2024 International Journal of Management Science Research, Vol.7, Issue 6, (Dec)
ISSN 2536-605X
DOI: 10.53469/ijomsr.2024.07(06).06

A Study on the Second-hand Sailboat Price Prediction Model Based on Principal Component Analysis and Multiple Linear Regression

Jiayi Hu

Zhongnan University of Economics and Law, School of Statistics and Mathematics, Financial Mathematics

Abstract: ***This paper proposes a second-hand sailboat price prediction model based on Principal Component Analysis (PCA) and Multiple Linear Regression (MLR) to address issues of non-transparent transactions and information asymmetry in the market. By analyzing four key indicators year, condition, appearance, and performance and categorizing national economic development levels using the Regional Economic Development Index (REDI), we transform these levels into dummy variables. Separate regression analyses were conducted for monohull and catamaran sailboats. The study shows that the year, appearance, and performance of sailboats are highly positively correlated with their prices, and regional economic development levels significantly affect prices. The model demonstrates excellent performance in predicting second-hand sailboat prices, aiding sellers in pricing accurately and promoting market transparency and development.***

Keywords: Second-hand sailboats; Price prediction; PCA; MLR; REDI.

1. INTRODUCTION

With economic development and the advancement of globalization, the international environment has become relatively stable, and the global shipbuilding industry is gradually recovering [1]. The decrease in the number of new ship orders has led to a more balanced supply and demand in the ship market [2].

However, the booming second-hand sailboat market has also brought various issues: non-transparent market transactions, information asymmetry between buyers and sellers, lack of channels for consumers to purchase second-hand sailboats, and the absence of uniform evaluation standards [3]. These problems hinder the development of the second-hand sailboat market [4]. Although relevant second-hand sailboat information can be found in various regions with the spread and development of the internet, current platforms and policies are not perfect, nor is there a unified evaluation standard [5]. Factors influencing the price of second-hand sailboats include the boat's age, onboard equipment, other self-loss factors, relevant policies in different countries and regions, the freight of the sailboat market, and the scope of sailboat operations [6]. In fact, the assessment of sailboat prices is limited not only by the evaluation of the sailboat's value but also by the asymmetry of collectible data and information.

Currently, research on predicting second-hand sailboat prices primarily focuses on using statistical models and machine learning algorithms. For instance, traditional Multiple Linear Regression (MLR) [7] and Support Vector Machines (SVM) [8] have been employed to predict the prices of second-hand vessels. However, these methods have certain limitations when dealing with high-dimensional data and multicollinearity issues [9] [10]. To improve prediction accuracy, researchers have attempted to incorporate Principal Component Analysis (PCA) to reduce data dimensionality and noise [11]. Although these approaches have achieved some success, significant challenges remain in data preprocessing, model selection, and parameter tuning, resulting in suboptimal prediction accuracy.

Presently, Multiple Linear Regression (MLR): MLR is a commonly used statistical method that establishes a linear relationship between multiple independent variables and a dependent variable to predict prices. Despite its advantages in interpretability and operability, MLR is prone to multicollinearity issues when dealing with high-dimensional data, affecting the model's stability and prediction accuracy [7]. Support Vector Machines (SVM): SVM is a supervised learning method based on statistical learning theory, particularly suitable for small sample sizes and high-dimensional data. SVM seeks to find the optimal hyperplane in a high-dimensional space to

achieve prediction. While SVM has advantages in handling complex data structures, its parameter selection and model training process are relatively complex and computationally expensive [8]. Random Forest (RF): RF is an ensemble learning method based on decision trees, improving prediction stability and accuracy by constructing multiple decision trees and averaging their results. RF performs well in handling nonlinear relationships and high-dimensional data, but its results are less interpretable, making it difficult to identify the specific impact of each variable [12]. Neural Networks (NN): NN is another commonly used machine learning method, especially suitable for handling complex nonlinear relationships. By using multilayer network structures and nonlinear activation functions, NN can capture deep patterns in data. However, NN requires a large amount of training data and computational resources, and its results often lack interpretability [13]. Principal Component Analysis (PCA): PCA is a dimensionality reduction technique that projects high-dimensional data into a lower-dimensional space to extract key features, thereby reducing data dimensionality and noise. PCA has significant advantages in addressing multicollinearity issues and improving model prediction accuracy, but its dimensionality reduction process may result in some information loss [14].

To overcome the limitations of existing methods, this study proposes a second-hand sailboat price prediction model combining PCA and MLR. By using PCA to reduce data dimensionality and extract key features, and employing MLR to establish a price prediction model, this study aims to improve prediction accuracy and stability. Specifically, the main tasks of this study are:

(1) Based on exploratory variables, determine four indicators-year, condition, appearance, and performance-using PCA.

