Proceedings of the 3rd International Conference on Mathematical Physics and Computational Simulation
DOI: 10.54254/2753-8818/105/2025.22574

# *Revealing Relevant Factors of Automobile Emissions Based on Linear Regression*

**Yuantian Liu**

*Institute of Statistics, Southwestern University of Finance and Econimics, Chengdu, China*
*42356001@smail.swufe.edu.cn*

***Abstract:*** Global warming has emerged as a critical global concern, with the control of greenhouse gas emissions becoming a paramount priority for nations worldwide. Carbon dioxide ($CO_2$), a significant greenhouse gas, comprises a substantial portion of vehicle exhaust emissions. To effectively mitigate $CO_2$ emissions from automobiles, it is imperative to identify and analyze the key determinants influencing these emissions. In this paper, the collected data were fitted into a model through multiple linear regression in R. The variance inflation factor (VIF) detection method was used to detect multicollinearity, and the stepwise regression method was employed to eliminate multicollinearity in the model to study the related factors of $CO_2$ emissions from automobiles. The findings indicate that engine size, number of cylinders, and combined fuel consumption are primary factors affecting $CO_2$ emissions from vehicles. Among them, the combined fuel consumption is the most significant influencing factor. These results offer valuable insights for automotive engineers and researchers, guiding efforts to enhance vehicle design and reduce $CO_2$ emissions.

***Keywords:*** Carbon Dioxide Emission, Multiple Linear Regression, Multicolinearity, Stepwise Regression.

## 1. Introduction

Greenhouse gases (such as $CO_2$, $CH_4$, $N_2O$, etc.) can absorb the long-wave infrared radiation reflected from the Earth's surface, preventing heat from escaping into space and creating a warming effect similar to that of a greenhouse. The natural greenhouse effect maintains a suitable temperature on Earth, but human activities have significantly increased the concentration of greenhouse gases, leading to excessive heat retention. Climate warming will result in an intensification of extreme weather events, accelerated glacial melting and rising sea levels, reduced agricultural productivity leading to potential food crises, and even the collapse of ecosystems. Ultimately, these changes pose a significant threat to human survival and well-being [1]. In the long run, developing clean energy is the fundamental solution to greenhouse gas emissions. However, in the short term, human beings still rely on fossil energy for production activities. Therefore, reducing the emissions of greenhouse gases such as carbon dioxide is a preoccupation at present.

With the rapid development of the automotive manufacturing industry, the number of cars in use has increased sharply. While cars have brought convenience to human life, they have also had a negative impact on the natural environment, the most significant of which is exhaust emissions [2]. In previous studies, most research focused on harmful gases in vehicle exhaust, such as carbon monoxide, hydrocarbons, nitrogen oxides, sulfur dioxide and particulate matter, while neglecting the

DOI: 10.54254/2753-8818/105/2025.22574

greenhouse gas carbon dioxide, which is not a harmful gas [3]. According to statistics, carbon dioxide accounts for over 99 percent of the carbon emissions from vehicle exhaust. Reducing carbon dioxide emissions from vehicle exhaust is of great significance in controlling greenhouse gas emissions in today's era of widespread car usage. To reduce carbon dioxide emissions from automobiles, it is necessary to identify the factors in cars that affect carbon dioxide emissions so as to make targeted improvements [4]. Therefore, this article will explore the relevant factors of carbon dioxide emissions from automobiles.

This paper first introduces the relevant concepts of multiple linear regression and multicollinearity, as well as the solutions to multicollinearity. After fitting the collected data to the original regression model, the VIF method is used to detect multicollinearity. Then, the stepwise regression method is employed to correct it and obtain the final model. Finally, hypothesis testing is conducted on the final model. At the same time, the relevant factors influencing the carbon dioxide emissions of automobiles are obtained.

## 2. Method and theory

### 2.1. Multiple linear regression

The statistical relationship between variables of objective things is the main research object of regression analysis. It is a statistical method that relies on a large number of experiments and observations of objective things to find the statistical regularity hidden in those seemingly uncertain phenomena. Multiple linear regression is a core method in statistics used to model the linear relationship between multiple independent variables and one dependent variable. After obtaining the multiple linear regression model, it is also necessary to conduct statistical tests on the regression model, including the significance test of the regression equation and the multicollinearity test of the explanatory variables, among which multicollinearity is the focus of this article.

