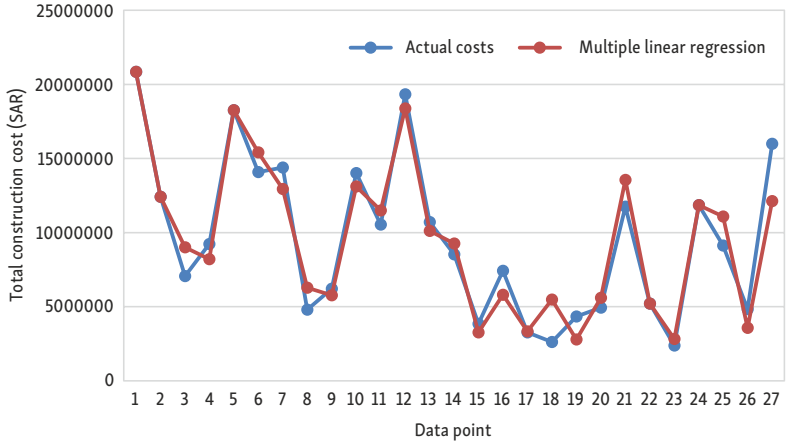

Figure 6. Performance of model training

of one of the three continuous parameters at a time while keeping the other two fixed at their means and the five categorical parameters fixed at their respective most frequent category. Figure 7 demonstrates the change of 10%, 20% and 30% of the range of each continuous parameter, above and below their respective means, and the effect of these changes on the construction cost.

It is inferred from Figure 7 that the model is most sensitive to change in the 'Area of structure', meaning that the 'Area of structure' affects the construction cost of industrial steel structures the most. Changes in the area, eave height, and date of construction can significantly impact the construction cost of industrial steel structures. When considering the area of the structure, enlarging it generally results in higher costs due to the increased requirement for steel beams, columns, and cladding materials. Larger areas may also necessitate more intricate structural designs and additional support systems, further driving up expenses. Conversely, reducing the area can lead to cost savings by requiring fewer materials and less labor. However, it's crucial to ensure that the reduced area still meets the operational needs of the industrial facility. Regarding eave height, taller structures typically incur greater costs as they require longer steel columns and beams, as well as larger roof and wall panels. Additionally, higher eave heights may necessitate the use of specialized equipment during construction, increasing labor costs. Lowering the eave height can reduce costs by requiring less steel material and simplifying construction processes. However, it's essential to consider how this change might affect the functionality and efficiency of the industrial space. The date of construction also influences costs, with older structures potentially requiring retrofitting or upgrades to meet modern safety and building code standards, adding to the overall expenses. Conversely, newer construction methods and materials may offer cost-saving opportunities through improved efficiency and durability. In summary, changes in area, eave height, and construction date can significantly impact the construction cost of industrial steel structures, requiring careful consideration to optimize budgeting and project feasibility.

4. Discussion of the findings

In this research, several statistical analysis tests are conducted to develop an accurate and reliable model for predicting the construction cost of industrial steel structure. The correlation test among the input variables indicated that all input factors are required to develop the proposed model. In addition, the effect of each factor on the output is measured using individual factor analysis and regression analysis. Although all the identified factors have a significant effect on the repose, the result indicated that the area of the structure and eave height of the structure are the most important factors that affect the construction cost of industrial steel. This is due to the fact that increasing the structure area usually requires the use of more steel, which raises the construction cost. Increasing the structure's height also requires the use of specialized equipment, skills, and tools. Furthermore, several regression models are tested based on data from actual projects. The best and most accurate model with an accuracy of 92.82% is selected to predict the construction cost of industrial steel structures. The training phase of the model produced satisfactory results, successfully avoiding both overfitting and underfitting. Moving on to the validation stage, the chosen model was tested on new and diverse datasets for forecasting. This entailed testing the model on three completely new projects to ensure its precision. These projects had never been attempted before, ensuring the model's adaptability to new data. The results of the developed model for estimating the cost of industrial steel structures demonstrate its dependability and accuracy. To increase its credibility, the model's performance was validated against actual project costs. This validation process demonstrates the model's ability to uncover hidden relationships within the dataset, which improves the accuracy of construction cost predictions. In addition, although two of these factors were removed due to a lack of data, the findings showed that the model is robust, with the value of R-seq being close to the value of R-Seq adjusted. This means that adding more factors has no effect on the model's accuracy.

