Research on the Influencing Factors of Automobile Price Based on Multiple Linear Regression Model

Caomengyu Xue *

School of Mathematics and Statistics, Hubei University of Education, Wuhan, China

* Corresponding author: Xuecmy@hue.edu.cn

Abstract. With the rapid development of the economy and technology, the living standard of the people has been continuously improved. The consumption level of residents has also been increasing. Because automobiles have generally become the major means of transportation for modern people, the price of cars has steadily caught people's attention. Until now, most individuals are unaware of the relevant parameters and indicators within the car. As a result, many consumers will be confused when selecting an automobile. This paper aims to address any issues potential car buyers could have. The research uses SPSS software to establish a multiple linear regression model of the factors affecting car prices based on the data set on Kaggle. The research methods of this paper are as follows: to test the significance of the results of the model, including the estimation and test of the regression parameters and equation; Diagnosis and treatment of multicollinearity and heteroscedasticity of the model results, and make certain statistical analysis; Eliminate the inconspicuous factors and factors that don't conform to economic significance. The research indicates that engine location, curb weight, number of cylinders, and drive wheels are examined as the significant factors that affect the car price. The research results of this paper provide a reference basis for consumers when purchasing cars, and help consumers to purchase the desired car products.

Keywords: automobile price, multiple linear regression, heteroscedasticity, multicollinearity.

1. Introduction

1.1. Background

With the booming of China's automobile market since the reform and opening up, automobile models and configurations have gradually become diversified in order to win the favor of consumers [1]. According to the statistics released by the China Association of Automobile Manufacturers, China's automobile production and sales increased steadily in 2022, with 27.021 million and 26.864 million vehicles completed respectively. The year-on-year growth was 3.4% and 2.1% namely. The main economic indicators of automobile production and sales continue to improve, showing strong development resilience, and playing an important role in stabilizing industrial economic growth [2]. The car is no longer the exclusive product of the rich but popularizes by every household. Nevertheless, because several consumers only know the brand and grade of the car and don't know the majority of the indicators that affect the car, they are unable to select the car that meets their desired price [3]. There is a large number of published studies that fully conclude the important factors affecting automobile pricing and propose countermeasures and suggestions for the automobile pricing mechanism [4].

1.2. Related research

A large and growing body of literature has investigated the relationship between automobile prices and their various influencing factors.

Cen Jiao, applies multiple linear regression, random forest, and fully connected neural network algorithms to study and analyze the automobile price data. The random forest algorithm reveals some advantages. The results indicate that the price of an automobile is mainly affected by the maximum power, maximum torque, maximum horsepower, safety performance, and other factors [3].

Xiaonan Li starts from the development status of China's automobile market and automobile consumer market. The research indicates that the automobile price in China is affected by five aspects: automobile power and handling, appearance and interior, automobile economical characteristic index, safety, brand, and reputation [4].

Chunjie Yang, from the perspective of consumers, classifies the characteristics of cars from five aspects, constructs three kinds of car price models: linear, semi-logarithmic, and logarithmic, and estimates and tests the models through multiple regression. The research indicates that these three models have good explanatory power, and the semi-logarithmic model is more suitable. The researcher also divides the car into low, middle, and high grades to build the car characteristic price model respectively, and pointed out that different characteristic attributes have different influences on the car price of different grades. The results demonstrate that low-grade cars pay attention to the economy and practicality of cars, middle-grade cars pay attention to safety, and high-grade cars pay attention to operability [5].

Ceyhun Ozgur, Zachariah Hughes, Grace Rogers, and Sufia Parveen utilize the method of multiple linear regression to investigate how seven factors they choose initially affect the price of the cars. After analysis, they examine that mileage and liters are the most useful metrics for estimating the price of a car [6].

Lujia Zhang selects the macro-level and industry-level indicators and establishes a multiple linear regression model between China's automobile price fluctuations and explanatory variables. The empirical results verify that China's automobile price changes have a positive correlation with the domestic steel price composite index, M2 ending balance, and the price index of means of production. They have a negative correlation with the 3-5 year loan interest rate and the exchange rate change of RMB against the US dollar [7].

1.3. Objection

This study aims to establish an accurate model to investigate the differences between various factors influencing car prices, assist consumers in referencing and purchasing, assist relevant management departments in regulating car prices, and standardize the car market [5]. Based on the relevant data of the specifications and insurance risk levels of imported cars and trucks on Kaggle in 1985 [8], this paper constructs a multiple linear regression model to estimate the dependent variables with the optimal combination of multiple independent variables. This study first conducts data processing to initially screen the independent variables for least squares estimation, then exploits stepwise regression to select the independent variables to eliminate the impact of multicollinearity, afterward performs residual analysis and heteroscedasticity diagnosis to test the regression results, and finally employs multivariate weighted least squares regression to deal with the heteroscedasticity problem, improving the model results [4].