Suppose that there are $n$ observation samples, each sample containing $k$ independent variables $X_1$,$X_2$,... ...,$X_k$ and one dependent variable $Y$. The model form is:

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \epsilon_i (\mathrm{i} = 1,2,\cdots,n), \quad (1)$$

where $\beta_0, \beta_1, \cdots, \beta_k$ are $k+1$ unknown parameters, $\beta_0$ are regression constants, $\beta_1, \beta_2, \cdots, \beta_k$ are regression coefficients. $Y$ is called the explained variable, while $x_1, x_2, \cdots, x_k$ are $k$ general variables that can be accurately measured and controlled, and are called the explained variables.

The multiple linear regression model contains four basic assumptions, namely, that each explanatory variable is not correlated with one another, the random error term is a random variable with an average or expected value of zero, the explanatory variables are not correlated with the random disturbance term, and the random disturbance term follows a normal distribution.

### 2.2. Multicollinearity and its discriminant method and influence

The sufficient and necessary condition to establish a multiple linear regression model is that the rank $r(X)$ of the design matrix $X = k+1$, which means that the column vectors of the sample matrix are linearly independent. However, in the actual production problem, when considering multiple influencing factors, due to the complexity of things, there is a certain correlation between most of the factors. Generally, when the correlation of independent variables is weak, it can be considered to meet the modeling requirements of regression models, but when the correlation of independent variables is strong, the modeling conditions of regression models are violated, and the model is said to have multicollinearity problems [5].

In the analysis of practical problems, it is more common that a linear relationship approximately holds, that is, there exists a set of constant $c_0, c_1, c_2, \cdots, c_k$ not all zero, such that:

DOI: 10.54254/2753-8818/105/2025.22574

$$c_0 + c_1 x_{i1} + c_2 x_{i2} + \cdots c_k x_{ik} \approx 0, i = 1,2,\cdots,n \tag{2}$$

When the independent variables $x_1, x_2, \cdots, x_k$have the relationship as shown in equation (2), it is said that there exists multicollinearity among the independent variables $x_1, x_2, \cdots, x_k$.

This article will introduce two currently more mainstream methods for discriminating multicollinearity, which are variance inflation factor (VIF) and eigenvalue diagnosis of multicollinearity [6]. The VIF diagnoses multicollinearity by measuring the impact of the correlation among independent variables on the variance of regression coefficients. The specific formula is:

$$VIF_i = \frac{1}{1 - R_i^2} \tag{3}$$

Among them, $R_i$ is the coefficient of determination when the $i$th independent variable is taken as the dependent variable and regressed with other independent variables. VIF greater than 10 indicates the presence of severe multicollinearity.

The eigenvalue diagnostic method is a multicollinearity discrimination method based on the eigenvalues and condition indices of the independent variable matrix. First, perform principal component analysis on the standardized independent variable matrix and calculate the eigenvalues. Then, calculate the condition index, which is defined as the ratio of the square root of the largest eigenvalue to that of the eigenvalue.

$$k_i = \sqrt{\frac{\lambda_m}{\lambda_i}}, i = 0,1,2,\cdots,k \tag{4}$$

Among them, $k_i$ is referred to as the condition number of the characteristic root $\lambda_i$. The determination rule is as follows: when $0 < k < 10$, $X$ has no collinearity problem; when $10 < k \leq 100$, $X$ may have a relatively strong collinearity problem; when $k \geq 100$, $X$ has a serious multicollinearity problem.

The variance inflation factor method can visually reflect the degree of collinearity between a single variable and other variables. Moreover, its calculation is relatively simple, and most mainstream statistical software supports direct calculation. At the same time, this method can pinpoint specific variables, facilitating subsequent processing. In comparison, although the eigenvalue diagnostic method has stronger capabilities in global analysis and complex multicollinearity detection, it has cumbersome calculation steps and lacks a direct variable location function. Therefore, this paper will adopt the variance inflation factor method to detect multicollinearity. The existence of multicollinearity will undermine the stability of parameter estimation. For instance, in the case of binary regression, the variance of the estimator is positively correlated with the correlation of the independent variable.