(2) Use cluster analysis methods to classify national indicators into three categories according to REDI levels and transform them into dummy variables.

(3) Conduct separate regression analyses for monohull and catamaran sailboats, showing that year, appearance, and performance are highly positively correlated with prices, consistent with economic theory.

By combining PCA and MLR, this study not only addresses high-dimensional data and multicollinearity issues but also improves the accuracy and stability of second-hand sailboat price predictions, providing strong support for accurate pricing by sellers and promoting market transparency and development.

2. PRELIMINARY

2.1 Assumptions

To simplify the problem, we made the following basic assumptions:

Assumption1: The wear and tear on the boat is positively correlated with its age, and the price given is the normal price after wear and depreciation.

Justification: Over time, sailboats inevitably experience wear and tear, leading to a decrease in their price. Any sailboat with significant damage to its appearance or performance will undergo substantial devaluation, regardless of its age. However, it is challenging to determine which sailboats have suffered severe damage based solely on available information, and it is unclear whether abnormal price changes are due to other factors. Therefore, this assumption allows us to explore general patterns in sailboat price changes.

Assumption2: The material of the boat indirectly affects the price of the sailboat through its appearance and performance.

Justification: Based on the data set we collected, most sailboats are made of fiberglass, and the impact on price is not significant but cannot be ignored. Therefore, we assume that different materials will cause differences in the appearance and performance of the sailboat, thus affecting its price.

Assumption3: We assume that second-hand sailboats are traded only domestically, without the impact of tariffs.

Justification: Since used sailboats are typically already in circulation and have paid related tariffs, taxes, and other fees, trading them within the same country or region usually does not require additional tariffs. Additionally, tariffs

vary by country and are easily influenced by policies, so this assumption simplifies the pricing model.

Assumption4: We assume the sailboat trading market is an oligopoly market where companies have pricing power.

Justification: Since businesses hold the pricing power, this article's pricing model focuses solely on the supply side, avoiding separate discussion of the demand side. According to the appendix, the main manufacturers are Beneteau (22%), Jeanneau (21.9%), Bavaria (14.1%), Hanse (7.6%), and Dufour (6.9%), all from France and Germany, indicating an oligopoly market.

Assumption5: Gaussian-Markov assumption.

Justification: When exploring factors affecting prices, a linear regression model is

established. If the data meets the assumption of Gaussian-Markov, we can use the OLS method to estimate and infer the regression coefficients.

2.2 Notations

The symbol description of the paper is shown in Table 1.

Table 1: Notations used in this paper

Symbol	Description	Unit
df_dummyState-1	economies with poor-level development	/
df_dummyState0	economies with middle-level development	/
df_dummyState-1	economies with high-level development	/

2.3 Data Description

The data for this study were sourced from Chinese and international sailboat trading websites [13], including parameters such as production status, freshwater capacity, fuel capacity, boat weight, draft depth, and materials. To ensure data quality, the following preprocessing steps were undertaken:

Outlier Handling: Outliers in continuous variables were screened using the 3 principle, while box plot methods were applied to discrete variables. Outliers were treated as missing values.

Interpolation: Continuous variables were interpolated using the inverse distance weighting method, and discrete variables were interpolated using the nearest neighbor method.

Standardization: Continuous variables were standardized using Z-score normalization to ensure stability in subsequent algorithms.

Through these steps, we ensured the completeness and accuracy of the data, providing a solid foundation for model construction.

3. PCA & MULTIPLE LINEAR REGRESSION MODEL AND SOLVE

When constructing the price prediction model, we first applied Principal Component Analysis (PCA) to reduce the dimensionality of the data, extracting two primary variables: appearance and performance. Next, we used the Regional Economic Development Index (REDI) to classify countries/regions, converting them into dummy variables. Finally, we employed Multiple Linear Regression (MLR) to develop a price prediction model based on year, appearance, performance, and regional economic development.

3.1 Analysis of Correlation Between Variables

In the exploratory analysis of variable characteristics, there are 15 types of explanatory variables in our dataset, making it relatively difficult to directly use stepwise regression analysis in calculations, with the risk of overfitting. Therefore, we considered using PCA to reduce the dimensionality of the data, fully explore the contribution rate and cumulative contribution rate of the principal components, screen the correlation between explanatory variables,

better extract the principal components that significantly affect the target variable, and avoid the problem of multicollinearity.

There are too many samples in the dataset, which is difficult to observe intuitively, and the types of variables are not single, and the covariance matrix is poorly plotted. Therefore, the correlation analysis between variables is carried out in this paper.

