ISSN 2522-3488 Vol. 9, Issue 1: 34-41, DOI: 10.25236/IJNDES.2025.090105

Research on ARIMA Model and Multiple Linear Regression Model of China's New Energy Vehicle Market Development Forecast Research

Wenzhuang Liu [1], Shaohui Han[2], Xiaowei Li[1], Rui Yan[3], Jiarun Cui[1], Jiayi Sun[4]

[1]Department of Architectural, North China University of Science and Technology, Tangshan, China, 063210
[2]Department of Metallurgical and Energy Sources, North China University of Science and Technology, Tangshan, China, 063210
[3]Department of Science, North China University of Science and Technology, Tangshan, China, 063210
[4]Department of Management, Shandong Second Medical University, Weifang, China, 261000

Abstract: *Faced with the global carbon neutrality goal and environmental pollution problems, new energy vehicles (NEVs) have become a key factor in achieving sustainable development due to their emission reduction and energy-saving characteristics. This study addresses the issue of predicting the growth trend of the NEV market in China by constructing a Pearson correlation analysis model to reveal the strong correlation between government R&D investment and NEV production and sales. The study also uses a multiple linear regression model to quantify the impact of different factors on NEV total sales, and an ARIMA model to forecast the development of the NEV market over the next ten years. The study's findings predict that the NEV market in China will show a stable growth trend from 2023 to 2032, with projected sales reaching 1.437,45 million by 2032. These findings provide important decision support for policymakers and industry stakeholders in strategic planning and resource allocation and have significant practical implications for promoting the healthy development of the new energy vehicle industry and responding to the global carbon neutrality goal.*

Keywords: *New Energy Vehicles, Growth Prediction, Pearson Correlation, Multiple Linear Regression, ARIMA Model*

1. Introduction

Vigorously developing new energy vehicles (NEV) has become an effective way to achieve carbon emission reduction and carbon neutrality [1]. China's new energy vehicle (NEV) industry faces great risks of overinvestment and overcapacity. Accurate prediction of China's future new energy vehicle market is significant for the reasonable growth and reasonable scale production of the Chinese government-controlled industry [2].

Lianyi Liu et al. proposed an optimized discrete gray scale power model to fit the nonlinear relationship between the gray information factor and time factor. The prediction results show that by 2025, the sales volume of new energy vehicles in China will reach 8.84 million, accounting for about 24% of the total sales volume of cars, exceeding the industry development target set by the Chinese government [3]. Nele Rietmann et al. have made a long-term forecast of EV inventory in 26 countries/regions on five continents through the logistics growth model and concluded that it will have an important impact on policymakers, marketers, and future research [4]. Renjie Hu et al. identified 25 technical topics using the LDA topic model, analyzed the evolution trend of the topics from the aspects of importance and popularity, and predicted the popularity and development trend of various technical topics in new energy vehicles from 2021 to 2025 by building an ARIMA model [5]. Siying Long et al. used the SARIMA model and exponential smoothing method to fit the sales data of new energy vehicles and used the SARIMA model with a better fitting effect to forecast sales. LASSO regression is used to determine the parameters that have a greater impact on the sales of new energy vehicles, and then K-means cluster analysis and multiple linear regression are used for correlation analysis to provide support for enterprise production [6].

Although various forecasting models such as SARIMA, ARIMA, and grey models have been adopted, these models may lack sufficient adaptability and scalability for the over-investment and overcapacity

ISSN 2522-3488 Vol. 9, Issue 1: 34-41, DOI: 10.25236/IJNDES.2025.090105

problems in China's NEV market. Each model may perform well under certain conditions, but it may not universally apply to all situations. Beyond this, existing research may not provide sufficiently dynamic optimization strategies to adapt to changing market conditions and policy environments. Therefore, it is necessary to optimize capacity planning and policy adjustments in real-time based on current and forecast market demand.By combining Pearson correlation analysis, multiple linear regression model, and ARIMA model, this paper provides a multi-dimensional analytical framework to forecast the development of China's new energy vehicle market. Applying this comprehensive model can capture the market dynamics and influencing factors more comprehensively than a single model, and improve the accuracy and reliability of the forecast.

