Quantitative Research on New Energy Vehicle Development Based on Multiple Linear Regression

Jinpeng Ye [1, *, #], Xiaoyan Fang [1, #], Yinglin Huang [2, #], Dongmin Xiao [1, #]

[1] College of mathematics and computer, Shantou University, Shantou, China, 515821

[2] College of Business, Shantou University, Shantou, China, 515821

* Corresponding Author Email: 22jpye1@stu.edu.cn

#These authors contributed equally.

Abstract. Analyzing the factors influencing the development of China's new energy vehicles is beneficial for understanding the progress of this sector and boosting its growth. This study utilizes gray correlation analysis and a multiple linear regression model to examine the relationship between diverse factors and the progress of new energy vehicles. To identify the primary determinants of new energy vehicle sales, data on pertinent indicators were collected. Initially, seven evaluation indicators affecting the sales of new energy vehicles were identified by reading many literatures. Subsequently, gray correlation analysis was employed to calculate the degree of association between these indicators and new energy vehicles. It was discovered that the gray correlation between the number of public charging piles and the sales of new energy vehicles is 0.922, and the corresponding gray correlations of the remaining six indicators are all 0.7-0.8, which means that the correlation between the indicators and the sales of new energy vehicles is obvious, and all of them are used to establish the multiple linear regression model. The results show that the number of public charging piles has the most significant effect, with a regression coefficient of 1×10^{-3}, which indicates that strengthening the charging infrastructure can effectively promote the growth of its sales.

Keywords: Grey Relational Analysis, Multiple Linear Regression, New Energy Vehicle Development, New Energy Vehicle Sales.

1. Introduction

Since 2011, the Chinese government has formulated a series of preferential policies for new energy vehicles [1], which has greatly promoted the development of new energy vehicles. The development of new energy vehicles is affected by many factors [2], and studying the degree of influence of these factors on new energy vehicles is conducive to further promoting the development of the new energy vehicle industry.

Zhou Yuping [3] used Spearman to carry out correlation analysis to calculate the correlation coefficient between the sales volume of new energy vehicles and each influential index, and analyze the correlation between them. Tang Wendi et al [4] used a multiple linear regression model to analyze the significance between residents' willingness to purchase new energy vehicles and each influential factor by calculating the significance P-value of each variable.

However, Spearman's correlation coefficient [5] may not be stable when calculating the results in small samples; while the new energy vehicles are developed in recent years, their annual sales data sample size is small, and the training effect of Spearman's correlation coefficient may not be good; while the gray correlation analysis may be better when analyzing the correlation relationship between the sub-sequence and the parent sequence in small data samples; the multiple linear regression model is simple and easy to analyze the relationship of multi-factors[6] The multiple linear regression model is simple and convenient in analyzing multi-factor relationships , and can accurately calculate the degree of regression fit, and then determine whether it is suitable for use, which is worth borrowing and using. This paper first uses gray correlation to analyze the degree of correlation between the sales of new energy vehicles and various indicators, extracts the indicators with high correlation to train the multiple linear regression model, and uses the regression coefficients to describe the impact of the indicators in a specific way. Compared with the previous model, the gray correlation analysis

model in this paper is not only suitable for small sample data [7], but also can screen out the main indicators for the regression model to improve the effect of the regression model; at the same time, the multiple linear regression model in this paper can quantify the specific impact of the changes in the main indicators on the sales of new energy vehicles, which can provide more specific reference for the government's macro-control.

2. Data and model principle

2.1. Data

2.1.1. Data sources

Analysis of the impact of new energy vehicle sales. Selected data sources are shown in Table 1.

Table 1. Data source.

Data	Website
Source 1	China Statistics Bureau (stats.gov.cn)
Source 2	China Association of Automobile Manufacturers (CAAM) (caam.org.cn)
Source 3	International Energy Agency (IEA) IEA-International Energy Agency
Source 4	China Energy Statistics Yearbook (nbsti.net)

2.1.2. Data processing

Due to the lack of data on per capita energy consumption for 2022, the paper is filled in by taking the average of three spline interpolation and three Elmit interpolation algorithms to ensure data completeness.

Similarly, due to the absence of data on the number of public charging piles from 2011 to 2013, the article estimated these figures based on known information. Specifically, it was known that the number of public charging piles in China increased significantly from 2010 to 2013, rising from 1,122 to 22,528 with an annual compound growth rate of 171.8%. Based on this information, it is estimated that there were approximately 3,049 public charging piles in China in 2011, which increased to 8,288 in 2012, and ultimately reached 22,528 by 2013.

2.2. Grey correlation analysis

The magnitude of the correlation between the factors of two systems that change over time or different objects is called their degree of correlation. Grey correlation analysis is based on the grey correlation degree to analyze the degree of correlation between the factors of the system by comparing the degree of similarity in the geometry of the curves and the geometric relationship of the data series [8].