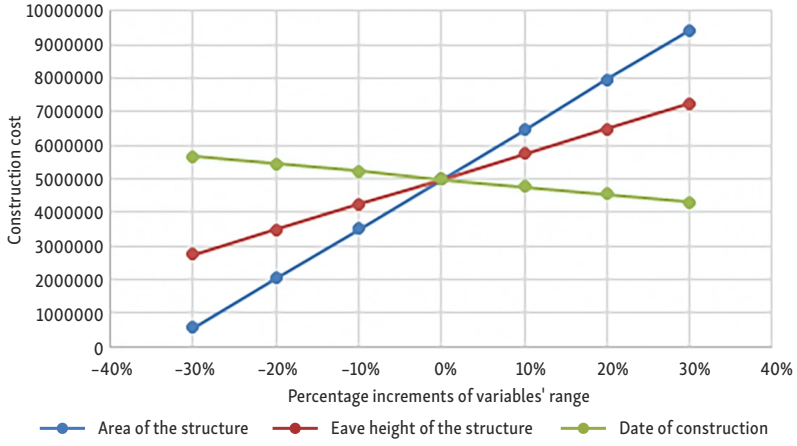

Figure 7. Sensitivity analysis of continuous parameters

The use of multiple regression models to forecast the cost of industrial steel structures is supported by statistical theory, particularly within the context of linear regression principles. Linear regression attempts to characterize the relationship between a dependent variable and one or more independent variables by fitting a linear equation to observed data. In this study, the relationship between the cost estimation and the factors affecting the output is linear as shown in the scatterplots. In addition, the fundamental theory supporting the use of multiple linear regression for prediction includes independence. This property is stratified in this study, where the observations used in the multiple linear regression model are independent of each other. In addition, the residuals of the data follow a normal distribution. This is clear from Figure 2. In this study, all of these assumptions are met. Therefore, the proposed model provided a robust framework for making predictions and understanding the relationship between variables. In addition, the developed mode can estimate the future cost with high accuracy. Furthermore, the proposed model's effectiveness is validated using entirely new data that the model had never seen before. The results show that the developed model maintains high accuracy throughout the validation phase. The study findings and insights can be summarized into the following points:

- The area and eave height of the structure have the major impact on the construction cost, as larger structure areas require more steel and greater heights necessitate specialized equipment and skills.
- The results indicate that the model is robust as evidenced by high values of R-seq and R-Seq adjusted. Thus, the considered factors are sufficient to predict the construction cost of steel structures.
- The cost estimation model demonstrates high accuracy of 92.82% which can be used effectively to estimate the construction cost of steel structures.
- The cost estimation model is useful for early-stage cost estimates in Saudi Arabia's industrial sector which offers business owners quick, accurate, and low-effort cost estimates compared to traditional cost estimation techniques.
- The cost estimation model is beneficial for assessing projects at the early stages of projects' life cycles. This is due to the fact that the pieces of data required as inputs to the developed model are simple and easily obtained at the projects' conceptual stages.
- The developed model is useful for business owners to assess the feasibility of projects in the industrial sector.
- The developed model can be generalized and adjusted to estimate the construction cost in other sectors such as private, public, and energy projects, etc.