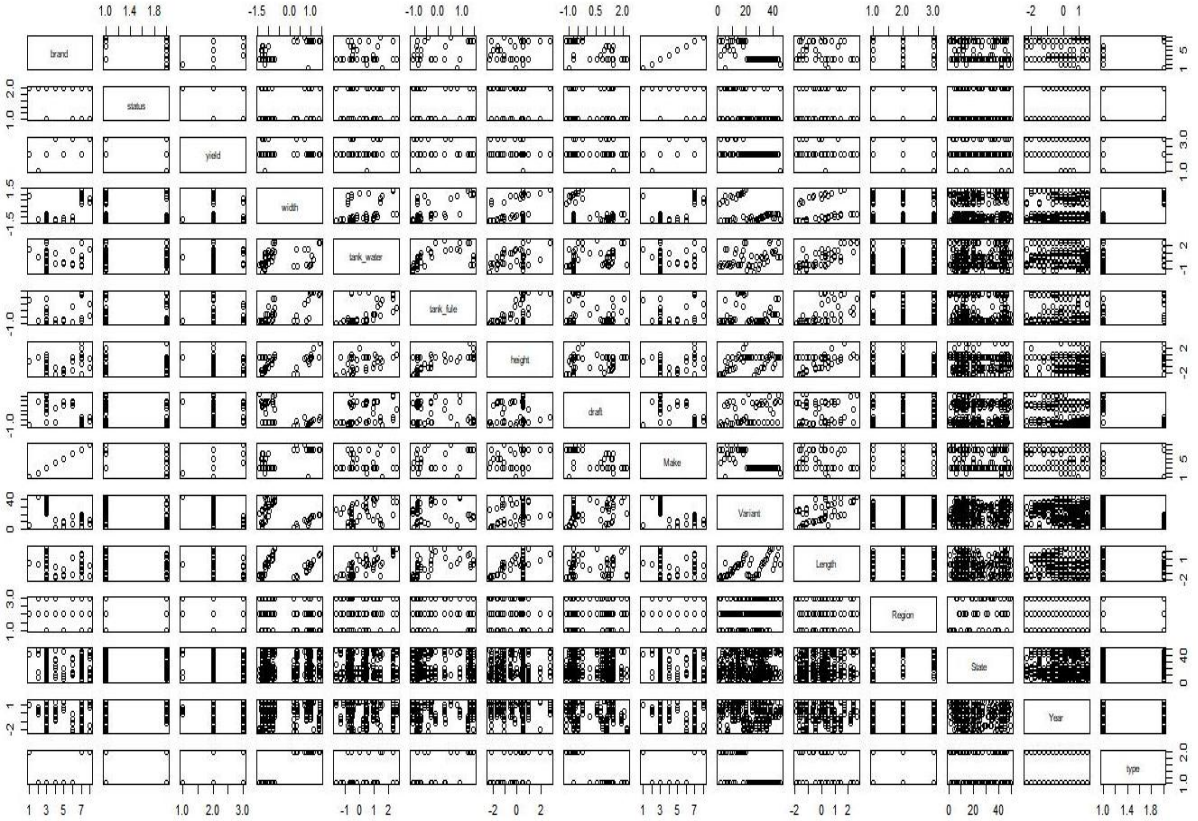


Figure 1: Regression trend of different variables on price.

Stepwise regression analysis

By analyzing Figure 1, we find that the state and output indicators do not have sufficient coverage and do not significantly affect prices. Therefore, they should be directly deleted. Under the type indicator, variables such as width, draft, and tank water show obvious stratification, indicating a high correlation between type and these indicators. The correlation coefficient between type and other variables is high, especially the correlation coefficients between width and draft, which are 0.9683464 and -0.8421405, respectively, showing strong positive and negative correlations. This confirms the necessity of separately studying monohull and catamaran sailboats. The width and fuel tank have a strong positive correlation with the price, with correlation coefficients of 0.7190577 and 0.7952913, respectively. These three variables should be considered simultaneously when establishing the regression model. There is a certain correlation between tank water and tank fuel, with a correlation coefficient of 0.1092401. However, since they are both related to tank variables, they will be classified together for further analysis. Future research can choose width and tank fuel as the main regression variables. Therefore, in subsequent analyses, type should be treated as a categorical variable, and monohull and catamaran sailboats should be studied separately.

Based on the results of the analysis in Figure 1, we excluded variables with poor attributes and classified the data according to type (i.e., monocoque and bicove).

Before performing regression analysis, the correlation between the dependent and explanatory variables needs to be analyzed.