2. Model

2.1 Correlation analysis

Correlation analysis is carried out for each indicator and development to determine whether it has a linear relationship. The most commonly used statistical measure is correlation, and the most commonly used correlation is Pearson correlation. Pearson correlation coefficient is calculated as follows.

$$\mathrm{r} = \frac{N\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{N\sum {x_i}^2 - (x_i)^2}\sqrt{N\sum {y_i}^2 - (y_i)^2}} \tag{1}$$

where r represents the Pearson correlation coefficient, where x and y represent two variables.

The values of r range from -1 to 1, where -1 means completely negative correlation, 0 means no correlation, and 1 means completely positive correlation. The closer r is to 1 or -1, the larger the absolute value of the correlation coefficient is, and the stronger the correlation is. The closer the correlation coefficient is to 0, the weaker the correlation is. The Pearson correlation coefficient can represent the degree and direction of correlation between random variables, as shown in the following table 1.

Table. 1. Pearson correlation coefficient correlation degree grade table

The size range of correlation coefficient	Degree of correlation	Correlation direction
$-1 < r < -0.5$	Strong correlation	negative
$-0.5 \le r \le 0.3$	Moderate correlation	negative
$-0.3 < r < 0$	Weak correlation	negative
0	uncorrelated	There is no
$0 < \mathrm{r} < 0.3$	Weak correlation	Forward direction
$0.3 \le r \le 0.5$	Moderate correlation	Forward direction
$0.5 < r < 1$	Strong correlation	Forward direction

2.2 The basic fundamental of multiple linear regression model

Multiple regression analysis is used to quantitatively describe the linear dependency between one dependent variable and several independent variables by the regression equation. Multiple linear regression analysis is a statistical method used to evaluate the relationship between a dependent variable and multiple independent variables. In the study of practical problems, the change of the dependent variable is often affected by several important factors, so it is necessary to use two or more influential factors as independent variables to explain the change of the dependent variable, which is multiple regression.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \mu_0 \tag{2}$$

where β_0, β_1, $\beta_2 \cdots$, β_n means the regression coefficient.

Here, both dependent and independent variables are known, and the ultimate goal of multiple linear regression is to find the multiple regression coefficients.

Suppose the regression parameter vector is, the independent variable matrix is, the prediction vector

ISSN 2522-3488 Vol. 9, Issue 1: 34-41, DOI: 10.25236/IJNDES.2025.090105

is obtained by multiplying the regression coefficient the independent variable is, and the actual dependent variable is.

$$\vec{\beta}=\begin{bmatrix}\beta_1\\ \beta_2\\ \cdots\\ \beta_n\end{bmatrix}\quad \mathrm{Y}=\begin{bmatrix}y^{(1)}\\ y^{(2)}\\ \cdots\\ y^{(n)}\end{bmatrix}\quad \overline{\mathrm{Y}}=\begin{bmatrix}\overline{y}^{(1)}\\ \overline{y}^{(2)}\\ \cdots\\ \overline{y}^{(n)}\end{bmatrix}\quad X=\begin{bmatrix}1 & x_1^{(1)} & \cdots & x_n^{(1)}\\ 1 & x_1^{(2)} & \cdots & x_n^{(2)}\\ 1 & \vdots & & \vdots\\ 1 & x_1^{(n)} & \cdots & x_n^{(n)}\end{bmatrix} \tag{3}$$

The optimal solution is the one with the smallest error between the actual value and the predicted value, so an objective function can be constructed.

$$J_\beta=\sum_{i=1}^{n}(\overline{y}^{(i)}-y^{(i)})^2 \tag{4}$$

Expansion equation.

$$\begin{aligned}J_\beta&=(\overline{\mathrm{Y}}-X\vec{\beta})^T(\overline{\mathrm{Y}}-X\vec{\beta})\\&=\overline{\mathrm{Y}}^T\overline{\mathrm{Y}}-\overline{\mathrm{Y}}^TX\vec{\beta}-\vec{\beta}^TX^T\overline{\mathrm{Y}}+\vec{\beta}^TX^TX\vec{\beta}\end{aligned} \tag{5}$$

To differentiate the equation.

