Research on the Price Prediction Technology of Power Grid Transformer Equipment Based on Least Squares-Multiple Linear Regression Model

Yan Li[1a], Yabo Shen[1b], Gaosheng Rong[1c], Suriya Dirisatii[1d], Xuehua Zhao[1e*]

[1]State Grid Xinjiang Economic Research Institute, Urumqi, Xingjiang, 830000, China

[a]e-mail:949761586@qq.com, [b]e-mail:ncepuyab@126.com, [c]e-mail:447381890@qq.com, [d]e-mail1247595932@qq.com, [e*]e-mail:304460740@qq.com

Abstract. This study addresses the challenges in predicting transformer equipment prices during the feasibility and design phases of power grid projects. By integrating historical data with a least-squares-based multiple linear regression model, this approach enhances prediction precision, enabling more effective pre-construction budgeting and dynamic cost management. The proposed methodology is designed to adapt to market fluctuations, offering robust guidance for cost control and investment planning in grid engineering.

Keywords: transformer, least squares, multiple linear regression, price prediction

1 Introduction

Market fluctuations in equipment and material prices challenge accurate cost predictions. Traditional methods are insufficient, necessitating advanced analytical techniques to integrate multifactor data for actionable insights.

Literature [1] analyzes the classification and influencing factors of blast furnace exhaust power generation engineering materials, identifying price influence paths to enhance bidding accuracy. Literature [2] categorizes material price influences into policy, economic, and enterprise factors, linking costs and profits to variables like supply-demand, competition, and raw material expenses for better cost control. Literature [3] proposes a material price management system for China Oil projects, introducing price libraries and budget tools to strengthen cost oversight. Literature [4] examines UHV equipment costs, focusing on copper, steel, R&D, and labor as key price determinants. Literature [5] correlates PPI fluctuations with material price changes, offering insights for power grid price forecasting. Literature [6] demonstrates high accuracy in multi-step predictions, but its complexity and computational cost make it less suitable for real-time applications in power grid projects. Similarly, the SVM-based multi-factor model. Literature [7] shows better alignment with real-world data compared to traditional neural networks, but it struggles with interpretability and

S. M. Ghadiri et al. (eds.), *Proceedings of the 2025 4th International Conference on Engineering Management and Information Science (EMIS 2025)*, Advances in Computer Science Research 121,
https://doi.org/10.2991/978-94-6463-736-6_28

requires careful tuning of hyperparameters. Literature [8] highlights ML-driven insulation material price prediction, achieving 95% accuracy and emphasizing GWP's impact for energy efficiency. Literature [9] evaluates synthetic battery material costs, identifying economical production strategies and research priorities. Literature [10] refines neural network predictors for energy raw material prices, enhancing accuracy by incorporating intrinsic signals and reevaluating lagged variable assumptions.

To sum up, this study proposes a least-squares-based multiple linear regression model with gray association analysis for variable selection. It balances prediction accuracy and efficiency, ensuring robustness in dynamic pricing, making it highly suitable for power grid transformer equipment cost prediction.

2 Analysis of the influencing factors of power grid transformer price

Gray association analysis calculates the association degree between factors to identify key drivers of transformer price fluctuations. In this study, raw material prices, labor costs, and monetary policy were selected based on qualitative analysis. The method quantifies the relationship between these variables and transformer prices, retaining the strongest predictors for regression analysis. This optimization enhances the model's accuracy and reliability.

Step 1 determines the analyzed sequence.

Based on qualitative analysis, dependent and independent variables were determined. Let the dependent variable data constitute the reference sequence X0, the respective variable data constitute the comparative sequence Xi′(I = 1, 2, ... n), n + 1 data series form the following matrix:

$$(X_0', X_1', ..., X_n') = \begin{bmatrix} x_0'(1) & x_1'(1) & ... & x_n'(1) \\ x_0'(2) & x_1'(2) & ... & x_2'(2) \\ ... & ... & ... & ... \\ x_n'(N) & x_n'(N) & ... & x_n'(N) \end{bmatrix}_{N\times(n+1)} \tag{1}$$

$X_i' = (x_i'(1), x_i'(2), ..., x_i'(N))^T, i = 0,1,2,...,n$ Where: N is the length of the variable sequence.

