Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 247 (2024) 617–623

Procedia Computer Science

www.elsevier.com/locate/procedia

The 11th International Conference on Applications and Techniques in Cyber Intelligence

Construction and Application of Cost Prediction Model Based on Multiple Linear Regression Analysis

Liuyan Lin[a], Wei Jiang[a], Biao Chen[b,*], Jing Yu[b], Chenhong Zheng[b]

[a]*State Grid Fujian Electric Power Co., Ltd., Fuzhou, China*
[b]*State Grid Fujian Economic Research Institute, Fuzhou, China*

Abstract

Production cost forecasting is the key link to control the production cost of products, is an important step for enterprises to scientifically plan the future cost level and cost objectives, and is also the basis for enterprises to make scientific decisions. Taking women's clothing as an example, this paper introduces a simple and practical product production cost prediction method based on multiple linear regression mathematical model, describes the establishment process and prediction steps of multiple linear regression mathematical model in detail, and makes a prediction empirical research. It is found that the design difficulty and production volume are the key factors affecting the production cost of women's clothing, while the unit cost of raw materials and the man-hour required for production are not significant. The sample equation was fitted by the least square method, and the mathematical model was established, and the empirical analysis was carried out by SPSS22.0 software. The results show that the model has a high degree of fitting and forecasting ability, and provides a scientific and practical cost forecasting tool for enterprises.

Keywords: Multiple Linear Regression, Cost Prediction, Model Interpretation, Actual Costs

1. Introduction

The prediction of product production cost has become a difficult kind of prediction because of its wide range, numerous influencing factors and complex procedures, but because of its important role in the production and

* Corresponding author. Tel.: +0-000-000-0000 ; fax: +0-000-000-0000 .
E-mail address: chenbiao351@163.com

10.1016/j.procs.2024.10.074

operation of enterprises, it has been paid attention to by enterprises, and has attracted the attention and research of many scholars and institutions. Traditional cost forecasting methods often rely on experience judgment and historical data, lack of sufficient flexibility and accuracy, and are difficult to adapt to rapid changes in the market environment. With the advent of the era of big data, multiple linear regression analysis, as a classical statistical method, has gradually become a hot topic in cost prediction research because of its superiority in dealing with the linear relationship between variables. Multiple linear regression analysis can provide more accurate prediction results by establishing a linear relationship model between multiple independent variables and dependent variables, which can consider various factors affecting the cost more comprehensively. The method for predicting manufacturing costs is grounded in a mathematical model derived from the multiple linear regression equation. This model is employed to construct a formula that can analyze and forecast the costs associated with product manufacturing. Among the various forecasting techniques, the one that utilizes a multiple linear regression mathematical model stands out for its distinct features, including a well-defined theoretical framework, straightforward architecture, uncomplicated computation, robust applicability, and excellent fit to the data. This paper introduces the product cost prediction method based on the multiple linear regression mathematical model and the establishment and calculation process of the mathematical model, and takes a computer product with a label produced by a computer company as an example to make a positive prediction study.