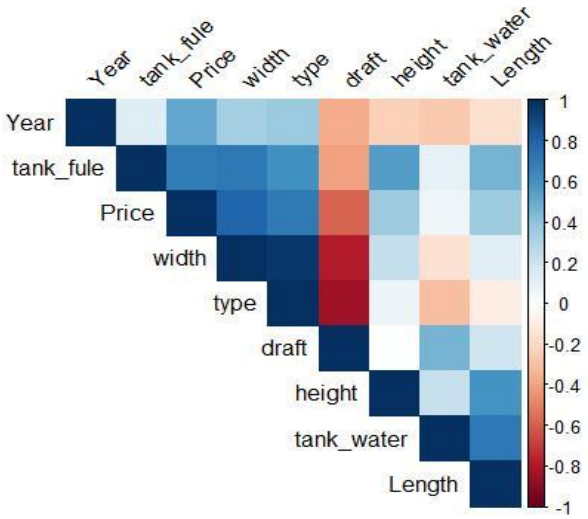


Figure 2: Covariance matrix of continuous explanatory variables

By analyzing Figure 2 and Figure 3, we find that the width, variable, and draft indicators are divided into two parts, possibly due to the influence of the type variable (all samples in the first half come from type 0 (monohull), and all samples in the second half come from type 1 (catamaran)). By analyzing Figure 3, it can be concluded that variables such as tank water, tank fuel, height, length, and year show obvious linearity and explanatory power for prices. The state indicator has a small impact on price. In subsequent analyses, other quantitative data on state (region) should be considered, and state should be analyzed after classification. The make variable is mainly concentrated in two insufficient sampling points and should be directly deleted. This phenomenon also indicates that these boats mainly come from two shipyards, both from France, supporting Assumption 4.

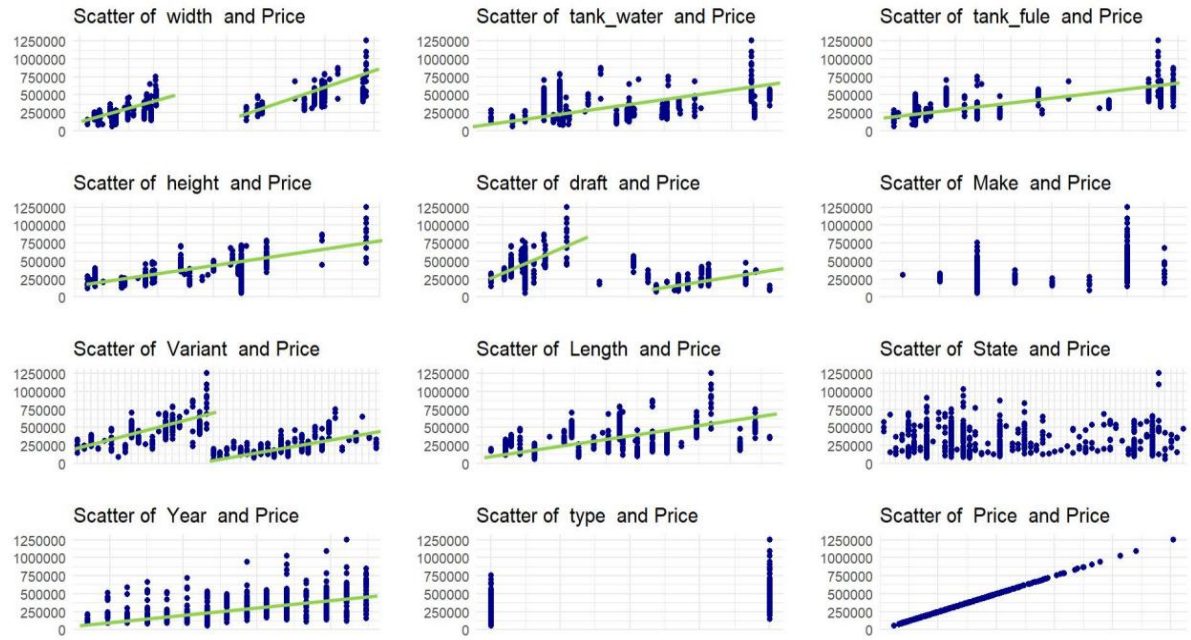


Figure 3: Scatter plot of each variable and price

3.2 Principal Component Analysis

Given that the pricing of sailboats is an econometric issue, variable selection must consider its economic significance rather than relying solely on its mathematical properties. Therefore, in this section, we will explore from both mathematical and economic perspectives and finally draw conclusions based on these two research results. We retained two core variables: the age (year) and region (state) of the boat. We used PCA to reduce the dimensionality of the dataset and integrated variables such as model, length, yacht width, draft, height, tank water capacity, and tank fuel capacity to obtain two new variables: appearance and performance. The reasons are as