2. Methodology

2.1. Source of data

The project utilizes the cross-sectional data set from Kaggle that contains 5330 loans. The data set includes 26 parameters which are either numerical or categorical, covering 206 representative models on the market and delegate products of 22 manufacturers. These 26 parameters are as follows: symboling: divided into - 2, - 1, 0, 1, 2, 3. The lower the risk level, the safer the vehicle risk is compared with the price level; Normalized losses: this column of data is the result of the standardization of annual insurance losses of vehicles. The data range is 65-256; Make: 22 manufacturers including Audi, Volvo, Toyota, Chevrolet, etc; Fuel type: includes gas and diesel; Aspiration: includes turbo and SuperCharge; Number of tools: two and four; Body style: including convertible, hatchback, sedan, and wagon; drive wheels: front, rear and four-wheel drives; Engine location: front engine and rear engine; Engine type: includes six types; Number of cylinders: there are seven types: 2, 3, 4, 5, 6, 8, 12; Fuel system: includes multi-point fuel injection system and 4

cylinder carburetor, etc; wheelbase, length, width, height, curb weight, engine size, bore, stroke, compression ratio, horsepower, peak-rpm,city-mpg,highway-mpg, price.

2.2. Data processing

Data processing is mainly used to check whether there are missing values and abnormal values in the data. First of all, this paper carries out descriptive statistical analysis on the entire sample data, finds out the mean, median, standard deviation, etc. of the data, and judges whether the average value of each sample point is beyond the reasonable range [9]. Secondly, this paper uses the box diagram to identify the abnormal value. The judgment standard is to calculate the minimum and maximum estimated values in the data. If the data exceeds this range, it may be an abnormal value [9].

2.3. Correlation analysis

For the correlation analysis of quantitative variables, Pearson correlation is used in this paper. If the correlation coefficient is greater than 0.5, it means that the two features have a strong correlation and their practical significance is similar. Only one special diagnosis can be retained or the two features can be combined.

2.4. Models

Regression analysis in statistical analysis methods is often applied to study the correlation between two or more variables. Because in practical problems, the price of cars is often affected by multiple variables, this paper establishes a multiple linear regression model to analyze the correlation between independent variables and dependent variables.

Assume that the explained variable Y has a linear relationship with multiple explanatory variables $X_1,X_2,...,X_3$, which is a multivariate function of explanatory variables, that is,

$$Y=\beta_0+\beta_1X_1+...+\beta_kX_k+\mu. \quad (1)$$

The partial regression coefficients $\beta_i(i=1,2,..., k)$ are k unknown parameters, β_0 is a constant term, and μ is a random error term. For n observations, the equation form is

$$Y_i=\beta_0+\beta_1X_{1i}+...+\beta_kX_{ki}+\mu_i. \quad (2)$$

The partial regression coefficients $\beta_i(i=1,2,..., k)$ are unknown but can be estimated using the sample observations $(X_{1i},X_{2i},...,X_{ki},Y_i)$. If the calculated parameter estimate is $\widehat{\beta_0},\widehat{\beta_1},...,\widehat{\beta_k}$, it can replace the position parameter in the population regression equation with the parameter estimate $\beta_0,\beta_1,...,\beta_k$. Then the multiple linear regression equation is

$$\widehat{Y_i}=\widehat{\beta_0}+\widehat{\beta_1}X_{1i}+...+\widehat{\beta_k}X_{ki}. \quad (3)$$

$\widehat{\beta_i}(i=1,2,...,k)$ is the parameter estimation value, and $\widehat{Y_i}(i=1,2,..., k)$ is the regression value or fitting value of Y_i.

2.5. Experimental steps

Step1. Data processing: to check and process dirty data in the data set[9].

Step2. Correlation analysis: to find out the degree of correlation between each quantitative variable that affects the price and the price.

Step3. Variable selection: to preliminarily determine the independent variables that affect the price.

Step4. Establishment and verification of model: to establish multiple linear regression model, and estimate and test the regression parameters and regression equations of the model.

Step5. Residual analysis: to test the normality, randomness and equivariance of the residual sequence.

Step6. Heteroscedasticity diagnosis: to test whether the model has heteroscedasticity and find out the variables with outliers.

Step7. Model optimization: to correct the influence of heteroscedasticity on the model by using multivariate weighted least squares estimation, carry out statistical analysis on the regression results, and check whether the heteroscedasticity problem has been solved. If the heteroscedasticity problem is solved, the regression result will be taken as the final result; Otherwise, adjust the weight category and the weight series to correct again.

3. Results And Discussion

3.1. Data Processing

First of all, 49 missing values are discovered during the descriptive statistical analysis of this data set. The descriptive statistical analysis of automobile price is shown in Figure 1. Specifically, Mean=13276.71057, Median=10295, Standard deviation=7988.852332.

Statistics		
price		
N	Valid	205
	Missing	8
Mean		13276.71057
Median		10295.00000
Mode		5572.000[a]
Std. Deviation		7988.852332
Variance		63821761.578
Minimum		5118.000
Maximum		45400.000
Sum		2721725.667
Percentiles	10	6619.40000
	20	7358.20000
	25	7781.50000
	30	8010.20000
	40	9007.40000
	50	10295.00000
	60	12553.40000
	70	15524.00000
	75	16509.00000
	80	17625.20000
	90	23125.00000
a. Multiple modes exist. The smallest value is shown		

Figure 1. Descriptive Statistics for Automobile Price (Photo Credit: Original)

Next, Because the unit dimension of each attribute is inconsistent, this paper first normalizes the data, and then draws the box diagram of quantitative indicators, as shown in Figure 2. Then replace it with the maximum or minimum value in the normal range [8].