$$\frac{\partial J_\beta}{\partial\vec{\beta}}=0-\frac{\partial\overline{\mathrm{Y}}^TX\vec{\beta}}{\partial\vec{\beta}}-\frac{\partial\vec{\beta}^TX^T\overline{\mathrm{Y}}}{\partial\vec{\beta}}+\frac{\partial\vec{\beta}^TX^TX\vec{\beta}}{\partial\vec{\beta}} \tag{6}$$

For vector variables x , the parameters α have the following relationship.

$$\frac{\partial(x^T*\alpha)}{\partial x}=\frac{\partial(\alpha^T*x)}{\partial x}=\alpha \tag{7}$$

It can be concluded that.

$$-\frac{\partial\overline{\mathrm{Y}}^TX\vec{\beta}}{\partial\vec{\beta}}-\frac{\partial\vec{\beta}^TX^T\overline{\mathrm{Y}}}{\partial\vec{\beta}}=-2X^T\overline{\mathrm{Y}} \tag{8}$$

For vector variables x , the parameter A is the following relation.

$$\frac{\partial(x^T*A*x)}{\partial x}=(A+A^T)*x \tag{9}$$

It follows that.

$$\frac{\partial\vec{\beta}^TX^TX\vec{\beta}}{\partial\vec{\beta}}=2X^TX\vec{\beta} \tag{10}$$

From the previous formula, we can deduce.

$$\frac{\partial J_\beta}{\partial\vec{\beta}}=-2X^T\overline{\mathrm{Y}}+2X^TX\vec{\beta} \tag{11}$$

Taking the derivative of 0 gives the solution.

$$X^T\overline{\mathrm{Y}}=X^TX\vec{\beta} \tag{12}$$

$$\vec{\beta}=(X^TX)^{-1}X^T\overline{\mathrm{Y}} \tag{13}$$

ISSN 2522-3488 Vol. 9, Issue 1: 34-41, DOI: 10.25236/IJNDES.2025.090105

2.3 The fundamentals of the ARIMA model

ARIMA (Autoregressive Comprehensive Moving Average) model is a classical time series analysis model used to model and predict time series data. The ARIMA model, which combines the characteristics of autoregressive (AR) and moving average (MA) models, can handle many different types of time series data and is widely used in time series analysis.

The ARIMA (autoregressive comprehensive moving average) model contains three parameters, namely (order of autoregressive analysis), (difference number), and (order of moving average components), which are commonly expressed as three parameters. Assume that the observed value of the study data satisfies the following formula.

$$z_t = \lambda_1 z_{t-1} + \lambda_2 z_{t-2} + \lambda_p z_{t-p} \tag{14}$$

where: $\lambda_t (i = 1,2,\ldots,)$ is the regression function and the number of lagging variables;p is a white noise process, then v_t is the observed value of linear data is an order autoregressive model, which can be expressed as $AR(P)$.

White v_t is represented by the hysteresis operator as follows.

$$v_t = \Lambda(L) z_t = (1 - \lambda_1 L - \lambda_2 L^2 - \cdots - \lambda_p L^p) z_t \tag{15}$$

where $\Lambda(L)$ is an autoregressive operator.

Therefore, the autoregressive operator can be expressed as:

$$\Lambda(L) = (1 - G_1^{-1} L)(1 - G_2^{-1} L) \cdots (1 - G_P^{-1} L) \tag{16}$$

If defined $G_1^{-1}, G_2^{-1}, \cdots, G_P^{-1}$ as the eigenroot of the autoregressive eigenequation. When the characteristic equation $\Lambda(L) = 0$ is satisfied, the IMA (autoregressive comprehensive moving average) model is stable in p order.

If the observed value z_t satisfies the following formula.

$$z_t = v_t + \theta_1 v_{t-1} + \theta_2 v_{t-2} + \cdots + \theta_q v_{t-q} \tag{17}$$

where $\theta_1, \theta_2, \cdots, \theta_q$ is the equation parameter, and v_{t-q} is the white noise corresponding to $t - q$. Observation z_t is the order q moving average model, expressed as $MA(q)$.

Then the autoregressive formula can be transformed into.

$$z_t = \theta(L) v_t = (1 + \theta_1 L + \theta_2 L^2 + \cdots + \theta_q L^q) v_t \tag{18}$$

where: $\theta(L)$ is the moving average operator.