5. Conclusions

The study proposed a multiple linear regression model for estimating the costs associated with industrial steel structures, a critical sector in the industrial landscape. Furthermore, the research aimed to identify the factors that influence the construction costs of these structures. A review of existing literature and consultations with SIDF experts led to the identification of eight parameters that influence construction costs. These factors were used as input parameters in the developed model. Three of the parameters, namely, the area of the structure, eave height of the structure and date of construction, are continuous parameters. The remaining factors, which are the existence of mezzanine in the structure, type of the structure, type of steel used in the construction, type of Insulation, and location of construction, are categorical parameters. Data from a total 30 projects executed under the supervision of the SIDF were collected to be utilized in model development. Out of 30 projects, data from a subgroup of 27 projects was used to develop a cost estimation multiple linear regression model. The developed model has a value of R^2 of 92.82%. This indicates that the proposed model is capable of predicting the cost of industrial steel structures. Unseen actual costs data form three projects were used to validate the regression model. For all three projects, the model adequately estimated the costs with an accuracy of more than 88%. Finally, a sensitivity analysis was carried out for the three continuous parameters. The results show that the area and eave height of the structure have the major impact on the construction cost of industrial steel structures.

This study advances the field of construction cost estimation by integrating data-driven techniques with industry-specific parameters. The developed model serves as a powerful tool for business owners, investors, and decision-makers in the industrial sector. By accurately predicting construction costs of steel structure of industrial projects, the model enables more informed budgeting and financial planning. Consequently, the developed model will be of substantial help for businessmen and decision makers looking to strategic investments in the industrial sector. For instance, policy-makers could leverage the findings to streamline funding allocation, enhance cost management strategies, and promote transparent bidding processes in industrial sector projects. Additionally, this study underscores the potential of data-driven decision-making in the construction industry, which could further refine cost estimation models. The study also forms a basis for future research work in the field of cost estimation of industrial steel structures in Saudi Arabia. The study can be extended in different dimensions. For example, the effect of uncertainty in some parameters can be addressed. This could involve employing techniques such as Bayesian inference or Monte Carlo simulations to account for variability and error in input parameters, thus providing more reliable predictions and decision-making frameworks.

In addition, collecting additional real-world data to train and validate the model would enhance the accuracy of the model. This expansion allows a more comprehensive understanding of system dynamics, addresses potential biases, and allows for better adaptation to changing trends. As a result, the model improves its predictive ca-

pabilities, allowing for more informed decision-making. Moreover, recognizing the presence of nonlinear relationships between variables is critical for developing accurate predictive models. While linear regression assumes a linear relationship, real-world data frequently shows nonlinear patterns. Future research could address this limitation through investigations of nonlinear regression models or other advanced techniques. Polynomial regression, generalized additive models, and machine learning algorithms such as decision trees and neural networks can help to better capture the data's complex relationships. By embracing these methods, this will improve model accuracy and gain a better understanding of the nuanced dynamics between variables.

Acknowledgements

The authors would like to express their gratitude to King Fahd University of Petroleum and Minerals, as well as the Saudi Industrial Development Fund, for their cooperation and assistance in conducting this study. This research has been funded by the Interdisciplinary Research Center of Construction and Building Materials at KFUPM through grant INCB2210.

Author's contributions

This work was equally contributed by all authors.

Disclosure statement

The authors reported no potential conflicts of interest.

Data availability

The data supporting the study's findings are accessible from the corresponding author upon reasonable request.

References

Aibinu, A. A., & Pasco, T. (2008). The accuracy of pre-tender building cost estimates in Australia. *Construction Management and Economics*, *26*(12), 1257–1269. https://doi.org/10.1080/01446190802527514

Alshamrani, O. S. (2017). Construction cost prediction model for conventional and sustainable college buildings in North America. *Journal of Taibah University for Science*, *11*(2), 315–323. https://doi.org/10.1016/j.jtusci.2016.01.004

Alshibani, A., & Alshamrani, O. S. (2017). ANN/BIM-based model for predicting the energy cost of residential buildings in Saudi Arabia. *Journal of Taibah University for Science*, *11*(6), 1317–1329. https://doi.org/10.1016/j.jtusci.2017.06.003

Arafah, M., & Alqedra, M. (2011). Early stage cost estimation of buildings construction projects using artificial neural networks. *Journal of Artificial Intelligence*, *4*(1), 63–75. https://doi.org/10.3923/jai.2011.63.75

Badawy, M. (2020). A hybrid approach for a cost estimate of residential buildings in Egypt at the early stage. *Asian Journal of Civil Engineering*, *21*, 763–774. https://doi.org/10.1007/s42107-020-00237-z

Badra, I., Badawy, M., & Attabi, M. (2020). Conceptual cost estimate of buildings using regression analysis in Egypt. *IOSR Journal of Mechanical and Civil Engineering*, *17*(5), 29–35.