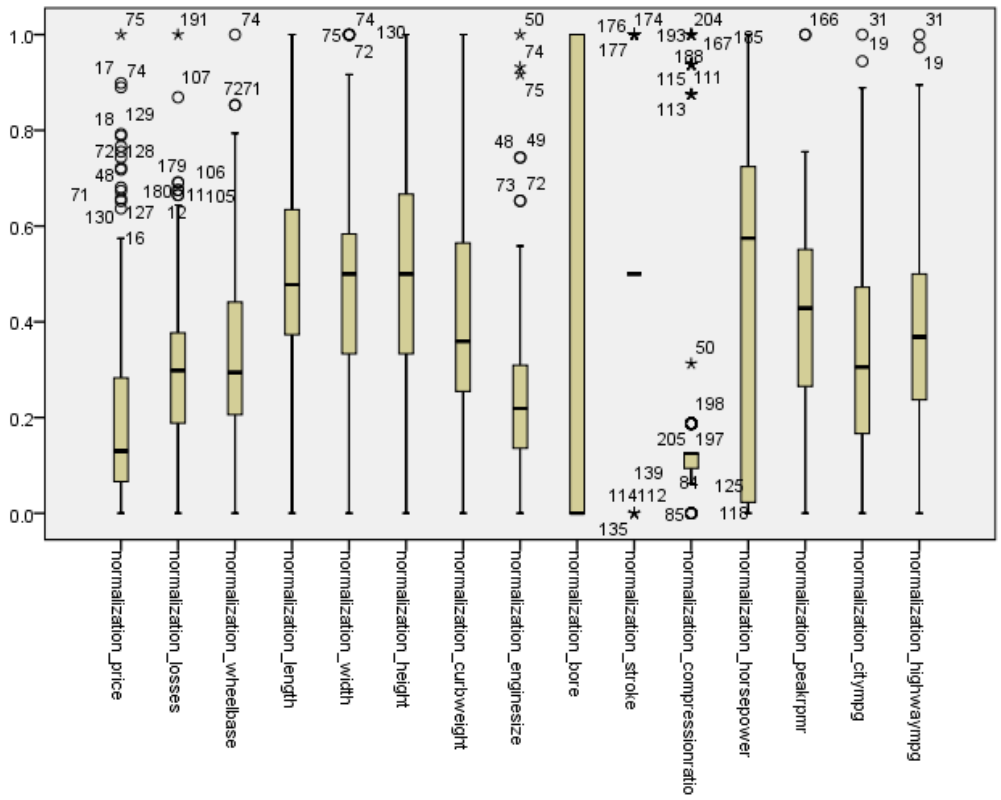


Figure 2. Box Diagram (Photo Credit: Original)

3.2. Correlation analysis

The correlation heat map of the quantitative variables of the influencing factors of automobile price is shown in Figure 3.

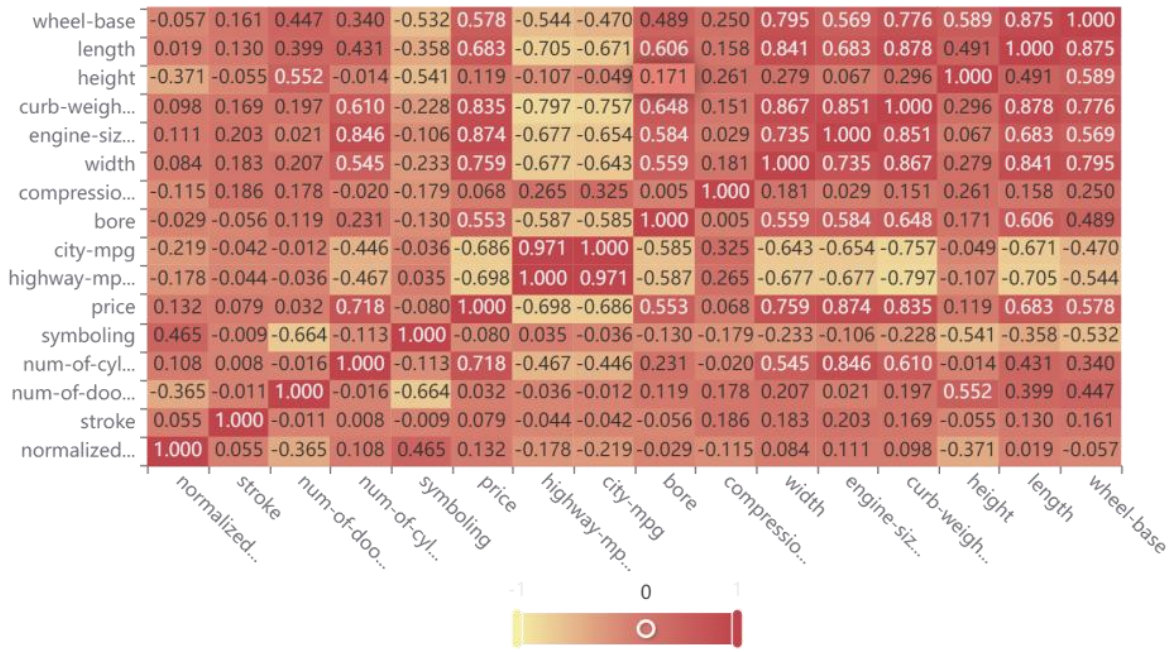


Figure 3. Correlation Heat Map (Photo Credit: Original)