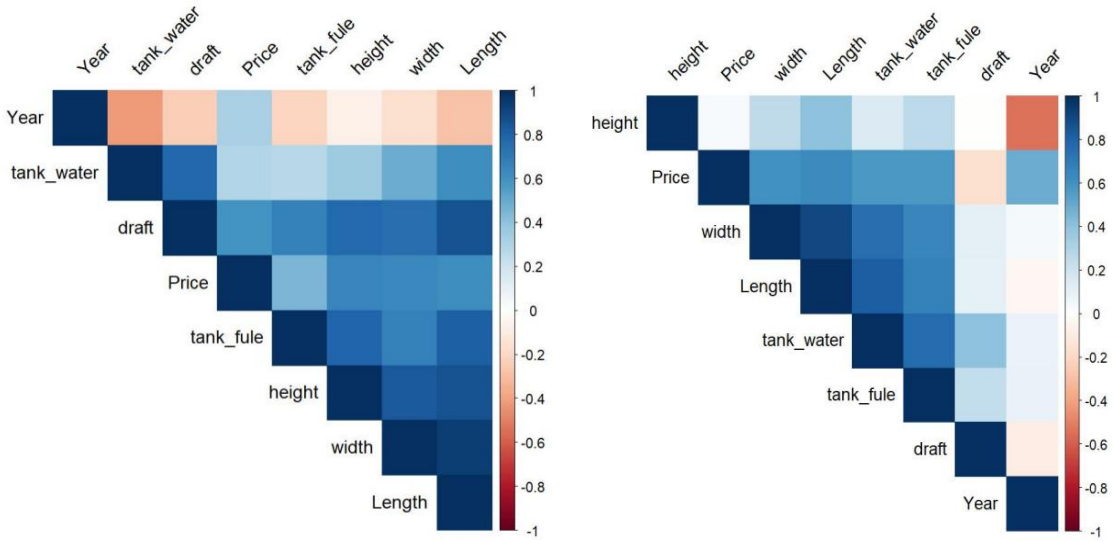
follows:

From the perspective of demand, consumers tend to prefer sailboats with excellent appearance and performance because sailboats themselves are luxury goods with higher social and self-image value than practical value. Excellent appearance and performance can increase consumer satisfaction and self-image, thus increasing demand for these goods and driving up prices.

From the perspective of supply, appearance and performance also affect the production cost and competitive advantage of sailboats, thereby affecting prices. Manufacturing sailboats with beautiful appearance and high performance may require more research and development, design, materials, and production costs, leading to higher prices. At the same time, if a brand's sailboats have significant advantages in appearance and performance, it may bring higher brand premiums and market share, enhancing competitive advantage and pricing power on the supply side, thereby driving up prices.

In microeconomics, the price of goods is determined by the relationship between supply and demand, including consumers' demand for product characteristics and the cost and competitiveness of producers supplying the goods. In industrial organization theory, appearance and performance are also considered important means for companies to gain competitive advantage and market share.

By analyzing the covariance matrix and scatter plot of each variable with price for monohull (Figure 4) and catamaran (Figure 5) sailboats, we find that variables such as length, height, width, and draft have strong correlations, while variables such as tank water and tank fuel have strong correlations.



Figures 4 and 5: Covariance matrices of continuous explanatory variables

Based on the above-selected strongly correlated variables, we conducted PCA and introduced two new variables, appearance and performance, using the following formulas (1) and (2) for monohull sailboats:

Table 2: The main component of the appearance of a monocoque shell

Principal Component	length	with	height	draft	tank_water	tank_fuel	length
Weight Coefficient	W_1	W_2	W_3	W_4	K_1	K_2	K_3

$$appearance = W_1 length + W_2 width + W_3 height + W_4 draft \quad (1)$$

$$performance = K_1\, tan\, k\_water + K_2 tan\, k\_fuel + K_3 length \quad (2)$$

Similarly, for catamarans, we obtained the following results, as shown in formulas (3) and (4):

Table 3: The main component of the appearance of the twin shell

Principal Component	length	with	height	draft	tank_water	tank_fuel	length
Weight Coefficient	J_1	J_2	J_3	J_4	L_1	L_2	L_3

$$appearance = J_1 length + J_2 width + J_3 height + J_4 draft \quad (3)$$

$$performance = L_1\,tank\_water + L_2\,tank\_fuel + L_3 length \quad (4)$$

Using machine learning in Python, we obtained the weight coefficients:

$[W_1,W_2,W_3,W_4]$ = [-0.13336412 , -0.46474554 ,-0.08189191 ,-0.87150405]
$[K_1,K_2,K_3]$ = [0.61266632,0.22797887,0.75674673]
$[J_1,J_2,J_3,J_4]$ = [0.19323485 ,0.78293582 ,0.10991818,0.58102478]
$[L_1,L_2,L_3]$ = [0.48291933 ,0.6500902,0.58666145]

3.3 State: Cluster Analysis Based on the Regional Economic Development Index (REDI)

For the categorical variable "state," we adopted a comprehensive REDI proposed by Hu Liang and Walter, which includes five dimensions: per capita income, employment, market size, infrastructure and public services, and technological innovation. REDI calculates the weighted average of these indicators to reflect a region's level of economic development. REDI can be used to evaluate a region's economic development level and classify it into three levels: high, medium, and low economic development.

After accessing the corresponding data for countries/regions from the World Bank website in the "2023_MCM_Problem_Y_Boats.xlsx" file, we cleaned the data and divided countries into high, medium, and low economic development categories based on their REDI values. The classification standard is: regions with the top 33.3% of the indicator scale are classified as high economic development countries, marked as state 1; the bottom 33.3% are classified as low economic development countries, marked as state -1. To facilitate regression analysis, we converted the categorical variable "state" into dummy variables, namely, dummy state -1, dummy state 0, and dummy state 1, corresponding to low, medium, and high economic development states, respectively.

Table 4: Dummy variable processing of "state" based on REDI

Index Rank	**top33.3%**	**middle33.3%**	**botttom33.3%**
Development_Level	high	medium	low
Name of Variable	State1	State0	State-1

3.4 Multiple Linear Regression and Results Analysis

Using appearance, performance, and state variables established through PCA and REDI, as well as year as prediction indicators, we established a multiple regression model for price. By analyzing scatter plots, we found that these four variables are linearly correlated with price. Therefore, we established equation (5):

$$\text{Price} = \beta_0 + \beta_1 Year + \beta_2 State + \beta_3 appearance + \beta_4 performance + \mu \quad (5)$$

3.4.1 Model Results Analysis for Monohull Sailboats

Model evaluation indicates that all independent variables, except for df_dummyState1, are significant (P-value < 0.05). The adjusted R-squared value for the regression analysis is 0.7756, demonstrating relatively good explanatory power for cross-sectional data. The conclusions are as follows:

(1) Year and performance have a significant positive impact on the price of sailboats (the dependent variable). This suggests that these two variables effectively explain price variations, with increases in their values corresponding to higher sailboat prices, in line with economic predictions.

(2) The coefficients for df_dummyState-1 and df_dummyState0 are negative, indicating that sailboat prices in these regions are lower than those in the baseline state (df_dummyState1). This implies that sailboat prices are lower in economically underdeveloped regions, which is consistent with economic expectations.

Table 5: Single-shell linear regression model

Variable	Estimate	Std.Error	t value	Pr(>|t|)
(Intercept)	295308.66	5768.556	51.192823	0.0000000
Year	63933.26	3603.793	17.740547	0.0000000
Df_dummyState-1	-77350.81	7661.405	-10.096165	0.0000000
Df_dummyState0	-42788.89	7805.703	-5.481747	0.0000001
performance	14289.70	5806.338	2.461052	0.0144103
appearance	-42656.49	6967.647	-6.122079	0.0000000

3.4.2 Model Results Analysis for Catamarans

Model evaluation demonstrates that coefficients such as year, region (expressed as dummy variables), product performance, and appearance significantly affect the dependent variable. The coefficients for year and product performance/appearance are positive, while the coefficient for the region is negative. Notably, the dummy variable for region -1 has the greatest and statistically significant impact on the dependent variable. Goodness-of-fit statistics reveal that the model explains 61.1% of the variance in the dependent variable, with a multiple R-squared of 0.611 and an adjusted R-squared of 0.605. The F-statistic is 102.7, with a P-value less than 0.05, indicating that the model possesses significant statistical significance and predictive value.

All coefficients of the explanatory variables align with expectations. Specifically, the year, performance, and appearance of the sailboat are positively correlated with its price. The year variable has the most significant impact; for each additional year, the price increases by 77,839.95 RMB.