Then the characteristic equation of the moving average operator is as follows:

$$\theta(L) = 1 + \theta_1 L + \theta_2 L^2 + \cdots + \theta_q L^q = 0 \tag{19}$$

Then, the moving average operator changes to:

$$\theta(L) = (1 - H_1^{-1} L)(1 - H_2^{-1} L) \cdots (1 - H_q^{-1} L) \tag{20}$$

where $H_1^{-1}, H_2^{-1}, \cdots H_q^{-1}$ is the feature root of the moving average feature equation.

Therefore, the observed values of the research values are as follows:

ISSN 2522-3488 Vol. 9, Issue 1: 34-41, DOI: 10.25236/IJNDES.2025.090105

$$z_t = \theta(L)^{-1} v_t = \left(\frac{k_1}{1-H_1 L} + \frac{k_2}{1-H_2 L} + \cdots + \frac{k_q}{1-H_q L}\right) v_t \ (k_1, k_2, \cdots, k_q \text{ is constant}) \tag{21}$$

The MA model is invertible at order q only if the eigenequation satisfies $\theta(L) = 0$.

Since the $ARMA$ model consists of a combination of the AR model and the MA model, the expression can be expressed as:

$$z_t = \lambda_1 z_{t-1} + \lambda_2 z_{t-2} + \cdots + \lambda_p z_{t-p} + v_t + \theta_1 v_{t-1} + \theta_2 v_{t-2} + \cdots + \theta_q v_{t-q} \tag{22}$$

In summary, the deformation formula of $ARMA$ can be obtained as follows:

$$\Lambda(L) z_t = \theta(L) v_t \tag{23}$$

3. Results

3.1 Analysis of Correlation results

The correlation coefficients between different indicators are obtained as shown in the following table.1 represents government investment in scientific research, 2 represents infrastructure construction (number of charging piles), 3 represents the average energy density of batteries, 4 represents the number of brand parties producing electric vehicles, 5 represents the industrial synergy rate, 6 represents the total manufacturing volume of electric vehicles, and 7 represents the total sales volume of electric vehicles. To visually present the structure, we drew thermal maps of the correlation coefficients between different indicators, as shown in Figure 1.

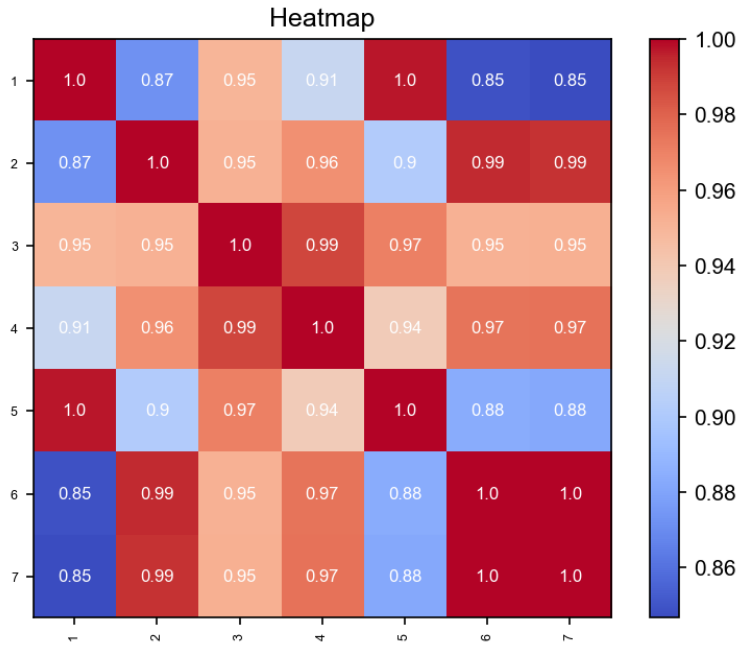

Figure 1. Thermal maps of correlation coefficients between different indexes

From Figure 1, it can be seen that the government's research and development investment has a high correlation with the total manufacturing volume and total sales volume of electric vehicles, with a correlation coefficient approaching 1, which indicates that the government's investment in research and development may have a significant positive impact on the manufacturing and sales of electric vehicles. The correlation between infrastructure construction (the number of charging piles) and the number of electric vehicle brand companies and the degree of industrial synergy is also high, which may mean that the improvement of charging facilities can help attract more brand companies into the market and promote synergy within the industry. The correlation between the average energy density of batteries and the total manufacturing volume and total sales volume of electric vehicles is also high, which indicates

ISSN 2522-3488 Vol. 9, Issue 1: 34-41, DOI: 10.25236/IJNDES.2025.090105

that the progress of battery technology may have a positive impact on the production and sales of electric cars.