Step 2 dimensionizes the sequence of variables

In general, the original variable sequence has different dimensions or orders of magnitude, which is necessary to ensure the reliability of the analysis results. Undimensionless, the factor sequences form the following matrix:

$$(X_0, X_1, ..., X_n) = \begin{bmatrix} x_0(1) & x_1(1) & ... & x_n(1) \\ x_0(2) & x_1(2) & ... & x_2(2) \\ ... & ... & ... & ... \\ x_n(N) & x_n(N) & ... & x_n(N) \end{bmatrix}_{N\times(n+1)} \tag{2}$$

The commonly used dimensionless methods include mean method, initial value method and so on.

$$x_i(k) = \frac{x_i'(k)}{\frac{1}{N}\sum_{k=1}^{N} x_i'(k)} \tag{3}$$

$$x_i(k) = \frac{x_i'(k)}{x_i'(1)}$$

Step 3 To find the difference sequence, maximum difference and minimum difference.

Calculate the absolute difference between columns to form the matrix:

$$\begin{bmatrix} \varDelta_{01}(1) & \varDelta_{02}(1) & \dots & \varDelta_{0n}(1) \\ \varDelta_{01}(2) & \varDelta_{02}(2) & \dots & \varDelta_{0n}(2) \\ \dots & \dots & \dots & \dots \\ \varDelta_{01}(N) & \varDelta_{02}(N) & \dots & \varDelta_{0n}(N) \end{bmatrix}_{N\times n} \tag{4}$$

among:

$$\begin{bmatrix} \varDelta_{01}(1) & \varDelta_{02}(1) & \dots & \varDelta_{0n}(1) \\ \varDelta_{01}(2) & \varDelta_{02}(2) & \dots & \varDelta_{0n}(2) \\ \dots & \dots & \dots & \dots \\ \varDelta_{01}(N) & \varDelta_{02}(N) & \dots & \varDelta_{0n}(N) \end{bmatrix}_{N\times n} \tag{5}$$

i=1,2,…, n; k=1,2,…N

The maximum and minimum numbers in the absolute difference array are the maximum difference and minimum difference:

$$\max_{\substack{1\le i\le n\\1\le k\le N}} \{\varDelta_{0i}(k)\} \underline{\underline{\ \varDelta\ }} \varDelta(\max) \tag{6}$$

$$\min_{\substack{1\le i\le n\\1\le k\le N}} \{\varDelta_{0i}(k)\} \underline{\underline{\ \varDelta\ }} \varDelta(\min)$$

Step 4 calculates the correlation coefficient.

The data in the absolute difference array are transformed as follows:

$$\begin{bmatrix} \xi_{01}(1) & \xi_{02}(1) & ... & \xi_{0n}(1) \\ \xi_{01}(2) & \xi_{02}(2) & ... & \xi_{0n}(2) \\ ... & ... & ... & ... \\ \xi_{01}(N) & \xi_{02}(N) & ... & \xi_{0n}(N) \end{bmatrix}_{N\times n} \tag{7}$$

In the formula, the resolution coefficient ρ is within (0,1). Generally, the data is taken from 0.1 to 0.5. The smaller the ρ, the higher the difference between the correlation coefficients. correlation coefficient ξ_{0i}(K) is a positive number not more than 1, $\triangle_{0i}$(K) The smaller, ξ_{0i}The larger (k) is, it reflects the i th comparison sequence X_i With the reference sequence X_0The degree of association in stage k.

Step 5 calculates the correlation degree.

Compare sequence X_i With the reference sequence X_0The degree of association is reflected by N correlation coefficients (, averaging X_i take part in X_0The degree of association.

$$\begin{bmatrix} \xi_{01}(1) & \xi_{02}(1) & ... & \xi_{0n}(1) \\ \xi_{01}(2) & \xi_{02}(2) & ... & \xi_{0n}(2) \\ ... & ... & ... & ... \\ \xi_{01}(N) & \xi_{02}(N) & ... & \xi_{0n}(N) \end{bmatrix}_{N\times n} \tag{8}$$

Step 6 to sort by correlation degree.

The correlation degree of the compared sequence and the reference sequence is ranked from large to small, and the greater the correlation degree is, the more consistent the trend between the compared sequence and the reference sequence changes is.

3 Model building

This study proposes a multiple linear regression model for predicting transformer equipment prices, prioritizing factors like raw material prices, labor costs, and monetary policy. While the model offers simplicity, it struggles with non-linear relationships.