It can be seen from Figure 2 that the eight variables, symbolizing, normalized losses, number of doors, height, stroke, compression-ratio, city-mpg, and highway-mpg, have little relationship with the price, thus they are eliminated directly. The correlation coefficients of fuel type, body style, and peak-rpm with price are -0.106, -0.084, and -0.086, respectively, therefore these three variables are also directly removed.

3.3. Variable selection

According to the above research, price is selected as the explained variable, and the preliminary selection of quantitative variables that affect the car price is the number of cylinders, wheelbase, length, width, curb weight, engine size, bore, and horsepower. For the quantification of qualitative variables, this paper selects five explanatory variables, namely, aspiration, drive wheels, engine location, engine type, and fuel system.

3.4. Establishment and verification of model

This paper imports the data set into SPSS software and estimates the independent variable coefficient of the data employing the least square method. From the result of the t-test, the variables curb weight and horsepower that is highly related to the price have not passed the significance test, and the sign of the regression coefficient of the bore is negative. Generally speaking, the larger the bore, the higher the price, but the result shows that the bore and the price have the opposite result. Considering that the VIFj of variables curbs weight and engine size is more than 10, it indicates that there is serious multicollinearity between this variable and other variables, and this multicollinearity may excessively affect the least squares estimation, so this paper chooses to apply stepwise regression to introduce other variables to select independent variables. According to the regression results, at the significance level of 5%, there are five independent variables, namely engine location, width, curb weight, number of cylinders, and drive wheels. Because the VIFj of engine size is 10.15, there is still a multicollinearity problem. This paper first removes the engine size variable and then re-establishes the regression model. The results are shown in Table 1 below.

Table 1. Results of Multiple Linear Regression

OLS Model	Unstandardized Coefficients	Standardized Coefficients	t	Sig.	VIF
(Constant)	-88766.864		-7.029	0.000	
Engine location	16527.748	0.249	-7.029	0.000	1.111
Width	868.694	0.233	8.490	0.000	4.215
Curb weight	5.846	0.381	4.083	0.000	5.276
Number of cylinders	2022.296	0.274	5.960	0.000	1.676
Drive wheels	1814.414	0.126	7.594	0.041	1.540
R	R Square	Adjusted R Square	F	Sig.	
0.931	0.846	0.842	218.414	0.000	

It can be seen from Table 2 that the R of the linear model is 0.931, R Square is 0.846, and Adjusted R Square is 0.842, which is close to 1, indicating that the explanatory power of these five factors on price is 84.2%, and the model fits well. F-value equals 218.414, and p-value equals 0.000<0.05, indicating that the regression equation is significant at the level of 5% significance, and there is a significant linear relationship between independent variables and dependent variables. After the multicollinearity treatment, the VIF-values of all independent variables are less than 10, which can be considered that there is basically no collinearity. The p-value of all independent variables is equal to 0.000<0.05, indicating that the regression coefficients of all variables are significant. The regression coefficient that is not standardized in the model represents the average characteristic price of the car, while the standardized coefficient reflects the difference in the impact of these five variables on the price. According to the non-standard regression coefficient, the five variables have a significant positive correlation with the price. The regression equation can be written as:

Price=-88766.864+16527.748Engine Location+2022.296Number of Cylinders+1814.414Drive Wheels.

3.5. Residual analysis

This paper draws a normalized residual histogram of the residual, as shown in Figure 3. It can be seen from Figure 4 that the left and right sides of the standardized residual are basically symmetrical, and the residual is an approximately normal distribution. It can be seen from Figure 5 that the scattered

points are basically scattered around the diagonal of the first quadrant, indicating that the residual normality result is good. It can be seen from Figure 6 that the distribution of standardized residuals shows a trend of diffusion with the increase of the value of variables. It can be considered that the residuals basically don't meet the homogeneity of variance.

