The Sailing Boat Price Study Based on Principal Component Regression Analysis

Yang Chen [*, 1, #], Zifeng Li [1, #], DaPeng Jia [2, #]

[1] School of Mathematics, South China University of Technology, Guangzhou, China, 510641

[2] School of Mechanical and Automotive Engineering, South China University of Technology, Guangzhou, China, 510641

* Corresponding Author Email: 202130322117@mail.scut.edu.cn

Abstract. This paper conducts research on the pricing of second-hand sailboats under certain assumptions. Initially, a substantial amount of potentially relevant indicators and data related to sailboat pricing were collected from the internet. The Pearson correlation coefficient was used to investigate the correlation between these indicators and to filter them. Due to the distinctive characteristics of sailboats, they were categorized into monohull sailboats and catamarans for separate analysis. Subsequently, the processed data were used for price prediction using a BP neural network, yielding preliminary results. To obtain an explicit method for calculating sailboat prices, this paper employed principal component analysis to calculate the pricing objective function for both types of sailboats. This led to the derivation of two multivariate linear regression equations, which were subjected to significance testing. The results indicate that the objective functions obtained in this paper can serve as references for pricing second-hand sailboats.

Keywords: Pearson Correlation Coefficient, Principal Component Regression Analysis, Multiple Linear Regression Equation, Pricing Function.

1. Introduction

The Pricing Issue is highly pervasive in people's daily lives, with the pricing of each product influenced by various factors, thereby exerting significant impacts on the market and its related industries. Currently, research on pricing issues primarily focuses on financial products, employing diverse research methods, including machine learning and weighting methods. However, there is a scarcity of research outcomes on the pricing of second-hand sailboats. To some extent, this paper aims to fill this gap by conducting research on the pricing of second-hand sailboats.

This paper collects various indicators that may have an impact on sailboat pricing, explores the relationships between these indicators, and then utilizes a neural network to obtain preliminary predictive results. The paper posits that there is a certain degree of linearity between sailboat pricing and the various indicators. Therefore, a multivariate linear function is chosen as the objective function for sailboat pricing. Based on the principal component analysis method, the paper calculates the parameters of the objective function, ultimately providing a reference objective function for the pricing of sailboats. (The paper's data is sourced from the following website: https://sailboatdata.com/.)

2. The Study of Sailboat Parameters

2.1. The Relationship between the Age of Sailboats and their Pricing

The paper hopes to observe the relationship between the age of the sailboat and its pricing.

In the dataset the paper has, the pricing variable of sailboats has a very large variance. Therefore, the paper considers the papering the following formula to perform a logarithmic transformation on the sailboat pricing data.

$$P_{log} = \log_2 P \tag{1}$$

After carrying out a logarithmic transformation and observing the data, the paper noticed that there may be a certain correlation between the price of sailboats and the year they were produced [1]. The paper made $P_{log} - Y$ boxplots for sailboat pricing and production year, as shown in Fig 1 and 2. Of course, the paper's charts are divided into monohull sailboats and catamarans.