The multivariate linear regression prediction model considers the relationship between dependent variables and multiple independent variables, and the linear regression is expressed as follows: Y $X_1, X_2, ..., X_n$

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n + \varepsilon \tag{9}$$

$\beta_0, \beta_1, ..., \beta_n$ $n+1$ ε $\varepsilon \sim N(0, \sigma^2)$ Where is a regression coefficient to be estimated, which is random. The key in this formula is to solve the regression coefficient and random variables, and substitute the initial observation sequence into formula (1) to obtain the following structure: β_n ε m $\{X_{k1}, X_{k2}, ..., X_{kn}, Y_k\}$, $k = 1,2,...,m$

$$Y_1 = \beta_0 + \beta_1 X_{11} + \beta_2 X_{12} + ... + \beta_n X_{1n} + \varepsilon \tag{10}$$

$$Y_m = \beta_0 + \beta_1 X_{m1} + \beta_2 X_{m2} + ... + \beta_n X_{mn} + \varepsilon_m$$

$\varepsilon_1, \varepsilon_2, ..., \varepsilon_m$ $N(0, \sigma^2)$ They oppose and obey each other, so formula (2) is expressed in the form of a matrix as follows:

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ ... \\ Y_m \end{bmatrix}, X = \begin{bmatrix} 1 & X_{11} & X_{21} & ... & X_{1n} \\ 1 & X_{21} & X_{22} & ... & X_{2n} \\ ... & ... & ... & ... & ... \\ 1 & X_{m1} & X_{m2} & ... & X_{mn} \end{bmatrix}, \beta = \begin{bmatrix} \beta_1 \\ \beta_1 \\ ... \\ \beta_m \end{bmatrix}, \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_1 \\ ... \\ \varepsilon_m \end{bmatrix} \tag{11}$$

In this paper, the least-squares estimate is used to solve the regression coefficient, assuming the least-squares estimate, so that the original observation sequence can be expressed as: $\beta_0, \beta_1, ..., \beta_n$ $\beta'_0, \beta'_1, ..., \beta'_n$ Y_k

$$Y_k = \beta'_0 + \beta'_1 X_{k1} + \beta'_2 X_{k2} + ... + \beta'_n X_{kn} + e_k \tag{12}$$

$k = 1,2,...,m$ e_k ε_k Where, then, is the estimate of a random variable. The estimate of the order is: Y_k Y'_k

$$Y'_k = \beta'_0 + \beta'_1 X_{k1} + \beta'_2 X_{k2} + ... + \beta'_n X_{kn} \tag{13}$$

$$Y_k - Y'_k = e_k$$

$\beta'_0, \beta'_1, \beta'_2, ..., \beta'_n$ Y_k Y'_k According to the principle of least squares estimation method, the estimated value of regression coefficient should meet the sum of squares of the deviation from the estimated value to reach the minimum, that is, the minimum sum of squares of all error terms:

$$Q = \sum_{k=1}^{m} \left[y_k - (\beta'_0 + \beta'_1 X_{k1} + \beta'_2 X_{k2} + ... + \beta'_n X_{kn}) \right]^2 \tag{14}$$

According to the extreme value principle, the regression coefficient satisfies: $\beta'_0, \beta'_1, \beta'_2, ..., \beta'_n$

$$\begin{cases} \dfrac{\partial Q}{\partial \beta_0} = -2\sum_{k=1}^{m}[Y_k - Y_k'] = 0 \\ \dfrac{\partial Q}{\partial \beta_1} = -2\sum_{k=1}^{m}[Y_k - Y_k']X_{k1} = 0 \\ \cdots \\ \dfrac{\partial Q}{\partial \beta_n} = -2\sum_{k=1}^{m}[Y_k - Y_k']X_{kn} = 0 \end{cases} \tag{15}$$

The above equation (7) is called the normal system of equations, substitute the formula (5) into the formula (7), and arrange it in matrix form:

$$(X^T X)\beta' = X^T Y \tag{16}$$

among:

$$X^T = \begin{bmatrix} 1 & X_{11} & X_{21} & \ldots & X_{1n} \\ 1 & X_{21} & X_{22} & \ldots & X_{2n} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & X_{m1} & X_{m2} & \ldots & X_{mn} \end{bmatrix}, Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \ldots \\ Y_m \end{bmatrix}, \beta' = \begin{bmatrix} \beta_0' \\ \beta_2' \\ \ldots \\ \beta_m' \end{bmatrix} \tag{17}$$

$(X^T X)$ With the coefficient matrix full rank, the least squares estimate of the regression coefficient is solved:

$$\beta' = (X^T X)^{-1} X^T Y \tag{18}$$

Therefore, the multiple linear regression prediction results can be obtained by taking the least-squares estimate of the final regression coefficient back to equation (1).

4 Empirical analysis

This paper focuses on a 220kV three-phase three-winding transformer, analyzing key factors like raw material prices, labor costs, and monetary policy as independent variables. Data from 2008 to 2018 were input into PASS software to establish a multiple regression model, with an R-squared of 0.903, indicating a strong correlation and good model fit. The specific model parameters are shown in Table 1.