2. Related Works

Dasakaetal. (2023) Multiple linear regression analysis of foundation soil reinforcement using wrapped end technology, providing a new perspective for geotechnical engineering [1]. Neethu et al. (2022) studied the multiple linear regression of a bioconvective MHD hybrid nanofluid flowing through an exponentially stretching plate, taking into account radiative and dissipative effects [2]. Cai et al. (2023) studied the local brittleness of rocks based on the multiple linear regression method, taking the Shahejie Formation as an example [3]. Yamouletal. (2022) used multiple linear regression analysis to predict solar installation energy demand for residential buildings in Morocco [4]. Shafi (2023) analyzed Malaysian household income by K-means clustering and multiple linear regression model [5]. Rim (2022) assessed the feasibility of the universal coefficient estimation technique utilizing multiple linear regression analysis [6]. Wang et al. (2022) developed a predictive model for the effluent from wastewater treatment plants, enhancing the methodology for data processing and the integration of process parameters [7]. Shorabehetal. (2022) created a decision-making model grounded in the decision tree and particle swarm optimization algorithm, aimed at pinpointing the most suitable sites for solar power plant installations in Iran [8]. Liu et al. (2022) proposed a combination forecasting model based on mixed interval multi-scale decomposition, which was applied to interval-valued carbon price forecasting [9]. Wu & Wang (2022) constructed a hybrid model for water quality prediction based on artificial neural network, wavelet transform and long short-term memory [10]. Li et al. (2022) proposed a new typhoon wind speed prediction model for offshore wind farms based on improved PSO-Bi-LSTM and VMD [11]. Alzain et al. (2022) applied artificial intelligence to predict real estate prices in Saudi Arabia [12]. To sum up, multiple linear regression analysis, as a powerful statistical tool, has shown its potential for prediction and analysis in many fields, but the in-depth study of specific industries such as cost prediction is still insufficient, which provides the necessity and direction for the study of this paper. Dasakaetal. (2023) Multiple linear regression analysis of foundation soil reinforcement using wrapped end technology, providing a new perspective for geotechnical engineering [1]. Neethu et al. (2022) studied the multiple linear regression of a bioconvective MHD hybrid nanofluid flowing through an exponentially stretching plate, taking into account radiative and dissipative effects [2]. Cai et al. (2023) studied the local brittleness of rocks based on the multiple linear regression method, taking the Shahejie Formation as an example [3]. Yamouletal. (2022) used multiple linear regression analysis to predict solar installation energy demand for residential buildings in Morocco [4]. Shafi (2023) analyzed Malaysian household income by K-means clustering and multiple linear regression model [5]. Rim (2022) evaluated the applicability of the universal coefficient estimation method by multiple linear regression analysis [6]. Wang et al. (2022) constructed the effluent prediction model of sewage treatment plant, and optimized the data processing method and process parameter integration [7]. Shorabehetal. (2022) established a decision model based on decision tree and particle swarm optimization algorithm to determine the optimal location of solar power plant construction in Iran [8]. Liu et al. (2022) proposed a combination forecasting model based on mixed interval multi-scale

decomposition, which was applied to interval-valued carbon price forecasting [9]. Wu & Wang (2022) constructed a hybrid model for water quality prediction based on artificial neural network, wavelet transform and long short-term memory [10]. Li et al. (2022) proposed a new typhoon wind speed prediction model for offshore wind farms based on improved PSO-Bi-LSTM and VMD [11]. Alzain et al. (2022) applied artificial intelligence to predict real estate prices in Saudi Arabia [12]. To sum up, multiple linear regression analysis, as a powerful statistical tool, has shown its potential for prediction and analysis in many fields, but the in-depth study of specific industries such as cost prediction is still insufficient, which provides the necessity and direction for the study of this paper.

3. Methods

3.1. Multiple Linear Regression

Regression analysis is a statistical technique that employs mathematics to handle the interconnections among variables. The core concept of this analysis is to identify a mathematical formula that most accurately captures the relationship between independent and dependent variables, even though this relationship is not always rigidly defined. Multiple regression analysis, a variant of this method, examines the interplay among several variables, illustrating how the quantity of a phenomenon or object is influenced by the quantitative changes of multiple other phenomena or objects.

Its characteristic is that no matter how many factors affect the analysis object, as long as the analysis object is determined, the most important factors can always be found out through regression analysis. And through the regression model to establish the relationship between them, and finally through the test to verify the accuracy of this relationship, more suitable for the study of dependent variables to determine the independent variables have more changes in the phenomenon or things. When only one independent variable, X, predicts the dependent variable, y, the goal of linear regression is to fit as close as possible to all the scatter points in the plot through a straight line. Similarly, when a dependent variable contains two or more independent variables, the multiple linear regression model is:

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_k x_k + \varepsilon \tag{1}$$

In the formula (1), y Is the dependent variable, which is related to the $x_i (i = 1,2,\cdots,k)$. There is a linear correlation. $x_i (i = 1,2,\cdots,k$ is an independent variable and affects the value of the dependent variable. α Indicates the intercept; $\beta_i (i = 1,2,\cdots,k)$ represents the slope of each independent variable, that is, the weight parameter of the independent variable; ε is a random error. Random error ε must be met: ε is a random variable; ε Obey normal distribution; For any value of x, ε has a constant variance.