The regression equation of the independent variables $x_1$ and $x_2$ and the dependent variable $y$ is:

$$y = \beta_1 x_1 + \beta_2 x_2 \tag{5}$$

Let $T_{11} = \sum_{i=1}^{n} x_{i1}^2$, $T_{12} = \sum_{i=1}^{n} x_{i1} x_{i2}$, $T_{22} = \sum_{i=1}^{n} x_{i2}^2$, and the correlation coefficient between $x_1$and $x_2$ is $r_{12} = \frac{T_{12}}{\sqrt{T_{11}}\sqrt{T_{22}}}$. After the calculation, it can be obtained that

$$VAR(\beta_1) = \frac{\sigma^2}{(1 - r_{12}^2)T_{11}} \tag{6}$$

$$VAR(\beta_2) = \frac{\sigma^2}{(1 - r_{12}^2)T_{22}} \tag{7}$$

DOI: 10.54254/2753-8818/105/2025.22574

From the above two formulas, it can be seen that when the correlation coefficient between the independent variables is larger, the variance of the parameter estimates increases, which significantly undermines the practical value of the regression model.

## 2.3. Solutions to multicollinearity

The mainstream solutions to multicollinearity include but are not limited to the following three: removing highly collinear variables, combing related variables, and stepwise regression.

The method of removing highly collinear variables refers to the process of detecting and eliminating redundant variables that are highly correlated with other variables in order to reduce multicollinearity. The detection index is usually VIF. The potential weakness of this method is that it might delete important information [7,8].

The method of combining related variables refers to the construction of comprehensive indicators by linearly combining or merging highly correlated indicators to reduce the number of variables and weaken multicollinearity. If $x_1$ and $x_2$ are highly correlated, then construct $Z$ such that:

$$Z = m_1 x_1 + m_2 x_2 \tag{8}$$

In the formula, $m_1$ and $m_2$ represent the weights of each variable, which are usually determined by the significance of each variable, empirical rules, and statistical methods. Stepwise regression is a variable selection technique that combines forward selection and backward elimination strategies to dynamically adjust the independent variables in the model through statistical tests [9]. The main process is as follows.

Model initialization refers to that the initial model is an empty model that only contains the intercept term:

$$Y = \beta_0 \tag{9}$$

The variable introduction refers to that each time, select the variable from those not yet included in the model that has the greatest explanatory contribution to the dependent variable and is statistically significant. The statistical test for variable introduction is to calculate the partial F-statistic.

$$F_{enter} = \frac{SSE_{reduced} - SSE_{full}}{df_{full} - df_{reduced}} \div \frac{SSE_{full}}{n - p_{full} - 1} \tag{10}$$

SSE represent Sum of Squared Errors, df represent Degree of Freedom, n represent Sample Size. If $F_{enter} > F_{critical}$, then introduce the variable.

Variable elimination refers to eliminate the variables that are not significant after introducing new variables. For the selected variable $X_j$, calculate its t-statistic.

$$t_j = \hat{\beta}_j / SE(\hat{\beta}_j) \tag{11}$$

If $|t_j| < t_{critical}$, then eliminate the selected variable $X_j$. Iterative loop refers to that continue this process iteratively until it is no longer possible to introduce any new variables, and all remaining variables in the model are significant, with none requiring removal.

DOI: 10.54254/2753-8818/105/2025.22574

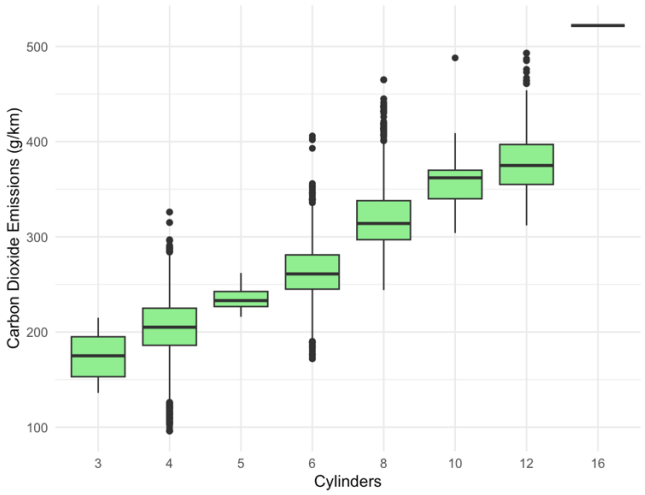


Figure 1: Distribution of $CO_2$ emissions by cylinder count

## 3. Results and applications

### 3.1. Data presentation

According to the survey, the potential factors related to the carbon dioxide emissions of automobiles are engine size, number of cylinders, urban road fuel consumption, highway road fuel consumption and combined fuel consumption. This article will explore the relevant factors of automotive carbon dioxide emissions through 7,385 sets of data containing the aforementioned variables.

Through R, as shown in Figure 1, it can be visually observed that there is a positive correlation between the number of cylinders and carbon dioxide emissions.