Table 1. The multiple linear regression model parameters.

model	Non-standardized coefficients		Standard coefficient	t	Sig
	B	standard error			
(constant)	749.662	179.758		4.169	.004
prices of raw and semifinished materials	.0019	.027	.308	.662	.531

labour cost	-.002	.002	-.489	-1.140	.293
monetary policy	13.403	11.531	.207	1.163	.284

According to Table 2, the multiple linear regression equations composed of raw material price, labor cost, monetary policy and transformer price are:

$$Y = 0.0019X_1 - 0.002X_2 + 13.403X_3 + 749.663 \tag{19}$$

From the above prediction model, the price of transformers from 2019 to 2022 is predicted, and the results are shown in Table 2 below.

Table 2. Prediction results and errors of multiple linear regression.

project	In 2019,	In 2020,	In 2021,	In 2022,
actual price	875	895	880	841
forecast price	886	891	867	838
fractional error	1.26%	0.45%	1.48%	0.36%

According to Table 2, the prediction error of the model constructed in this paper is controlled at 1.50%, and the prediction effect is good.

The proposed least-squares-based multiple linear regression model has significant potential for real-world applications in power grid enterprises. For instance, the model's outputs can be integrated into budget management software to streamline the pre-construction cost planning process. This integration would enable decision-makers to generate real-time pricing forecasts for transformer equipment, thereby improving resource allocation and minimizing the risks of budget overruns.

Additionally, the model can be adapted to account for different types of transformer equipment by recalibrating regression coefficients based on specific material costs, manufacturing processes, and regional price trends. Such adaptability ensures the model's relevance across diverse operational scenarios, enabling enterprises to respond effectively to varying equipment specifications and market demands. By aligning the model's predictions with practical budgeting and procurement systems, it can serve as a robust tool for enhancing financial planning and investment strategies in power grid construction projects.

5 Conclusion

Equipment material price fluctuations will bring great risk to power grid construction project engineering cost control, therefore, this paper proposes a least squares-multivariate linear regression model for predicting grid transformer equipment prices. Empirical analysis shows a prediction deviation within 1.50%, demonstrating good accuracy. However, the model's reliance on historical data limits its adaptability to unprecedented market shocks and extreme fluctuations. Additionally, the assumption of linear relationships oversimplifies complex interactions. Future research could incorporate advanced machine learning techniques like XGBoost or neural networks to model nonlinear relationships, handle larger datasets, and improve prediction robustness, particularly in volatile markets.

References

1. He Zanli, Lai Siwei, Zhao Jiawei, et al. Analysis of influence factors of blast furnace gas power generation engineering [J]. Metallurgy and Materials, 2023,43 (09): 26-28.
2. Lu Yanchao, Wen Weining, Zhao Biao, et al. Analysis of influencing factors of power grid engineering equipment materials [J]. Electric Power Construction, 2013,34 (04): 74-78.
3. Zeng Yuxia, Wen Yumei, Chen Luyao. Research on the Whole Process Management of Engineering Equipment and Material Price [C] / / Natural Gas Professional Committee of Chinese Petroleum Society. Proceedings of the 33rd National Natural Gas Academic Annual Conference (2023) (06 Comprehensive). Sichuan China you Group Co., Ltd.; Sichuan Chuanang Gas Co., Ltd.; 2023:26. DOI:10.26914/c.cnkihy.2023.070035.
4. He Yilin, Wan Zhengdong, Huang Yan, et al. Research on the price formation factors of large-scale equipment under the Action Plan of Large-scale equipment Update —— Based on the calculation and analysis of UHV primary equipment material cost, research and development expenditure and labor cost [J]. Price theory and Practice, 2024, (07):138-142.
5. Zhang Huiling, Shen Yabo, Wang Shiyi. Correlation analysis between equipment and material information price and production price index [J]. China Electric Power Enterprise Management, 2023, (12): 56-57.
6. Zhang Jigang, Wu Liangzheng, Qiao Huiting, et al. Multi-Step Price Prediction of Main Materials in Power Grid Projects Based on CEEMDAN-GRU [J]. Control Engineering, 2023, 30(11): 2134-2142.
7. Huang Longshan. Research on Power Material Procurement Price Prediction Based on SVM Algorithm [J]. China New Technologies and Products, 2023, (21): 126-129.
8. Papachatzis K. Machine learning-based price prediction for thermal insulation materials: A holistic approach integrating thermophysical, technical, and environmental attributes in the Greek construction market [J]. Energy & Buildings,2024,324114899-114899.
9. D. T G, L. M P, Paul A. Cost and price projections of synthetic active materials for redox flow batteries [J]. Journal of Power Sources, 2021, 499
10. Matyjaszek M, Valverde F G, Krzemień A, et al. Optimizing Predictor Variables in Artificial Neural Networks When Forecasting Raw Material Prices for Energy Production[J]. Energies,2020,13(8):