The calculation steps are as follows:

(1) Step 1: Determine the parent series and the characteristic series.

The comparison sequence is:

$$[X_1'\ X_2'\ \cdots\ X_n'] = \begin{bmatrix} x_{(1)}'(1) & x_{(2)}'(1) & \cdots & x_n'(1) \\ x_{(2)}'(2) & x_2'(2) & \cdots & x_n'(2) \\ \vdots & \vdots & & \vdots \\ x_1'(m) & x_2'(m) & \cdots & x_n'(m) \end{bmatrix} \tag{1}$$

Then the parent sequence is:

$$X_0' = (x_0'(1), x_0'(2), \cdots x_0'(m))^T \tag{2}$$

(2) Step 2: Normalization of indicator data.

$$X'_{ij} = \frac{X_{ij} - \min(X_{1j}, X_{2j}, \cdots X_{nj})}{\max(X_{1j}, X_{nj}, \cdots X_{nj}) - \min(X_{1j}, X_{nj}, \cdots X_{nj})} \tag{3}$$

(3) Step 3: Calculate the correlation coefficient.

$$\gamma(x_0(k), x_i(k)) = \frac{\Delta\min + \rho\Delta\max}{\Delta ik + \rho\Delta\max} \tag{4}$$

$$\Delta\min = \min_i \min_k |x_0(k) - x_i(k)| \tag{5}$$

$$\Delta\max = \max_i \max_k |x_0(k) - x_i(k)| \tag{6}$$

$$\Delta ik = |x_0(k) - x_i(k)| \tag{7}$$

P is the discrimination coefficient, which usually takes 0.5.

(4) Step 4: Define the grey correlation between x_0 and x_1 as

$$y(x_0, x_i) = \frac{1}{n}\sum_{k=1}^{n} y(x_0(k), x_i(k)) \tag{8}$$

The value of correlation ranges from (0, 1), and a larger value indicates a stronger correlation with the parent series.

2.3. Multiple linear regression (MLR)

Multiple linear regression aims to investigate the linear relationship between multiple independent variables and a continuous dependent variable to determine the quantitative link between multiple variables [9]. However, multiple linear regression may have the problems of multicollinearity and heteroscedasticity: multicollinearity describes the linear or near-group linear relationship between the explanatory variables, which may lead to the instability of the regression coefficients; heteroscedasticity refers to the unequal variances of the individual perturbation terms, and the perturbation term with larger variance destabilizes the model to a greater extent. Both of them may reduce the stability and prediction accuracy of the model [10]. Therefore, a rigorous test of multicollinearity and heteroscedasticity is also needed to ensure the stability and prediction accuracy of the model.

(1) Step 1: Training regression equation

$$\mathbf{Y} = \beta_0 + \beta_1\mathbf{X_1} + \beta_2\mathbf{X_2} + \cdots + \beta_7\mathbf{X}_7 + \epsilon \tag{9}$$

(2) Step 2: Multicollinearity test

$$VIF_m = \frac{1}{1 - R^2_{1-k\backslash m}} \tag{10}$$

Where VIF_m refers to the variance inflation factor of the m variable, when VIF >10, it indicates the existence of multicollinearity, which needs to be considered for indicator elimination and correction.

(3) Step 3: White's test - testing for heteroscedasticity

$$\hat{S} - s^2 S_{XX} = \frac{1}{n}\sum\nolimits_{i=1}^{n}(e_i^2 - s^2)x_i x_i' \xrightarrow{p} \mathbf{0_{K\times K}} \tag{11}$$

When P < 0.05, there is heteroscedasticity and robust standard errors need to be considered; when P > 0.05, there is no heteroscedasticity and the regression can be performed directly.

3. Results

3.1. Gray correlation analysis results

According to the grey correlation analysis, the following correlation results can be obtained, as shown in Table 2.

Table 2. Correlation result.

Evaluation item	Relatedness	Rankings
Number of charging stations	0.922	1
Gross domestic product of the secondary industry	0.778	2
Disposable income per capita	0.776	3
Energy consumption per capita	0.767	4
CO_2 emission	0.767	5
Diesel imports	0.75	6
Electricity production grows	0.733	7

From the basic principle of gray correlation analysis, it can be seen that the closer the gray correlation degree is to 1, which means the stronger the correlation with the parent series. In the table, the correlation between the number of charging piles and the sales of new energy vehicles is 0.922, which is a strong correlation; while the correlation between other indicators and the sales of new energy vehicles is between 0.7 and 0.8, which is a strong influence (among them, the Gross domestic product of the secondary industry has the second largest influence), so all of them are trained in multiple linear regression.