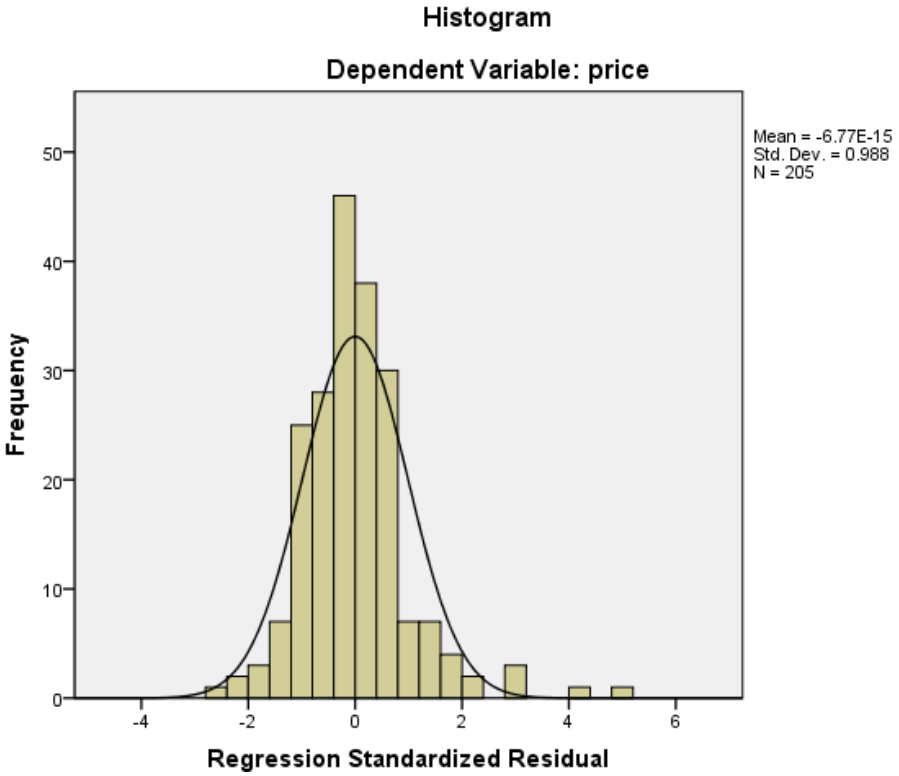


Figure 4. Standardized Residual Histogram (Photo Credit: Original)

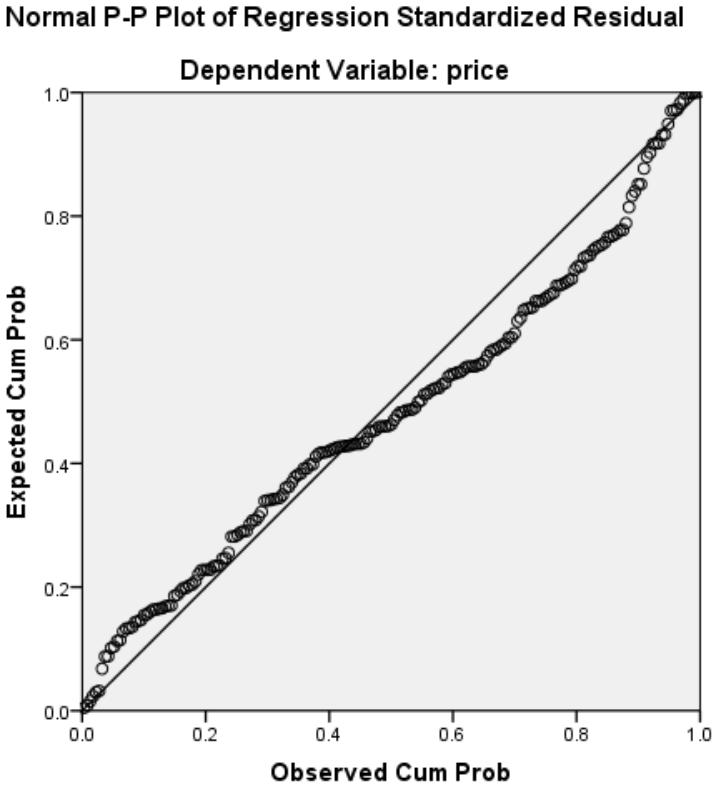


Figure 5. Normal P-P Plot (Photo Credit: Original)