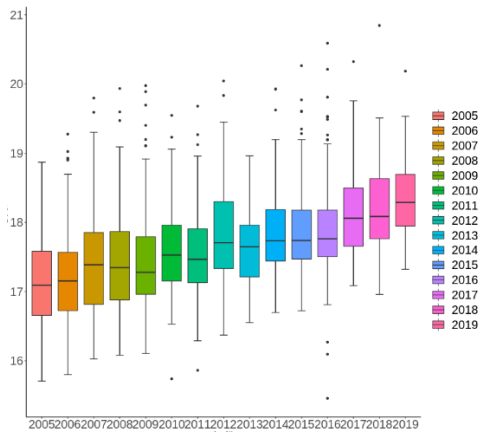

Fig 1: monohulled sailboats boxplot

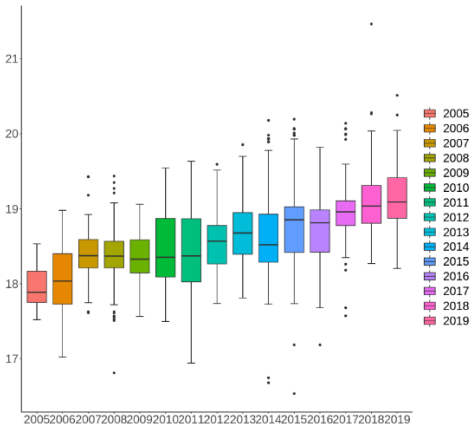

Fig 2: catamarans boxplot

Observing the above two box plots, the paper could see intuitively that the paper's hypothesis was somewhat correct. At the same time, the paper could see that there were some sailboat prices that were outliers in certain years, indicating that there were other factors that influenced sailboat pricing besides year. Therefore, the paper decided to consider building a multivariate function to describe the relationship between the data.

2.2. Correlation Analysis between Variables

Considering the large amount of data the paper have and the unclear correlation between the data, the paper first calculated the Pearson correlation coefficient between the data and created a correlation heatmap based on the coefficient [2-4], as shown in Fig 3.

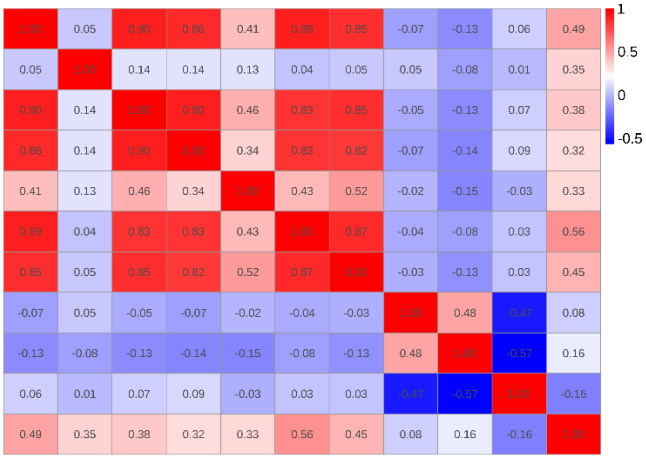

Fig 3: Pearson Corheatmap

Note: The variables in the image are listed from left to right: Length, Year, LWL, Beam, Draft, Displacement, Sail Area, Average Cargo Throughput, GDP per capita, The total logistics cost as a percentage of GDP, Listing Price

From the correlation heatmap, it can be seen that there is a strong correlation between some variables, such as Length and LWL, Beam, Displacement, and Sail Area. This indicates that the paper can approximately treat these variables as linearly correlated with each other. From this perspective, the paper can the use one of these indicators instead of the others for subsequent model building [5-7].

In addition, the paper can also see from the charts that there is very little correlation between all the variables and price, indicating that each indicator alone does not show a good linear relationship with price [8].

3. Multivariate Linear Function Model for Sailboat Pricing

After obtaining the above results, the paper presents the form of the multivariate linear function for sail- boat pricing and characteristics that the paper needs as the objective function to be solved, as follows.

$$P = a_1 L + a_2 D + a_3 Y + a_4 ACT + a_5 TLP + a_6 D \quad P_{pc} + k \tag{2}$$

3.1. Establishment of Sailboat Pricing-Characteristic Function

After data processing, the dimension of the paper's indicators was reduced from 10 to 6. Next, the paper will first establish a model for monohull sailboats.

The paper first provides a statistical definition of the principal components in 6 indicators [9].