Cho, H.-G., Kim, K.-G., Kim, J.-Y., & Kim, G.-H. (2013). A comparison of construction cost estimation using multiple regression analysis and neural network in elementary school project. *Journal of the Korea Institute of Building Construction*, *13*(1), 66–74. https://doi.org/10.5345/JKIBC.2013.13.1.066

Chopra, P., Sharma, R. K., & Kumar, M. (2014). Regression models for the prediction of compressive strength of concrete with and without fly ash. *International Journal of Latest Trends in Engineering and Technology*, *3*(4), 400–406.

Chro, A. (2021). Early cost estimation models based on multiple regression analysis for road and railway tunnel projects. *Arabian Journal of Geosciences*, *14*, Article 972. https://doi.org/10.1007/s12517-021-07359-x

Dang, C. N., & Le-Hoai, L. (2018). Revisiting storey enclosure method for early estimation of structural building construction cost. *Engineering, Construction and Architectural Management*, *25*(7), 877–895. https://doi.org/10.1108/ECAM-07-2015-0111

Dharwadkar, N. V., & Arage, S. S. (2018). Prediction and estimation of civil construction cost using linear regression and neural network. *International Journal of Intelligent Systems Design and Computing*, *2*(1), 28–44. https://doi.org/10.1504/IJISDC.2018.092554

El-Sawah, H., & Moselhi, O. (2014). Comparative study in the use of neural networks for order of magnitude cost estimating in construction. *ITcon*, *19*, 462–473.

Fragkakis, N., Marinelli, M., & Lambropoulos, S. (2015). Preliminary cost estimate model for culverts. *Procedia Engineering*, *123*, 153–161. https://doi.org/10.1016/j.proeng.2015.10.072

Gujarati, D. N. (2003). *Basic econometrics* (4th ed.). McGraw-Hill.

Günaydın, H. M., & Doğan, S. Z. (2004). A neural network approach for early cost estimation of structural systems of buildings. *International Journal of Project Management*, *22*(7), 595–602. https://doi.org/10.1016/j.ijproman.2004.04.002

Gunduz, M., Ugur, L. O., & Ozturk, E. (2011). Parametric cost estimation system for light rail transit and metro trackworks. *Expert Systems with Applications*, *38*(3) 2873–2877. https://doi.org/10.1016/j.eswa.2010.08.080

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning*: *Data mining*, *inference*, *and prediction* (2nd ed.). Springer. https://doi.org/10.1007/978-0-387-84858-7

Hegazy, T., & Ayed, A. (1998). Neural network model for parametric cost estimation of highway projects. *Journal of Construction Engineering and Management*, *124*(3), 210–218. https://doi.org/10.1061/(ASCE)0733-9364(1998)124:3(210)

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. Springer. https://doi.org/10.1007/978-1-4614-7138-7

Kamarthi, S., Sanvido, V., & Kumara, S. (1992). Neuroform—Neural network system for vertical formwork selection. *Journal of Computing in Civil Engineering*, *6*(2), 178–199. https://doi.org/10.1061/(ASCE)0887-3801(1992)6:2(178)

Kim, G. H., An, S. H., & Kang, K. I. (2004). Comparison of construction cost estimating models based on regression analysis, neural networks, and case-based reasoning. *Building and Environment*, *39*(10), 1235–1242. https://doi.org/10.1016/j.buildenv.2004.02.013

Kim, G.-H., Shin, J.-M., Kim, S., & Shin, Y. (2013). Comparison of school building construction costs estimation methods using regression analysis, neural network, and support vector machine. *Journal of Building Construction and Planning Research*, *1*, 1–17. https://doi.org/10.4236/jbcpr.2013.11001