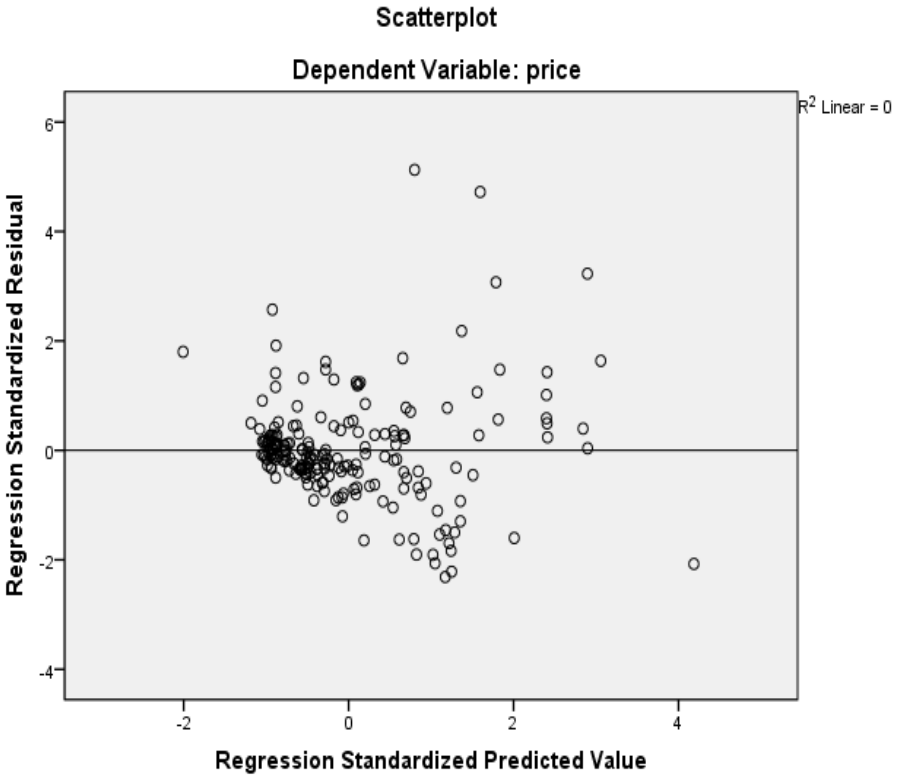


Figure 6. Standardized Residual Scatter Diagram (Photo Credit: Original)

3.6. Heteroscedasticity diagnosis

Then, this paper further diagnose whether the model has heteroscedasticity by utilizing the Spearman correlation coefficient method, and finds that the p-values of residual and curb weight and width are 0, less than 0.05, indicating that the residual and curb weight and width have significant correlation and heteroscedasticity[10]. The residuals of these two variables are shown in Figure 7 and Figure 8. All residuals don't change randomly around e=0 but tend to increase with the increase of independent variables.

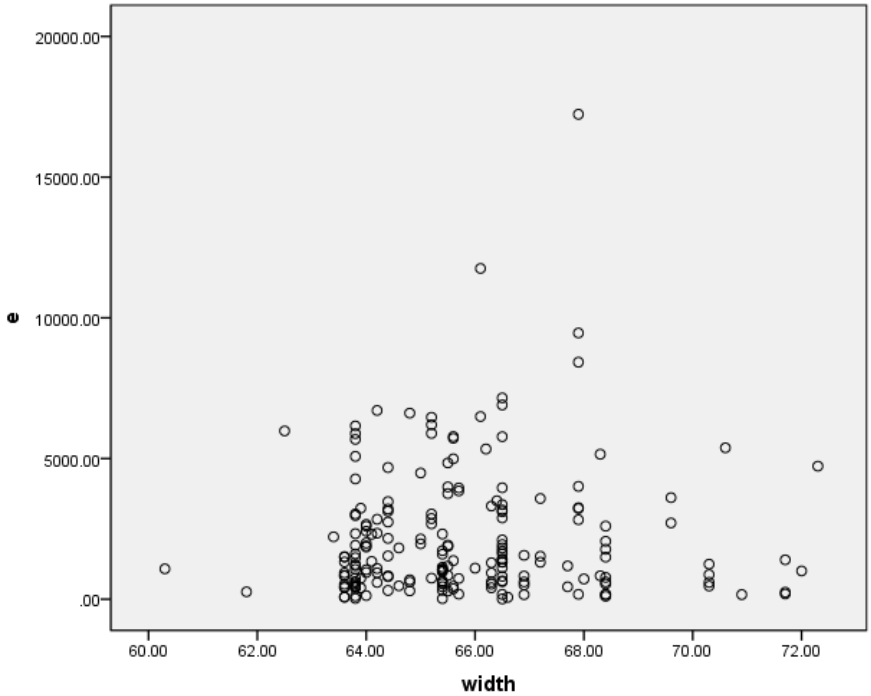


Figure 7. Residual Scatter Diagram of Width (Photo Credit: Original)

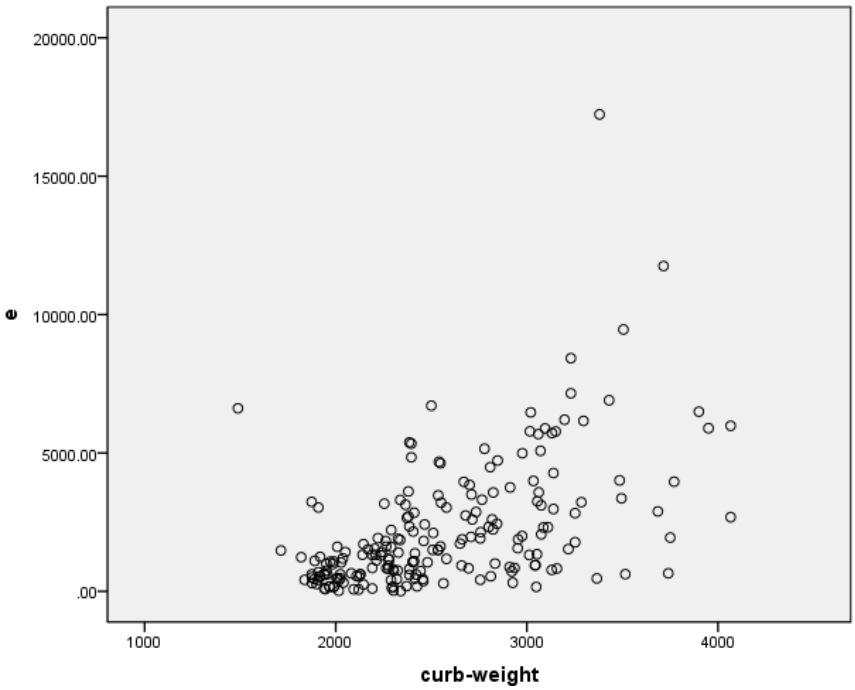


Figure 8. Residual Scatter Diagram of Curb Weight(Photo Credit: Original)

3.7. Model optimization

In order to eliminate the heteroscedasticity of the model, this paper exploits multivariate least squares estimation to modify the model. In this paper, the correlation coefficient between the residual and other variables is shown in Table 2 through the Spearman correlation coefficient method.