3.2 Analysis of multiple linear regression model results

By using Python to perform multiple linear regression analysis, we can obtain correlation coefficients of multiple linear regression, the multiple linear regression equation can be obtained as follows.

$$y = 0.1324x_1 - 0.8506x_2 - 2.3867x_3 - 0.0012x_4 + 0.6193x_5 + 2.2607x_6 \quad (24)$$

where y is the total sales volume of electric vehicles, where x_1 means the investment of government scientific research funds, where x_2 represents the infrastructure construction (the number of charging piles), where x_3 means the average energy density of batteries, where x_4 is the average energy density of batteries, where x_5 is the industrial synergy rate, and where x_6 is the total manufacturing volume of electric vehicles. After regression, R^2 is 0.9273, indicating a good regression effect. The fitting effect is shown in the figure below.

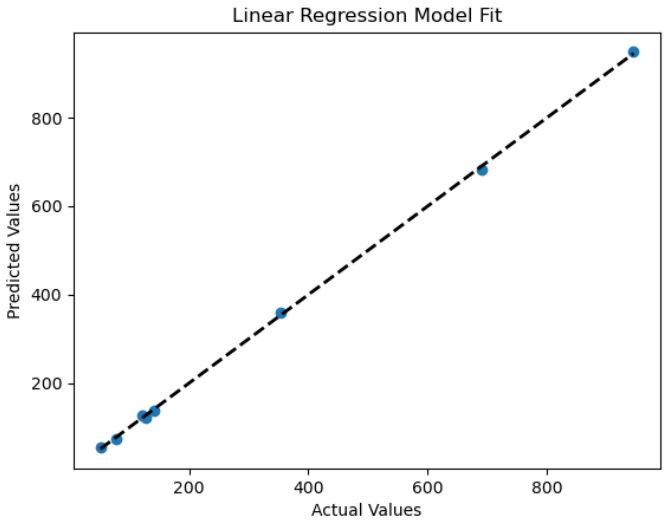

Figure 2. Multiple linear regression fitting diagram

From Figure 2, we can see that the analysis shows that the development of new energy vehicles is more affected by the technological development of the new energy vehicle industry and the level of infrastructure, so we should vigorously develop the technology of new energy vehicles, increase the average energy density of batteries, increase the total manufacturing capacity of electric vehicles, and accelerate the construction of infrastructure, especially charging piles.

3.3 Analysis of ARIMA model results

If the order of regression analysis =1, the difference number=1, and the order of moving average component =2 are taken, then three parameters representing are selected for prediction.

The forecast results are shown in the following table 2.

Table. 2. Forecast results of new energy vehicle development

A given year	Predicted value
2023	1093.156
2024	1195.013
2025	1265.470
2026	1314.208
2027	1347.921
2028	1361.256
2029	1388.984
2030	1405.691
2031	1424.213
2032	1437.454

ISSN 2522-3488 Vol. 9, Issue 1: 34-41, DOI: 10.25236/IJNDES.2025.090105

Taking the logarithm of x and y respectively, we get Figure 3.