Table 6: twin shell linear regression model

Variable	Estimate	Std.Error	t value	Pr(>|t|)
(Intercept)	469191.04	12568.753	37.329960	0.0000000
Year	77839.95	5837.973	13.333385	0.0000000
Df_dummyState-1	-34502.35	13747.337	-2.509748	0.0125645
Df_dummyState0	-12802.93	14595.178	-0.877203	0.3810204
performance	25661.65	8230.821	3.117751	0.0019845
appearance	49542.28	8290.042	5.979119	0.0000000

3.5 Accuracy Testing and Visualization

Observed from Table7 and Table 8, This is shown in the table below, the model has high significance, with an F-statistic of 213.926 ($p < 0.01$), an R-squared of 0.779, and an adjusted R-squared of 0.776. This indicates that the model explains most of the variance in price changes and fits the observed data well. Additionally, the coefficients of year, performance, and appearance are all significant ($p < 0.05$ or lower), meaning they significantly impact predicting product prices. Compared with the reference level, states -1 and 0 have significant negative impacts on prices, while state 1 does not have a significant impact. Overall, the model can be used to predict the price of a given entity, as it correctly explains most of the variance in the existing data.

Table 7: Single-shell Indicator

Residual standard error:53980 on 303 degrees freedom
Multiple R-squared:0.7793 **Adjusted R-squared:**0.7756
F-statistic:213.9 on 5 and 303 DF **p-value:**<2.2e-16

Table 8: twin shell Indicator

Residual standard error:75740 on 327 degrees freedom
Multiple R-squared:0.611 **Adjusted R-squared:**0.605
F-statistic:107.7 on 5 and 327 DF **p-value:**<2.2e-16

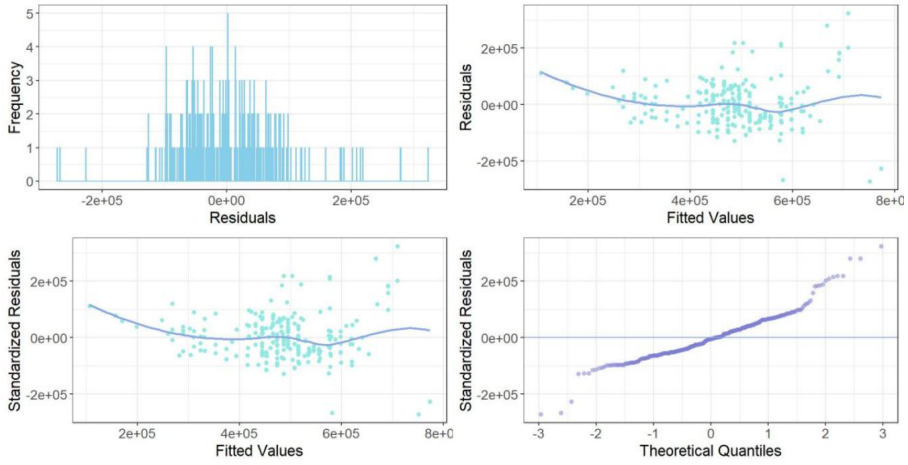


Figure 6: monohulled sailboats residual analysis diagram

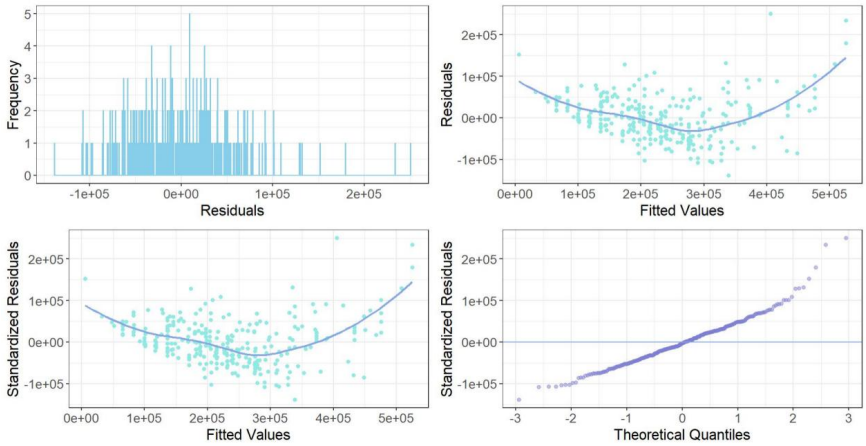


Figure 7: catamarans residual analysis diagram

Observing Figures 6 and 7, it is found that the residual plots show systematic errors in certain areas of the model.

The relationship between fitted values and residuals indicates that the model may have heteroscedasticity (the variance of the residuals changes with the fitted values).

The standardized residuals and the theoretical distribution in the Q-Q plot show that the residuals approximately follow a normal distribution, but there are deviations in the extreme values, suggesting the presence of some outliers or that the model has not fully captured certain data characteristics.