Table 2. Correlation Coefficient

	e	engine location	width	curb-weight	number-of-cylinders	drive-wheels
e	1.000	-0.012	0.446	0.565	0.313	0.491

Because the rank correlation coefficient between curb weight and residual is the largest, the curb weight is applied as a weight variable for multivariate weighted least squares estimation. Since the optimal value of the power exponent m=2.0 is reached at the boundary, this paper increases the selection range of the power exponent m, and the optimal power exponent of the calculated data reaches m=5.5, and the log-likelihood value is - 1890.370, which reaches the maximum. The regression results of the multivariate weighted least squares estimation are shown in Table 3.

Table 3. Results of Multiple Weighted Least Squares Regression

WLS Model	Unstandardized Coefficients	Standardized Coefficients	t	Sig.
(Constant)	-34745.543		-3.484	0.001
Engine location	17659.640	0.301	8.688	0.000
Width	34.019	0.013	0.202	0.840
Curb weight	7.313	0.621	9.186	0.000
Number of cylinders	896.643	0.111	2.800	0.006
Drive wheels	2011.642	0.180	4.860	0.000
R	R Square	Adjusted R Square	F	Sig.
0.890	0.792	0.787	151.844	0.000

It can be seen from Table 4 that the Adjusted R Square of the linear model is 0.787, which is close to 1, indicating that the model fits well. The F-value of the regression model is equal to 151.844, and the p-value is equal to 0.000<0.05, indicating that the regression equation is significant at the 5% significance level, and the independent variable and dependent variable have a significant linear

relationship. Since the p-value of width is 0.840>0.05, it indicates that the significance test of width is not passed and has no statistical significance, so this variable is excluded. The p-value of other independent variables is less than 0.05, indicating that the regression coefficient of other independent variables is significant. From the non-standard regression coefficient, it can be concluded that there is a significant positive correlation between the independent variables and prices. The multivariate weighted regression equation can be written as:

Price*=-34745.543+17659.640Engine Location+7.313Curb Weight+896.643Number of Cylinders+2011.642Drive Wheels.

In this paper, heteroscedasticity diagnosis is performed again after completing the weighted least squares estimation, the residual scatter diagram of weighted least squares of width and curb weight is shown in Figure 9 and Figure 10. It can be seen that the heteroscedasticity problem has been basically solved.

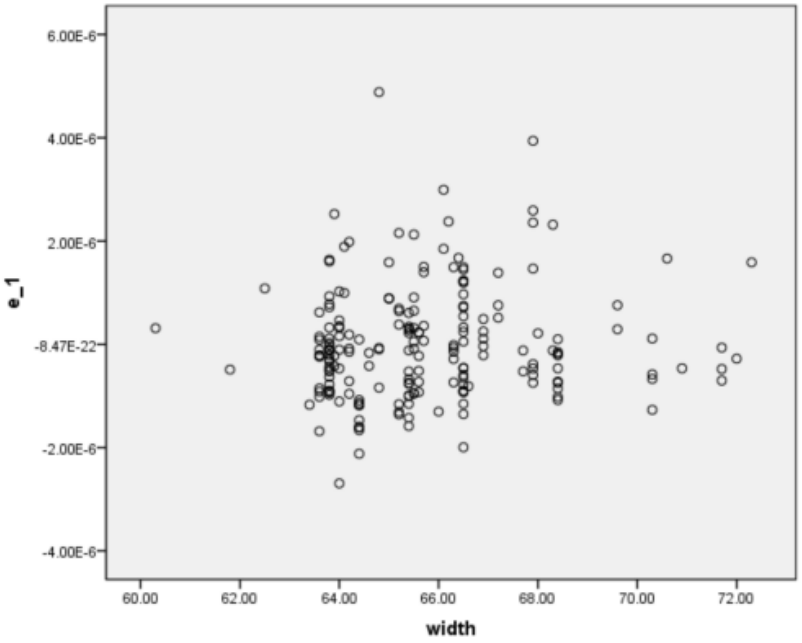


Figure 9. Weighted Least Squares Residual Scatter Diagram of Width (Photo Credit: Original)