Assuming that $c_1, c_2, c_3, c_4, c_5, c_6$ represent the weights of $x_1, x_2, x_3, x_4, x_5, x_6$ respectively. Let $X_1, X_2, X_3, X_4, X_5, X_6$ denote the random variables with sample observations $x_1, x_2, x_3, x_4, x_5, x_6$ respectively. If the paper can find $c_1, c_2, c_3, c_4, c_5, c_6, x_1, x_2, x_3, x_4, x_5, x_6$ such that the value of:

$$\mathrm{Var}(c_1 X_1 + c_2 X_2 + c_3 X_3 + c_4 X_4 + c_5 X_5 + c_6 X_6) \tag{3}$$

reaches its maximum, since variance reflects the degree of difference in data, it means that the paper has captured the maximum difference in these six variables.Obviously, there is a constraint $c_1 + c_2 + c_3 + c_4 + c_5 + c_6$=1, and the paper need to find the optimal solution under this constraint. This solution is a unit vector in a six-dimensional space, representing a "direction", namely the principal component direction.

First,let $y_0(y_1, y_2, y_3, y_4, y_5, y_6)^T$,where $y_i(i = 1, \ldots, 6)$ in this problem represents the pricing of sailboats [10].

Construct design matrix X_0 and matrix Ones with all elements in the first column being 1.

Therefore, the paper can define the least squares regression coefficient matrix.

$$hg_1 = {X_0}^{-1} \cdot y_0 \tag{4}$$

Calculate the Pearson correlation coefficient matrix **R** of the design matrix X_0.

Standardize the design matrix $\boldsymbol{X_0}$ and y_0, let **xd** be the standardized matrix of $\boldsymbol{X_0}$,and **yd** be the standardized matrix of $\boldsymbol{y_0}$.

Use the *pcacov*-function to compute the eigenvectors vec_1, eigenvalues and the proportions of variance explained by each principal component rate for the matrix X. That is:

$$\mathrm{rate} = (ra_1, ra_2, ra_3, ra_4, ra_5, ra_6)^T \tag{5}$$

Obtain the proportion $rate_i$ of ra_i by

$$rate_i = \frac{ra_i}{ra_1 + ra_2 + ra_3 + ra_4 + ra_5 + ra_6} \tag{6}$$

If there exists a small ra_i, it means that the i-th principal component has a lo contribution to the model, and the paper can remove it. If each part is relatively large, i.e., each part's contribution rate is not negligible, the paper can also choose not to remove it. Here the paper chooses not to remove any principal component.

In theory, the sum of all components of the eigenvector should be positive, otherwise it has no practical meaning. Therefore, the paper uses the *remat*-function to construct a matrix L with the same dimension as the eigenvector, whose elements are ±1. Then the paper can modify the eigenvector $\boldsymbol{vec_1}$ to $\boldsymbol{vec_2}$, so that the sum of all components of $\boldsymbol{vec_2}$ is positive.

$$\mathrm{vec}_2 = \mathrm{vec}_1 \cdot \mathrm{F} \tag{7}$$

Then, the paper can calculate the regression coefficients of the principal components.

$$\mathrm{hg}_{21} = (\mathrm{xd} \cdot \mathrm{vec}_2)^{-1}\mathrm{yd} \tag{8}$$

Calculate regression equation coefficients for standardized variables.

$$\mathrm{hg}_{22} = \mathrm{vec}_2 \cdot \mathrm{hg}_{21} \tag{9}$$

The paper substitutes the monohull boat data and calculate:

$$\mathrm{rate} = (33.3602,21.5082,15.2648,14.0375,8.6569,7.1724) \tag{10}$$

The smallest contribution rate to the principal components is also 7%, so the paper does not need to remove any of the components and still choose all six.

Finally, the paper calculates the residual standard errors of the least squares regression analysis and principal component regression analysis, which are $\mathrm{rmse}_1 = 1.1697\mathrm{e}+05, \mathrm{rmse}_2 = 1.1694\mathrm{e}+05$. So $\mathrm{rmse}_2 < \mathrm{rmse}_1$, passes the significance test.