For the calculation of parameters α and β, the least square method is used to fit the sample equation. The specific way is: for the independent variable X and the factor. The change of the variable y is observed for many times, and n groups of observation data are obtained after n times of observation. There are n equations in the simultaneous equations, and the parameters can be obtained by using the least square method to solve the regression coefficients, and then the multiple linear regression mathematical model can be constructed by substituting the parameters into the corresponding equations. This mathematical model still needs to be tested, and we will not discuss the test process here. After passing the significance test, the mathematical model can be used to predict. It is worth emphasizing that the production cost value obtained by prediction is not an accurate value but only a fuzzy value, and the accuracy of prediction is related to the number and accuracy of independent variables. When k in the equation $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_k x_k + \varepsilon$ in tends to infinity, that is, when the number of independent variables is infinite, the predicted value tends to the true value.

3.2. Model Method of Cost Prediction

In enterprises, there are so many factors affecting the production cost of products that it is impossible to find an accurate mathematical model to express their interrelationship. However, because the research object-the production

cost of products (that is, the dependent variable) is determined and the influencing factors (independent variables) are changed, it is suitable to use the multiple linear regression method to study it. Based on the above considerations, we can use multiple linear regression equation to establish a mathematical model to discuss the mathematical relationship between the relevant variables, and then use the calculated mathematical model to predict the production cost of products. The following will discuss the establishment and calculation process of the mathematical model of the product cost prediction method based on the multiple linear regression mathematical model. If we regard the production cost of a product as a dependent variable (represented by y) and the influencing factors as independent variables (represented by X), the mathematical relationship between the production cost and the influencing factors can be simply written as:

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon \tag{2}$$

y represents the cost of the product.x_1, x_2, x_3, x_4 represent the unit cost of raw materials, the difficulty of design, the man-hours required for production and the production volume. Among the formulation,$\beta_1, \beta_2, \beta_3, \beta_4$ are the coefficients corresponding to the respective variables, which reflect the degree to which each factor contributes to the total cost.αIs the intercept term of the model, which represents the cost basis value when all the independent variables are zero. The unit cost of raw materials directly affects the basic cost of products, because raw materials are the basic elements of products. The higher the cost of raw materials, the higher the cost of products. The complexity and innovation of the design process can increase the cost of research and development, including the man-hours of designers, the techniques and tools used, etc. The more difficult the design is, the more research investment is required, and the higher the cost is. The number of man-hours required in the production process reflects productivity and labor costs. Longer hours usually mean higher labor costs, which increases the cost of the product. Production economies of scale can reduce the cost per unit to some extent, because fixed costs can be spread over more products. However, if the production volume exceeds a certain threshold, it may lead to higher costs, such as the need for additional equipment or management costs.

4. Experimental Process and Result

4.1. Experimental Process

Gathering information is the first step. We need to identify key variables-such as production volume, labor hours, design difficulty, and raw material costs. We selected a range of women's clothing from a clothing company to obtain quantitative measures, including production units and labor hours, and recorded actual costs. The temporal aspect is considered to capture changes over time. The design difficulty adopts the scoring system. Different women's clothing designs have different difficulties. We use a technical director and a designer to score and take the average to get the design difficulty. Strict quality control measures are implemented, including data validation checks, and data sets are pre-processed to handle any anomalies. The data collection strategy produced effective results in the area of cost prediction through multiple linear regression analysis. The results of data collection are shown in Table 1.