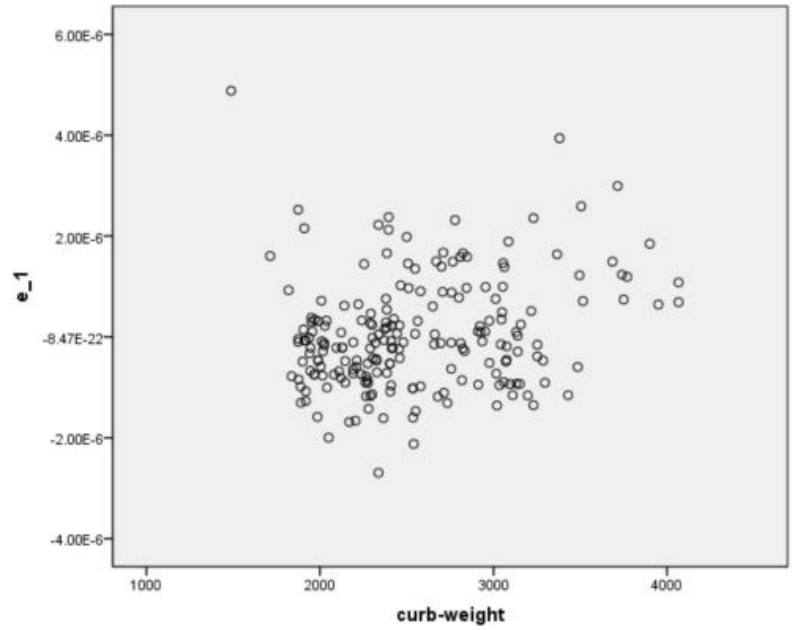


Figure 10. Weighted Least Squares Residual Scatter Diagram of Curb Weight (Photo Credit: Original)

3.8. Limitation

The current research still has some deficiencies in many aspects as follows:

(1) The data selected in this paper is cross-sectional data, which doesn't take into account the change in automobile attributes over time. However, in reality, the product attributes will change over time and are also affected by the market, society, culture, and other factors of the automobile.

(2) The data selected in this paper covers the hardware indicators and performance indicators of the automobile. Nevertheless, with the update and progress of the times and the extensive development of the network, the software systems of the car, such as ABS (anti-lock) system, reversing radar system, etc., also become the reference basis for customers.

(3) This paper only establishes a linear model of vehicle price and its influencing factors. In fact, there are various forms of functions. There are many forms of models that can be considered. It cannot be determined that the linear model has the highest degree of superiority.

Future research should consider the time factor, consider how to quantify the market, society, culture, and other factors, introduce more indicators, establish different functional models for analysis, and find the best model.

4. Conclusion

In this paper, the independent variables are selected through data processing and stepwise regression, and a multiple linear regression model is established. This paper uses the variance expansion factor to test the multicollinearity of the regression results, uses the residual analysis and Spearman correlation coefficient method to diagnose the heteroscedasticity of the model, and uses the multiple weighted least squares regression to correct the heteroscedasticity of the regression results and analyze the regression results of the model. It is concluded that the significant factors affecting the automobile price are engine location, curb weight, number of cylinders, and drive wheels. This is because the distance between the position of the engine and the center of gravity of the vehicle will determine whether the weight is concentrated at the front or rear of the vehicle, affecting the vehicle's balance weight ratio, so affecting the vehicle's handling and driving stability. The curb weight is related to the safety of the vehicle. Under the same cylinder diameter, the more cylinders, the larger the displacement, the more complex the production process, the higher the cost, and the higher the price. Different drive wheels have different driving modes, which affect the traction, steering, and flexibility of the vehicle. The main contribution of this paper is to provide a reference for consumers who have little knowledge of car properties of cars when purchasing cars and help consumers buy the car products, they like by studying the factors that affect the price of cars.

References

[1] Wei, HJ. Literature review on the influencing factors of second-hand car price evaluation in China [J]. Times Auto, 2022, (24): 166 - 168.

[2] Industry Information Department of China Automobile Association. Production and sales of the automobile industry in 2022. January 12, 2023. Retrieved on March 10, 2023. Retrieved from: http://www.caam.org.cn/chn/4/cate_32/con_5236639.html

[3] Cen, J. Research on Influencing Factors of Automobile Price Based on Stochastic Forest and Neural Network [C]. Suzhou University, 2020.

[4] Li, XN. An Analysis of Impacting Factors on Domestic Automobile Price Based on Hedonic Price Theory[J]. Prices Monthly, 2018.

[5] Yang, CJ. Hedonic Price for Cars: An Application to the China Car Market[C]. Chongqing Normal University, 2011.

[6] Ozgur C, Hughes Z, Rogers G, et al. Multiple Linear Regression Applications Automobile Pricing. International Journal of Mathematics and Statistics Invention, 2016.

[7] Zhang, LJ. Empirical Analysis of Factors Affecting Price Fluctuation in the Domestic Automobile Industry [J]. Mall Modernization,2016, (13): 8 - 9.

[8] Ramakrishnan Srinivasan. Automobile Dataset. Retrieved on March 10, 2023. Retrieved from: https://www.kaggle.com/datasets/toramky/automobile-dataset.

[9] Yang, J; Li, H; Zhang, YH. Prediction of electric vehicle price based on support vector machine [J]. Enterprise Technology and Development, 2022, (01): 79 - 81.

[10] Jiang, JM. Research on heteroscedasticity in multiple linear regression model [C]. Guilin University of Technology, 2022.