Calculate the coefficients of the original variable regression equation to obtain the final monohull sailboat pricing function. Then, the paper substitutes the variables defined earlier into the equation to obtain the objective function for single-hull sailboats.

$$P_M = 658.981509L_M + 57790.632663D_M + 12482.372183Y_M - 0.000240ACT_M$$
$$-2659.740646TLP_M + 2.349469GDP_{pcM} - 25338353.362482 \tag{11}$$

To test the fitting degree of the function, the paper plotted a scatter graph of the predicted values and the actual values of the function, as shown in Fig 4. The results indicate that, under the assumption of the model, the fitting degree of the pricing function for individual sailboats is quite good.

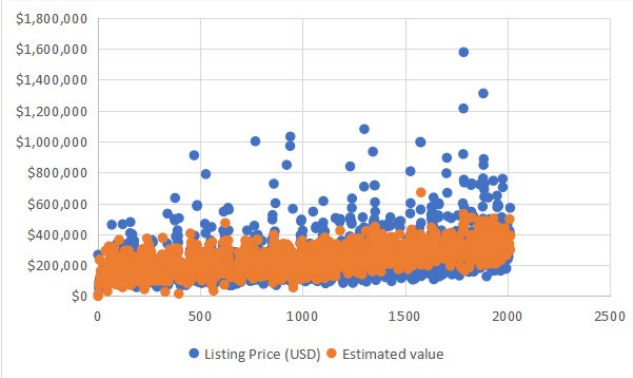

Fig 4: M function check

Similarly, for catamarans, the paper also doesn't need to exclude any principal components. By calculating $rmse_1 = 1.1260e+05, rmse_2 = 1.1255e+055$ with the data of catamarans, so $rmse_2 < rmse_1$ the paper can obtain the pricing function for catamarans. Thus, the paper obtains the objective function for the catamaran:

$$P_C = 34113.304516L_C + 4.207799D_M + 23246.747163Y_M + 0.000225ACT_C$$
$$+3346.960120TLP_C + 1.156529GDP_{pcC} - 47907268.344616 \tag{12}$$

Similarly, the paper plotted a scatter graph of the predicted values and the actual values to test the fitting degree of the function for catamarans, as shown in Fig 5. Visually, the fitting degree of the function for catamarans is also satisfactory.

Fig 5: C function check

3.2. The Impact of Sailboat Variant on Pricing Estimation Accuracy

According to the model assumption, the paper used sailboat length and draft as proxies for sailboat type.

Based on the coefficient features of the function:

$$P_M(f_M) = A_M {f_M}^\top + k_M \tag{13}$$

$$P_C(f_C) = A_C {f_C}^\top + k_C \tag{14}$$

it can be seen that the coefficients of the sailboat length and draft, denoted by L and D respectively, are both positive in the P_M and P_C functions.

3.3. The Impact of Regional Indicators on Sailboat Pricing

Similar to the analysis approach in the previous section, the paper first used the average cargo throughput, the proportion of total logistics costs to GDP, and per capita GDP as proxies for regional characteristics, based on the paper's assumptions.

However, the paper observes from the data that:

$$a_{4M} = -0.000240, a_{5M} = -2659.740646, a_{6M} = 2.349469$$

$$a_{4C} = 0.000225, a_{5C} = 3346.960120, a_{6C} = 1.156529 \tag{15}$$

Obviously, for monohull sailboats, the values of ACT and TLP are negatively correlated with pricing. Conversely, for catamarans, their pricing is negatively correlated with the values of ACT and TLP.

4. Analysis of Sailboat Pricing in the Hong Kong(SAR)

4.1. Sailboat pricing function in Hong Kong(SAR)

Firstly, the paper selected 5 types of monohull sailboats for analysis in the sample, as shown in the Table.1.