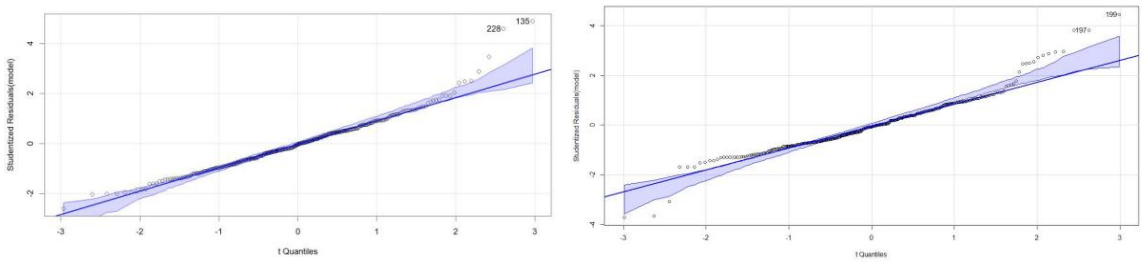


Figure 8, 9: Normalized Residual QQ Plot

From the Figure8,9, we find that the sample points on the graph generally form a straight line. This indicates that the residuals of the model follow a normal distribution. The correspondence between sample quantiles and theoretical quantiles is good. However, the applicability of this model in different market environments requires further validation. Additionally, the model may exhibit bias when predicting extreme prices. Future research should consider more variables that influence price to enhance the model's generalizability and predictive accuracy.

Overall, the charts indicate that performance and appearance have a significant impact on price. Additionally, the residual analysis shows that there are some biases in the model, suggesting the need for further optimization to improve prediction accuracy.

4. CONCLUSION

This study successfully identified the key factors influencing second-hand sailboat prices, including year, appearance, performance, and regional economic development level, by constructing a price prediction model based on PCA and MLR. This model holds significant practical value, as it can assist sellers in pricing more accurately and promote transparency in the market.

However, this study also has certain limitations. For instance, the data collection scope is limited, and the applicability of the model in different market environments requires further validation. Future research should expand the data scope and consider more variables that influence price to enhance the model's generalizability and predictive accuracy. Additionally, future studies could incorporate other predictive methods, such as machine learning algorithms, to further improve price prediction accuracy.

REFERENCES

[1] Xavier J R, Ramesh B. A study on the effect of multifunctional tantalum carbide nanofillers incorporated graphene oxide structure in the epoxy resin for the applications in the shipbuilding industry. Materials Science and Engineering: B, 2023, 289: 116234.
[2] Lazim H M, Abdullah J. Malaysia shipbuilding industry: A review on sustainability and technology success factors. Journal of Advanced Research in Applied Sciences and Engineering Technology, 2022, 28(3): 154-164.
[3] Liu T. Porosity reconstruction based on Biot elastic model of porous media by homotopy perturbation method. Chaos Solitons & Fractals, 2022, 158: 112007.
[4] Nam H S, De Alwis N, D'agostini E. Determining factors affecting second-hand ship value: Linkages and implications for the shipbuilding industry. WMU Journal of Maritime Affairs, 2022, 21(4): 493-517.
[5] Emelianov V, Zhilenkov A, Chernyi S, et al. Application of artificial intelligence technologies in metallographic analysis for quality assessment in the shipbuilding industry. Heliyon, 2022, 8(8): e10002.
[6] Kou Y, Liu L, Luo M. Lead-lag relationship between new-building and second-hand ship prices. Maritime Policy & Management, 2014, 41(4): 303-327.
[7] Peng W H, Adland R, Yip T L. Investor domicile and second-hand ship sale prices. Maritime Policy & Management, 2021, 48(8): 1109-1123.
[8] Liu T. Parameter estimation with the multigrid-homotopy method for a nonlinear diffusion equation. Journal of Computational and Applied Mathematics, 2022, 413: 114393.
[9] Lee C, Park K. Deep learning-based modeling of second-hand ship prices in South Korea. IAES International Journal of Artificial Intelligence, 2022, 11(3): 886.
[10] Lim S S, Lee K H, Yang H J, et al. Panamax second-hand vessel valuation model. Journal of Navigation and Port Research, 2019, 43(1): 72-78.
[11] Liu T, Yu J, Zheng Y, et al. A nonlinear multigrid method for the parameter identification problem of partial differential equations with constraints. Mathematics, 2022, 10(16): 2938.
[12] The Price Elasticity of Luxury Demand: Evidence from the United States and China. Journal of Business Research, Vol. 69 No. 1 (Jan. 2016), pp. 332-337.
[13] The Price Elasticity of Luxury Demand: Evidence from the United States and China, "Journal of Business Research, Vol. 69, No. 1 (Jan., 2016), pp. 332-337