Table 1. Data Collection for Multiple Regression

Serial number	Unit cost of raw materials (yuan/unit)	Design difficulty (1-10 points)	Man-hours required for production (hours/unit)	Production (unit)	Production cost (yuan)
/	x1	x2	x3	x4	y
1	5.5	6	2.5	100	3250
2	7.25	8	3	150	6300
3	4.75	7	1.8	80	2228
4	6	9	3.2	200	8323
5	3.75	5	1.2	50	1095
6	8.5	7	4	300	12744
7	4.25	4	1.5	60	1170
8	5.75	8	2.8	120	4536
9	6.5	10	3.5	250	12259
10	4	6	2	90	2160

4.2. Experimental Results

We put it into SPSS22.0 and selected “multiple linear regression”, whose coefficients, standard errors, t statistics and P values are shown in Table 2.

Table 2. Coefficients and Standard Errors of the Linear Regression Equation

	Coefficients	Standard error	tStat	P-value
Intercept	-3276.075375	1208.792	-2.71021	0.042266
x1	237.3935032	364.0084	0.652165	0.543082
x2	374.8595609	173.0057	2.166746	0.082481
x3	-1057.909895	999.6465	-1.05828	0.338351
x4	53.00635875	6.845721	7.742992	0.000574

The multiple regression model is:$y = -3276.1 + 237.3x_1 + \beta_2 x_2 - 1057.9x_3 + 53.0x_4 + \varepsilon$

The intercept of the regression equation was -3276. The standard error of the 075375 was 1208.792. The t-statistic is -2. 71021, corresponding to a p-value of 0. 042266. This means that the intercept is statistically significant (usually P-value < 0.05 is considered significant), but the P-value is close to 0.05, indicating that its significance is not very strong. Intercepts, while significant, are usually of little practical significance in multiple linear regression unless all independent variables are zero, which rarely happens in practice.

X1 (raw material cost) and X3 (man-hours required to generate) do not have a significant impact on production cost, and more data or further analysis may be needed to determine their relationship to cost. X2 (design difficulty) shows some significance and has an impact on the production cost. In fact, designers need to keep up with fashion trends and predict and understand the needs of consumers. This requires designers to have a keen market insight and a deep understanding of fashion elements. The design of women's clothing needs to take into account the wearing needs of women of different body types to ensure that the clothing is both beautiful and comfortable. This requires designers to have a deep understanding of human body structure and be able to design clothes suitable for different body types. In the design process, designers need to find a balance between creativity and cost to ensure that the design is both competitive in the market and cost-effective. Therefore, the design difficulty has a great impact on the cost. X4 significantly affected the production cost, and the t-statistic was high, indicating that the production cost increased by an average of 53.0 yuan per unit increase in X4.

It can be seen from Table 3 that the fitting degree of the model is very high. Multiple R (coefficient of determination) is 0.996, a value very close to 1, indicating a strong linear relationship between the independent and dependent variables. R Square (coefficient of determination) is 0.991, which means that the model explains 99.1% of the variation in the dependent variable, and the model fit is very high. Even after considering the degrees of freedom, the fitting degree of the model is still very high.

Table 3. Fit of Model

MultipleR	0.995552846
RSquare	0.991125468
AdjustedRSquare	0.984025843
Standard error	553.6286837
Observation value	10

Table 4. Residual Analysis of Regression

Observation value	Forecast Y	Residual error
1	2934.607396	315.3926
2	6221.128138	78.87186
3	2811.831581	-583.832
4	8737.981779	-414.982
5	869.2741302	225.7259
6	13036.05437	-292.054
7	825.80194	344.1981
8	4486.4291	49.5709
9	11564.48306	694.5169
10	2577.408501	-417.409

The model has high predictive power, but because of the high P-value of XVariable1 and XVariable3, it may be necessary to reconsider whether these variables should be included in the model. Although the overall fit of the model is high, given the non-significance of XVariable1 and XVariable3, further model diagnostics or variable selection may be needed to simplify the model and improve the explanatory power. As shown in Table 4, the residuals of the observations range from -583.832 to 694.5169, which indicates that there is some deviation between the predicted value and the actual value. However, these deviations are relatively small due to the high RSquare values.