Table 1. Chosen Variant M

Group	Variant	Year
1	Beneteau Oceanis 40	2023
2	Beneteau Oceanis 51.1	2018
3	Beneteau Oceanis 43	2012
4	Beneteau Oceanis 38	2014
5	J Boats 111	2012

The paper assumes that the pricing model for sailboats in the Hong Kong market follows the sailboat pricing function. Based on the model, the monohull sailboat pricing function is:

$$P_M = 658.981509L_M + 57790.632663D_M + 12482.372183Y_M - 0.000240ACT_M$$
$$-2659.740646TLP_M + 2.349469GDP_{pcM} - 25338353.362482 \quad (16)$$

Among them, ACT, TLP, and GDP_{pc} are three indicators related to the region. The paper found that the average cargo throughput in Hong Kong is 2,138,000 tons, the per capita GDP is $48,400, and the average proportion of total logistics costs to GDP is 13.7%. By plugging these three data into the above equation, the paper obtained the pricing function of a single sailboat in the Hong Kong market as follows:

$$P_{HKM} = -25262000 + 658.981509L_{HKM} + 57790.632663D_{HKM} + 12482.372183Y_{HKM} \quad (17)$$

Now the paper substitutes those 5 types of sailboats into the function to obtain their predicted values, and collect the listed prices under the Hong Kong market as shown in the Table 2:

Table 2. Predictive Value and Actual Value M

Group	1	2	3	4	5
Predictive Value	$377,799	$361,355	$218,197	$216,173	$291,602
Actual Value	$386,587	$373,556	$220,000	$230,000	$290,000

Compare them and create a clustered bar chart. As shown in Fig 6:

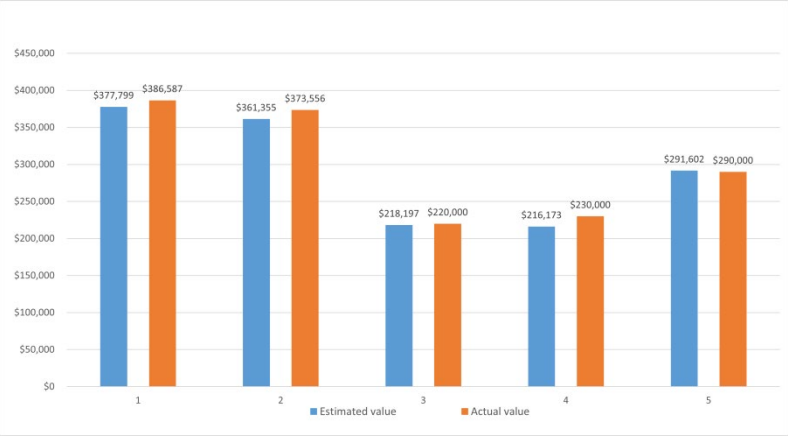

Fig 6: M Hong Kong

The paper found that the predicted values are very consistent with the actual values. Therefore, the paper assumes that the model proposed is applicable.

Next, the paper selected 4 types of double-hull sailboats as shown in the Table 3:

Table 3. Chosen Variant M

Group	Variant	Year
1	Lagoon 46	2019
2	Lagoon 450	2016
3	Lagoon 450	2017
4	Lagoon 40	2023

The pricing function of catamarans is:

$$P_C = 34113.304516L_C + 4.207799D_M + 23246.747163Y_M + 0.000225ACT_C$$
$$+3346.960120TLP_C + 1.156529GDP_{pcC} - 47907268.344616 \quad (18)$$

Similarly, by substituting the regional indicators data of Hong Kong into the equation, the paper obtained the pricing function for catamarans in the Hong Kong market as follows:

$$P_{HKC} = -47805000 + 23246.747163L_{HKC} + 4.207799D_{HKC} + 34113.304516Y_{HKC} \quad (19)$$

Substituting the sailboat parameters into the equation, the paper obtains the sailboat pricing predicted values. At the same time, the paper also found the corresponding listing prices for these four types of sailboats in the Hong Kong market, as shown in the Table 4:

Table 4: Predictive Value and Actual Value C

Group	1	2	3	4
Predictive Value	$699454	$622892	$646139	$536591
Actual Value	$716932	$560000	$50000	$290,000

Make a clustered bar chart as shown in Fig 7:

Fig 7: M Hong Kong

Thus, under the assumption, the paper obtains a model that is applicable to sailboat pricing in the Hong Kong market.