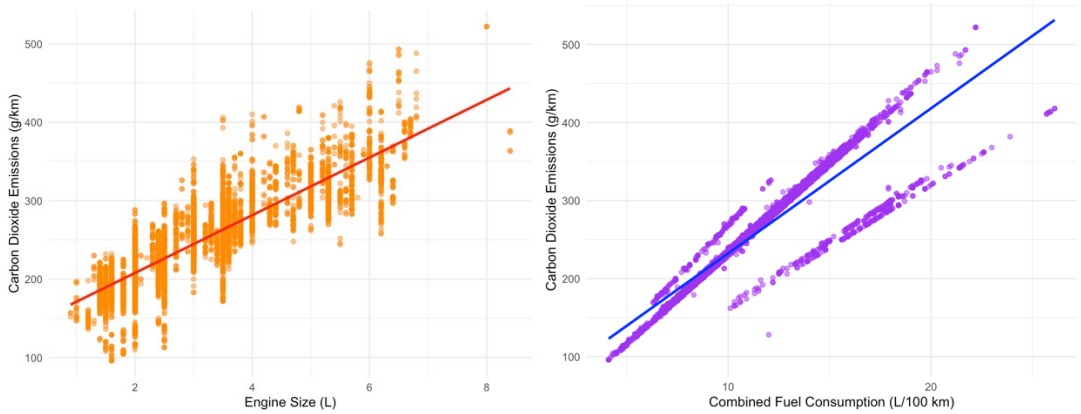


Figure 2: Engine size vs. $CO_2$ emissions and fuel consumption vs. $CO_2$ emissions

As shown in Figure 2, in each data sets, the carbon dioxide emissions corresponding to the engine size are distributed on both sides of the regression line, demonstrating a positive correlation between engine size and carbon dioxide emissions. The graphical representation indicates that the carbon dioxide emissions associated with the combined fuel consumption in each data set are dispersed around the regression line, thereby confirming a positive correlation between fuel consumption and $CO_2$ emissions. The initial regression model was obtained by fitting through the lm function in R [10]. It is found that the $CO_2$ emissions is

$$CO_2 \text{ Emissions} = 50.8179 + 5.5090 \cdot \text{Engine Size} + 6.5421 \cdot \text{Cylinders}$$

DOI: 10.54254/2753-8818/105/2025.22574

$$+1.2212 \cdot \text{Fuel consump(city)} + 1.4350 \cdot \text{Fuel consump(highway)} + 10.6892 \cdot \text{Fuel consump(combined)} \quad (12)$$

## 3.2. Detection of multicollinearity

As shown in the Figure 3, the heat map shows that there may be correlations between engine size and the number of cylinders, as well as between city fuel consumption and highway fuel consumption, which leads to strong multicollinearity. Find highly collinear variables by calculating VIF in R.

Table 1: VIF of each variable

| Variable | VIF |
|---|---|
| Engine Size | 8.528664 |
| Cylinders | 7.572818 |
| Fuel Consumption(city) | 2059.8866997 |
| Fuel Consumption(highway) | 556.043118 |
| Fuel Consumption(combined) | 4625.521279 |

If the VIF exceeds 10, it indicates the presence of significant multicollinearity. Consequently, city fuel consumption, highway fuel consumption, and combined fuel consumption are identified as highly collinear variables, whereas the number of cylinders and engine size exhibit relatively lower correlations, as shown in Table 1.

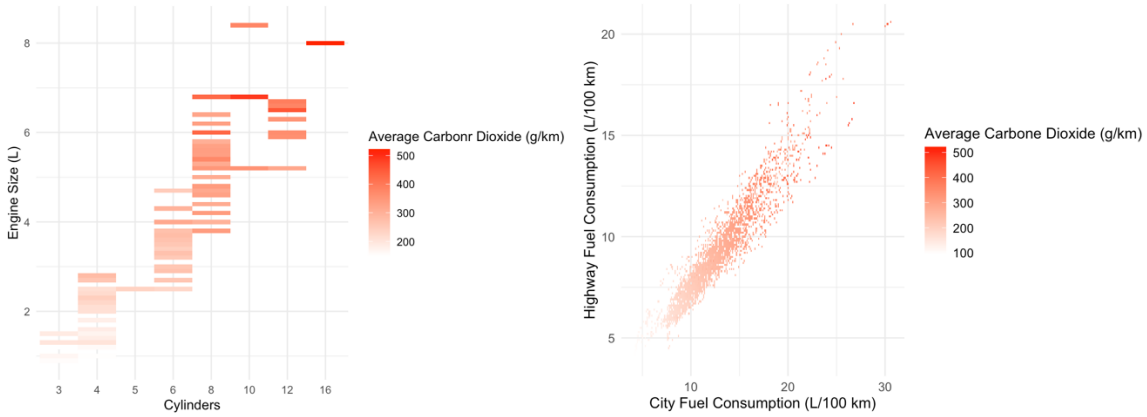


Figure 3: Heatmap of engine size vs. cylinders vs. $CO_2$ emissions and city vs. highway fuel consumption vs. $CO_2$ emissions