From common sense, when the public charging piles and other infrastructure is perfect, fast charging and reasonable charging prices can stimulate consumers' desire to buy new energy vehicles; and with the increase of the gross domestic product of the secondary industry, it has led to the development of infrastructure construction, which also promotes the government's investment and support for the development of new energy vehicle industry. New energy vehicle charging infrastructure continues to improve, for the popularity of new energy vehicles to provide a great help, and increase the support and investment in new energy vehicles, so that enterprises to develop a more energy-efficient, more convenient new energy vehicle products to meet the needs of consumers when the consumer demand for new energy vehicles will promote consumer demand for new energy vehicles, thus driving the growth of new energy vehicle sales.

3.2. Multiple linear regression results

After the collinearity test, the final training results are as Table 3.

Table 3. Training result.

Sales of new energy vehicles	Non-normalization factor B	Normalization factor Beta	P	VIF	R^2	Adjust R^2
Electricity production grows	3.93	0.06	0.53	1.24		
Diesel imports	-0.39	-0.11	0.29	1.29	0.94	0.92
Number of charging stations	1×10^{-3}	0.96	0.000***	1.05		
_Cons	-1.34	-	0.98	-		

The goodness of fit R^2=0.94, the fitting effect is good, and the effect diagram is as Figure 1.

Figure 1. Fitting graph.

Among them, the P value of heteroscedasticity test was 0.2428 > 0.05, indicating that there was no heteroscedasticity.

The following linear regression equation can be obtained:

$$\mathbf{Y} = 3.93\mathbf{X_1} - 0.39\mathbf{X_2} + 10^{-3}\mathbf{X_3} - 1.34 \tag{12}$$

Eventually, it was found that after excluding the indicators with multicollinearity such as the number of public charging piles, the number of charging stations has the largest significant p-value of 0.00, which matches the results of the gray correlation analysis.

The regression coefficient of the Number of charging stations is 1×10^{-3}, the regression coefficient of diesel imports is -0.39, and the regression coefficient of the growth rate of electric power is 3.93. Their β-values are all greater than 0, i.e., they are all positively correlated: for every increase of 1,000 units of number of charging stations, there will be roughly an increase of 1 unit of sales volume (10,000 units); for every increase of 1 unit of diesel imports (10,000 tons), will decrease roughly 3900 new energy vehicle sales; every 1% increase in the growth rate of electricity, will increase roughly 39,300 new energy vehicle sales. The model's goodness-of-fit R^2 is 0.94, which is a good fit and can well present the relationship between the indicators and new energy vehicle sales.

4. Conclusions

This paper focuses on the influence of many factors such as the number of charging piles on new energy vehicles in China. Based on the existing historical data, this paper uses grey correlation analysis and multiple linear regression to carry out in-depth quantitative analysis. However, due to insufficient data and limited training samples, there may be some deviation between the quantitative results and the actual situation. With the continuous accumulation of data in the future, it is believed that the quantified results will be more accurate. In the real world, the development of the new energy-electric vehicle industry is also affected by a variety of factors. Therefore, new energy automobile enterprises should fully consider these factors to better promote and develop new energy vehicles and further promote the sustainable development of China's new energy automobile industry.

References

[1] Liu Cheng. Research on the impact of industrial policy support on the development of new energy vehicles [D]. Jiangxi University of Finance and Economics, 2023.

[2] Tang Lichun, Yan Miao. Main influencing factors of new energy development in China. Market Research, 2015(02):19-23.

[3] Zhou Yuping. Research on the impact of policy incentives and media richness on consumers' new energy vehicle purchases [D]. Nanchang University, 2023.

[4] Tang Wendi, Yu Sitong, Wang Jiahui et al. Research on the influence mechanism of new energy vehicle purchase intention of Shandong province residents. China Collective Economy, 2024(02):54-57.

[5] Yu Qun, Huo Xiaodong, He Jian et al. Trend prediction of power outages in China based on Spearman's correlation coefficient and system inertia. Chinese Journal of Electrical Engineering, 2023, 43(14):5372-5381.

[6] Sun Lezheng, Zhao Mengdi, Du Shaobo et al. Prediction of game difficulty and influencing factors based on multiple linear regression model. Digital Technology and Application, 2023, 41(12):85-87.

[7] Zhang Tianwen. Evaluation of innovation and entrepreneurship ability of college students based on gray correlation analysis model. Journal of Changchun Engineering College (Natural Science Edition), 2023, 24(04):124-128.

[8] Xie Juanjuan. Gray correlation analysis of agricultural industry structure optimization in Gansu Province. Research on Land and Natural Resources, 2024(02):41-44.

[9] Li Hongchao. Measurement and prediction of real estate credit risk of commercial banks--Based on multiple linear regression analysis. Shanghai Real Estate, 2023(11):40-45.

[10] Yao Wang. Multiple linear regression based on multiple covariance correction [D]. Yili Normal University, 2024.