4.2. The impact of regional indicators on sailboat pricing in Hong Kong(SAR)

From the analysis of the impact of regional indicators on sailboat pricing, the paper knows that the impact of regions is different for monohull sailboats and catamarans, this difference is mainly reflected in the constant term:

$$K_{HKM} = -25262000, K_{HKC} = -47805000 \quad (20)$$

Under the model assumption, the regional impact indicators in the Hong Kong regional sailboat pricing function are degraded to a constant term, so the regional effect in the Hong Kong region has different impacts on the pricing of the two types of sailboats.

5. Conclusions

This paper utilizes the Pearson correlation coefficient to select the impact indicators for sailboat pricing, and then, based on principal component analysis, derives the pricing objective functions for two categories of sailboats: monohulls and catamarans. The work undertaken in this paper, to some extent, addresses the existing gap in the pricing of second-hand sailboats and introduces a new perspective for research in the field of pricing. However, due to the uniqueness of sailboat pricing, there are still areas for improvement in this study. The research in this paper is built on certain assumptions, such as the absence of brand effects in sailboat pricing and the possibility of substituting non-numeric indicators with numeric ones. Consequently, if real-world conditions do not align with these assumptions, the predicted prices obtained using the functions in this paper may exhibit significant deviations from actual prices. In future research, not only will the search for better pricing models be necessary, but also factors like brand effects need to be considered during modeling.

References

[1] Xiao Rongge, Zhuang Qi, Jin Shuaishuai, Liu Bo, Liu Guoqing. Evaluation of influencing factors of pipeline wax deposition strength based on principal component analysis [J]. Petroleum Science and Technology,2023,41(6).

[2] Xu Changjin, LiuZixin, Liao Maoxin, Li Peiluan, Xiao Qimei, Yuan Shuai. Fractional-order bidirectional associate memory (BAM) neural networks with multiple delays: The case of Hopf bifurcation [J]. Mathematics and Computers in Simulation,2021,182.

[3] Xiao Rongge, Zhuang Qi,Jin Shuaishuai, Liu Bo, Liu Guoqing. Evaluation of influencing factors of pipeline wax deposition strength based on principal component analysis [J]. Petroleum Science and Technology,2023,41(6).

[4] VillaAleman Eliel, Christian Jonathan H, Darvin Jason R, Foley Bryan J, Dick Don D,FallinBrent, Fessler Kimberly A S. Diffuse Reflectance Spectroscopy and Principal Component Analysis to Retrospectively Determine Production History of Plutonium Dioxide.[J]. Applied spectroscopy,2023.

[5] Beynon Malcolm J., Jones Paul, Pickernell David. Evaluating EU-Region level innovation readiness: A longitudinal analysis using principal component analysis and a constellation graph index approach [J]. Journal of Business Research,2023,159.

[6] Larson, R., & Edwards, B. (2017). Calculus. Cengage Learning.

[7] Ross, S. M. (2014). Introduction to probability models. Academic press.

[8] Li Shengshi, Zou Yonghua, Wang Guanjun, Lin Cong. Infrared and Visible Image Fusion Method Based on a Principal Component Analysis Network and Image Pyramid [J]. Remote Sensing,2023,15(3).

[9] Ahmed Hassen Youssef, Engy Saeed Mohamed, Shereen Hamdy Abdel Latif. Handlingmulti-collinearity using principal component analysis with the panel data model [J]. EUREKA: Physics and Engineering,2023(1)

[10] Shoukui Si and Xiqing Sun. Mathematical Modeling Algorithms and Applications. Beijing: National Defense Industry Press,2022.