## 3.3. Correction of multicollinearity

In terms of carbon dioxide emissions from automobiles, there is a lack of expert solutions for integrating urban fuel consumption and highway fuel consumption as variables, so the combined variable method cannot be used to eliminate multicollinearity. When dealing with redundant variables, to ensure that the model still has a high explanatory power after reducing collinearity, which means achieving a balance between explanatory power and complexity, the stepwise regression method is used instead of the variable removal method. An improved model was obtained through stepwise regression using R.

$$CO_2 \text{ Emssions} = 51.1704 + 5.4971 \cdot \text{Engine size} + 6.4875 \cdot \text{Cylinders} + 13.2688 \cdot \text{Fuel consumption(combined)} \quad (13)$$

DOI: 10.54254/2753-8818/105/2025.22574

Through stepwise regression, the two redundant variables of city fuel consumption and highway fuel consumption were removed.

### 3.4. Model validation and evaluation

The validation and evaluation of the multiple linear regression model begin by examining the fundamental assumptions of linear regression, specifically linearity, normality, and homoscedasticity.

Verify and visualize whether the model conforms to the basic assumptions of the linear regression model through R. The residuals vs. fitted value plot is used to check whether the variance of the residuals is constant and whether there is a nonlinear relationship in the model. When the residuals are randomly distributed around the horizontal line $y = 0$ without any obvious pattern, it indicates that the model conforms to the assumptions of linearity and homoscedasticity.

Both the quantities-quantities plot and the residual histogram are used to test the normality of residuals. When the points in the quantities-quantities plot are roughly distributed along the diagonal line and the residual histogram is close to a bell-shaped curve, it indicates that the model conforms to the normality assumption. As shown in the Figure 4, it can be seen from the results that all the assumptions of the linear regression model hold true.

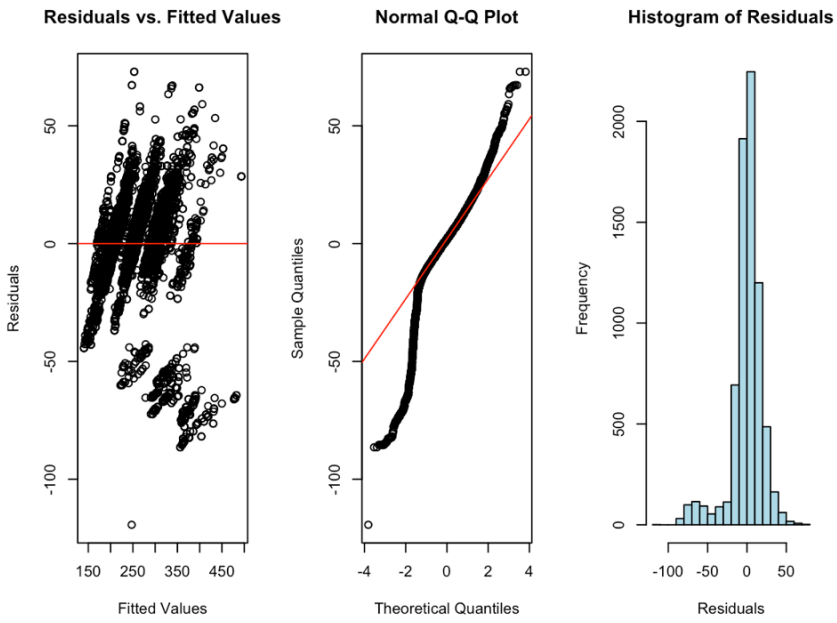


Figure 4: Residual analysis

As shown in the Figure 5, the comparison chart of expected values and actual values shows that the expected values can reflect the changing trend of the actual values. After calculation and modification, the residual standard error of the regression model is 0.8792, close to 1, which also indicates a good fitting effect of the model.