5. Conclusion

The research in this article indicates that multiple linear regression analysis is an effective production cost prediction method, especially suitable for predicting the cost of products such as women's clothing. The empirical research results indicate that design difficulty and production volume are significant factors affecting the production cost of women's clothing, while the influence of unit cost of raw materials and required production hours is not significant. The high fitting degree and predictive ability of the model provide scientific cost control and decision support for enterprises. The multiple linear regression mathematical model cost prediction method introduced in this article has the advantages of clear structure and simple calculation process. It is used for predicting the production cost of products and has good scientific and fitting properties. However, it has two disadvantages. Firstly, it is required that the production status of the production enterprise must be stable and there should be no significant fluctuations. If the production fluctuation is severe, its accuracy will be affected; Secondly, production data must be

accurate. If the accuracy of production data is poor, the accuracy of predictions will be compromised. With changes in market environment and production conditions, continuous monitoring and updating of models are key to ensuring prediction accuracy. By applying multiple linear regression analysis, enterprises can better plan costs, optimize resource allocation, and improve market competitiveness.

References

[1] Dasaka S M, Jaiswal S, Chauhan V B. Multiple linear regression analysis of foundation soil reinforced with geogrid using wraparound ends technique[M]//Smart Geotechnics for Smart Societies. CRC Press, 2023: 773-778.

[2] Neethu T S, Sabu A S, Mathew A, et al. Multiple linear regression on bioconvective MHD hybrid nanofluid flow past an exponential stretching sheet with radiation and dissipation effects[J]. International Communications in Heat and Mass Transfer, 2022, 135: 106115.

[3] Cai M, Wang Y, Zhao W, et al. Study on Local Brittleness of Rock Based on Multiple Linear Regression Method: Case Study of Shahejie Formation[J]. Geofluids, 2023(1): 2-14.

[4] Yamoul N, Dlimi L, Chakir B A. Prediction of Residential Building's Solar Installation Energy Demand in Morocco Using Multiple Linear Regression Analysis[J]. Energy Engineering, 2022, 119: 2135-2148.

[5] Shafi M A. K-means clustering analysis and multiple linear regression model on household income in Malaysia[J]. Int J Artif Intell, 2023, 12(2): 731-738.

[6] Rim C S. Evaluation of applicability of pan coefficient estimation method by multiple linear regression analysis[J]. Journal of Korea Water Resources Association, 2022, 55(3): 229-243.

[7] Wang R, Yu Y, Chen Y, et al. Model construction and application for effluent prediction in wastewater treatment plant: Data processing method optimization and process parameters integration[J]. Journal of Environmental Management, 2022, 302: 114020.

[8] Shorabeh S N, Samany N N, Minaei F, et al. A decision model based on decision tree and particle swarm optimization algorithms to identify optimal locations for solar power plants construction in Iran[J]. Renewable Energy, 2022, 187: 56-67.

[9] Liu J, Wang P, Chen H, et al. A combination forecasting model based on hybrid interval multi-scale decomposition: Application to interval-valued carbon price forecasting[J]. Expert Systems with Applications, 2022, 191: 116267.

[10] Wu J, Wang Z. A hybrid model for water quality prediction based on an artificial neural network, wavelet transform, and long short-term memory[J]. Water, 2022, 14(4): 610.

[11] Li J, Song Z, Wang X, et al. A novel offshore wind farm typhoon wind speed prediction model based on PSO–Bi-LSTM improved by VMD[J]. Energy, 2022, 251: 123848.

[12] Alzain E, Alshebami A S, Aldhyani T H H, et al. Application of artificial intelligence for predicting real estate prices: The case of saudi arabia[J]. Electronics, 2022, 11(21): 3448.