Latief, Y., Wibowo, A., & Isvara, W. (2013). Preliminary cost estimation using regression analysis incorporated with adaptive neuro fuzzy inference system. *International Journal of Technology*, *4*(1), 63–72. https://doi.org/10.14716/ijtech.v4i1.102

Lowe, D. J., Emsley, M. W., & Harding, A. H. (2006). Predicting construction cost using multiple regression techniques. *Journal of Construction Engineering and Management*, *132*(7), 750–758. https://doi.org/10.1061/(ASCE)0733-9364(2006)132:7(750)

Lu, M., AbouRizk, S. M., & Hermann, U. H. (2000). Estimating labor productivity using probability inference neural network. *Journal of Computing in Civil Engineering*, *14*(4), 241–248. https://doi.org/10.1061/(ASCE)0887-3801(2000)14:4(241)

Mahamid, I. (2011). Early cost estimating for road construction projects using multiple regression techniques. *Australasian Journal of Construction Economics and Building*, *11*(4), 87–101. https://doi.org/10.5130/AJCEB.v11i4.2195

Mahamid, I. (2013). Conceptual cost estimate of road construction projects in Saudi Arabia. *Jordan Journal of Civil Engineering*, *7*(3), 285–294.

Mahamid, I., & Bruland, A. (2010). Preliminary cost estimating models for road construction activities. In *FIG Congress 2010. Facing the Challenges – Building the Capacity*, Sydney, Australia.

Moselhi, O., & Siqueira, I. (1998). Neural networks for cost estimating of structural steel buildings. In *Proceedings of the AACE International Transactions*, *IT/IM.06*. American Association of Cost Engineers (AACE), Morgantown, WV.

Ofori-Boadu, A. N. (2015). Exploring regression models for forecasting early cost estimates for high-rise buildings. *The Journal of Technology, Management, and Applied Engineering*, *31*(5).

Petroutsatou, C., Lambropoulos, S., & Pantouvakis, J.-P. (2006). Road tunnel early cost estimates using multiple regression analysis. *Operational Research*, *6*, 311–322. https://doi.org/10.1007/BF02941259

Roxas, C. L. C., & Ongpeng, J. M. C. (2014). An artificial neural network approach to structural cost estimation of building projects in the Philippines. In *DLSU Research Congress 2014*, Manila, Philippines.

Sanni-Anibire, M. O., Zin, R. M., & Olatunji, S. O. (2021). Developing a preliminary cost estimation model for tall buildings based on machine learning. *International Journal of Management Science and Engineering Management*, *16*(2), 134–142. https://doi.org/10.1080/17509653.2021.1905568

Shin, Y. (2015). Application of boosting regression trees to preliminary cost estimation in building construction projects. *Computational Intelligence and Neuroscience*, *2015*(4), Article 149702. https://doi.org/10.1155/2015/149702

Siqueira, I. (1999). *Neural network-based cost estimating* [Master thesis]. Concordia University, Montreal, Quebec, Canada.

Sonmez, R. (2004). Conceptual cost estimation of building projects with regression analysis and neural networks. *Canadian Journal of Civil Engineering*, *31*(4), 677–683. https://doi.org/10.1139/l04-029

Sonmez, R., & Ontepeli, B. (2009). Predesign cost estimation of urban railway projects with parametric modeling. *Journal of Civil Engineering and Construction Management*, *15*, 405–409. https://doi.org/10.3846/1392-3730.2009.15.405-409

U.S. Department of Defense. (1995). *Parametric cost estimating handbook*. Department of Defense, Arlington, VA, USA.

Wang, X.-Z., Duan, X.-c., & Liu, J.-y. (2010). Application of neural network in the cost estimation of highway engineering. *Journal of Computers*, *5*(11), 1762–1766. https://doi.org/10.4304/jcp.5.11.1762-1766

Xue, X., Jia, Y., & Tang, Y. (2020). Expressway project cost estimation with a convolutional neural network model. *IEEE Access*, *8*, 217848–217866. https://doi.org/10.1109/ACCESS.2020.3042329