DOI: 10.54254/2753-8818/105/2025.22574

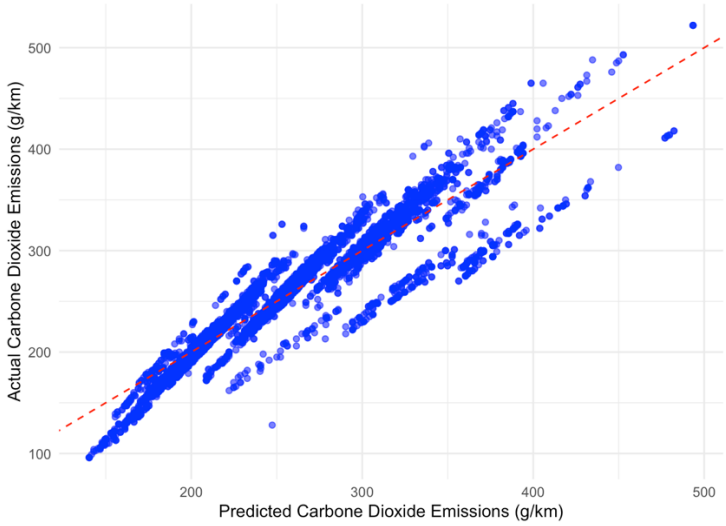


Figure 5: Predicted vs. actual $CO_2$ emissions

## 4. Conclution

This article introduces the application of multiple linear regression in multivariate fitting models, as well as the impact of possible multicollinearity in the model and its identification and correction. Through the fitting and correction of existing data on carbon dioxide emissions from automobiles, it is concluded that engine size, cylinder number and combined fuel consumption are significant influencing factors of carbon dioxide emissions. Moreover, since the coefficient of comprehensive fuel consumption is greater than that of the other variables, it can be known that combined fuel consumption plays a dominant role in the impact on carbon dioxide emissions. To reduce carbon dioxide emissions from automobiles, car manufacturers can focus their research on engine size, the number of cylinders and combined fuel consumption.

In future research, more comprehensive data can be collected to explore whether more factors influence the carbon dioxide emissions of automobiles. Additionally, based on this article, it can be investigated which factors have lower improvement costs to achieve more efficient improvements, because identifying low-cost, high-impact interventions can guide policymakers and manufacturers in making informed decisions that lead to more efficient improvements in emission reduction. At the same time, other methods for addressing multicollinearity can be explored to better eliminate collinearity while maintaining the model's explanatory power. Techniques such as principal component analysis, ridge regression, or even machine learning algorithms might offer promising solutions.

## References

[1] *Liu Honglai. (2008). Global Climate Change and $CO_2$ Emission Reduction. The 18th Branch Conference of the Chinese Association for Science and Technology Annual Conference: $CO_2$ Emission Reduction, Green Utilization and Development Symposium.*

[2] *SHI Xianghui. (2010). A Brief Analysis of Harm and Control on Automotive Exhasut Emission.Automoblie Parts, (07),84-86.*

[3] *Chen Ruilang & Qiao Yongping. (2012).Carbon Dioxide Gas from Automobile Exhaust Emission.Science & Technology Information, (26),99+101.*

[4] *Wang Wei & Jing Sutong. (2009). Discussion on Carbon Dioxide Emissions and Control Measures of Light Vehicles. Journal of Traffic Energy Conservation and Environmental Protection,(01),32-35.*

[5] *Wang Yao. (2023). Multiple Linear Regression Based on Multiple Cointegration Correction (Master's Degree Thesis,Yili Normal University).*

DOI: 10.54254/2753-8818/105/2025.22574

[6] *Ma Xiong wei. (2008). Diagnosis and Empirical Analysis on Multicollinearity in Linear Regression Model.Journal of Huazhong Agriculture University(Social Science Edition),(02),78-81+85.*

[7] *Lin Shilian (2016). Comparison and application of correction method of multicollinearity (Master's Degree Thesis, Guangdong University of Finance & Economics).*

[8] *Zhang Fenglian. (2010). The Discussion on Solutions of Multicollinearity in Multilinear Regression Models (Master's Degree Thesis, South China University of Technology).*

[9] *Yang You & Li Xiao hong. (2006). Instance Analysis of Multi-Collinearity by Stepwise Regression.Journal of Chongqing Three Gorges University,(03),39-41.*

[10] *XIE Chunyu & WAN Wenjun. (2022). Analysis on Influnec Factors of Stock Price in Secondary Market—Multiple Linear Regression Based on R.China Academic Journal,(15),99-102.*