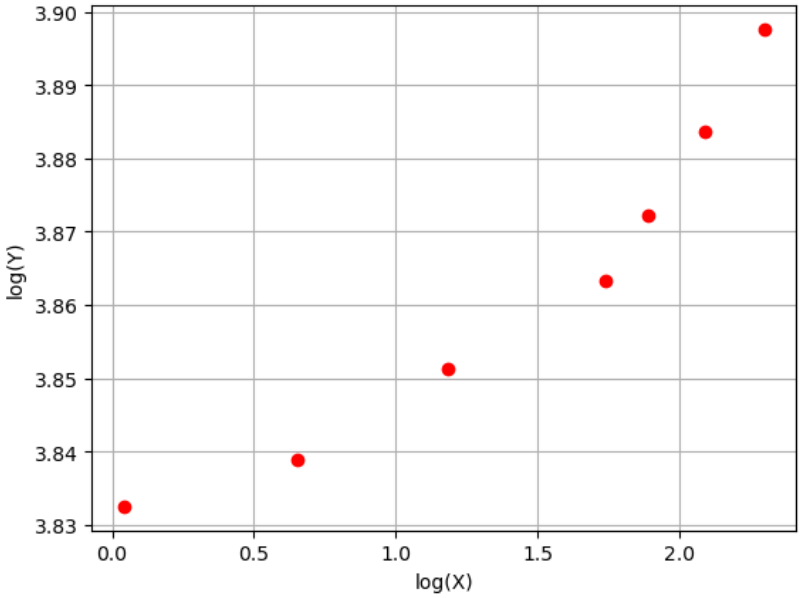

Figure 3 Logarithmic results

As can be seen from Figure 3 and Table 2, the development of new energy vehicles will continue to rise in the next 10 years. From 2023 to 2032, the forecast for new energy vehicles shows a stable growth trend. In 2023, the forecast is 109.156 million vehicles, and by 2032, the number is expected to reach 143.745 million. This indicates that the market for new energy vehicles will maintain a strong growth momentum in the next decade.

4. Conclusions and outlooks

In this study, the prediction of the development trend of new energy vehicles is comprehensively analyzed. We used Pearson correlation analysis, multiple linear regression, and ARIMA models to understand and predict the factors influencing the growth of new energy vehicle development trends. The analysis shows that there is a high correlation between government R&D investment and the manufacturing and sales of new energy vehicles, highlighting the important impact of government support on industry growth, and the importance of technological progress and infrastructure construction in driving the new energy vehicle market. The significance of this study is to provide practical application value for policymakers and industry stakeholders. By identifying the key drivers of NEV growth, our model can inform strategic planning and resource allocation. In addition, the model established in this paper can be generalized to other regions or industries experiencing similar growth dynamics, providing a valuable tool for forecasting and decision-making.

Although this study has made significant progress in predicting the development of new energy vehicles, some limitations need to be acknowledged. First, the model is based on historical data, and future projections may be affected by unforeseen factors such as market conditions, technological advances, and policy changes. Second, ARIMA models, while effective in time series prediction, may not capture all the details of nonlinear relationships that exist in complex systems. Our model can be enhanced by incorporating additional variables such as global economic trends, consumer behavior, and technological breakthroughs. In addition, exploring alternative or hybrid models that combine the strengths of different forecasting techniques can improve the accuracy and robustness of predictions.

References

[1] Li X, Xiao X, and Guo H. A novel grey Bass extended model considering price factors for the demand forecasting of European new energy vehicles[J]. Neural Computing and Applications, 2022, 14(34): 11521-11537.
[2] Zeng B, Li H, Mao C, et al. Modeling, prediction, and analysis of new energy vehicle sales in China using a variable-structure grey model[J]. Expert systems with applications, 2023, 213: 118879.

ISSN 2522-3488 Vol. 9, Issue 1: 34-41, DOI: 10.25236/IJNDES.2025.090105

[3] Liu L, Liu S, Wu L, et al. Forecasting the development trend of new energy vehicles in China by an optimized fractional discrete grey power model[J]. Journal of Cleaner Production, 2022, 372: 133708.
[4] Rietmann N, Hügler B, Lieven T. Forecasting the trajectory of electric vehicle sales and the consequences for worldwide CO2 emissions[J]. Journal of Cleaner Production, 2020, 261: 121038.
[5] Hu R, Ma W, Lin W, et al. Technology topic identification and trend prediction of new energy vehicle using LDA modeling[J]. Complexity, 2022, 2022(1): 9373911
[6] Long S, Liu Q. Research on New Energy Vehicle Sales Forecast and Product Optimization Based on Data Mining[C]//2021 2nd International Conference on Electronics, Communications and Information Technology (CECIT). IEEE, 2021: 1019-1024.