Proceedings of the 2nd International Conference on Mathematical Physics and Computational Simulation
DOI: 10.54254/2753-8818/38/20240641

The sailing boat price study based on two-step multiple regression analysis

Jingwen Zhang

Haidian Foreign Language Academy, Beijing, 100095, China

zhangjw2006@126.com

Abstract. This paper presents a novel approach to the valuation of boats through the utilization of a two-step multivariate linear regression analysis. Initially, a comprehensive data set comprising multiple boat pricing records is examined to estimate local demand and utility functions. This estimation is guided by prior experiences and theoretical frameworks in marine economics. In the initial stage of the regression analysis, the aforementioned functions are employed to establish localized demand and preference profiles for individual boats. Subsequently, a second regression analysis is conducted, leveraging the aforementioned profiles to predict the market prices of boats with greater accuracy. The integration of a dual-regression model enables this study to refine the predictions of boat valuations while elucidating the complex dynamics between local market conditions, consumer preferences, and boat characteristics. The methodology and findings of this paper provide enhanced insights for stakeholders in the marine market, offering a robust tool for assessing boat prices in varied market environments.

Keywords: Boat Valuation, Demand Function Estimation, Two-step Multivariate Linear Regression, Utility Function Evaluation.

1. Introduction

The valuation of boats is influenced by a range of factors, including their physical attributes, maintenance history, and market conditions. This paper aims to examine the significant factors that affect boat pricing. To this end, a Multivariate Linear Regression (MLR) model is employed to identify and quantify the key variables. Physical characteristics such as size, age, and type of boat are primary aspects that influence prices [1-3]. For instance, larger boats typically command higher prices due to their enhanced amenities and capacity, while newer models might attract premium pricing due to advanced features. Moreover, the condition of the boat, including the extent of upkeep and the recentness of upgrades, is of great consequence in determining its market value. Economic factors also exert a significant influence on boat pricing. Fluctuations in the economy can alter consumer spending power, which in turn affects demand and consequently the prices of luxury items such as boats. Tsolakis et al developed a regression error correction model based on a theoretical supply and demand model, highlighting how interest rates, charter rates, and order size impact ship prices have a significant impact on shipbuilding costs and expected profits on ship prices [4]. Engstrom et al enhanced the Pissarides-Weber method using registration measures as consumption proxies in Sweden and Finland to assess tax evasion's effect on ship pricing [5]. Jiao, Y et al analyzed nearly 100,000

DOI: 10.54254/2753-8818/38/20240641

social media travel reviews to determine how tourists' perceptions affect cruise values, finding notable differences in perceptions among cruise types but not in tangible qualities [6]. Zaw et al estimated the external economic value of Icelandic whale-watching boats, showing that switching to electric boats reduces costs and environmental pollution [7]. Jessica G. et al concluded that the shipbuilding industry's price elasticity depends on the reliability and quality of the ships produced [8]. These insights set the stage for the two-step multiple linear regression model used to analyze sailboat pricing in this paper, which is further discussed in the next sections. The remainder of the paper is organized in the following manner: Chapter 2 presents the sailboat parameters used for the first multiple regression analysis, and Chapter 3 develops the multiple linear regression mathematical model used for the second step multiple regression analysis of sailboat pricing. The sailboat pricing analysis is then given in Chapter 4. Finally, comments and suggestions for future research are given in Chapter 5.

2. First-Step Multiple Regression Analysis on Sailboat Parameters

The data set for the paper encompasses a diverse range of parameters, including type, boat class, year of manufacture, condition, length in feet, beam width in feet, dry weight in pounds, hull material, fuel type, number of engines, total horsepower, maximum and minimum engine years, engine category, seller ID, city, and the dates and years when the listings were created. The data in this paper is sourced from the following website: https://www.kaggle.com/.

2.1. Demand Function Evaluation

The demand function essentially captures how quantity demanded varies with price and other factors. For sailboats, the demand can be hypothesized to depend primarily on: price of the boat, boat class, length and beam, etc. The demand function can be modeled as follows:

$$\mathcal{D}(p,x) = \alpha_0 + \alpha_1 p + \alpha_2 \mathcal{C} + \alpha_3 \mathcal{L} + \alpha_4 \mathcal{B} + \alpha_5 \mathcal{H} + \alpha_6 \mathcal{Y} + \epsilon \quad (1)$$

where, $\mathcal{D}$ represents the demand function, x denotes the vector of other influencing variables, and $\mathcal{C}, \mathcal{L}, \mathcal{B}, \mathcal{H}$ and $\mathcal{Y}$ highlight the specific categories or key features like class, length, beam, horsepower, and age of the boat, respectively. ϵ stands for residual.

2.2. Utility Function Evaluation

The utility function will evaluate the subjective satisfaction a buyer obtains from purchasing a sailboat, which depends on attributes that directly affect their experience:

$$\mathcal{U}(x) = \beta_0 + \beta_1 \mathcal{C} + \beta_2 \mathcal{N} + \beta_3 \mathcal{L} + \beta_4 \mathcal{B} + \beta_5 \mathcal{W} + \beta_6 \mathcal{M} + \beta_7 \mathcal{F} + \beta_9 \mathcal{E} + \beta_9 \mathcal{H} + \epsilon \quad (2)$$

In this utility function, $\mathcal{U}$ emphasizes the utility derived, x again represents the vector of boat attributes, and symbols like $\mathcal{N}, \mathcal{W}, \mathcal{M}, \mathcal{F}, \mathcal{E}$ stand for condition, dry weight, hull material, fuel type, and number of engines, which are all presented in a stylized format to distinguish them.

3. Second-Step Multiple Regression Analysis for Sailboat Pricing

3.1. Explicit Multiple Linear Regression

The basic expression for sailboat pricing linear multivariate regression using the least squares method is:

$$\min Q^b(\beta) = \| \varepsilon \|^2 = \| y - X\beta_s \|^2 = (y - X\beta_s)^{\mathrm{T}} (y - X\beta_s) \quad (3)$$

where Q^b is the sum of squares of the residuals ε, y is the dependent variable (price of boat) of the regression, and β_s represents the parameters of the regression independent variables. The parameters can be calculated as follows:

$$\widehat{\beta}_s = (X^{\mathrm{T}} X)^{-1} (X^{\mathrm{T}} y) \quad (4)$$

DOI: 10.54254/2753-8818/38/20240641

The objective function of ridge regression adds a regularization term to the base of ordinary linear regression. For example, in ridge regression corresponding to the L^2 norm:

$$\min \boldsymbol{Q^b(\beta_\lambda)} = \| \boldsymbol{\varepsilon} \|^2 = \| \boldsymbol{y - X\beta_\lambda} \|^2 + \boldsymbol{\lambda} \| \boldsymbol{\beta_\lambda} \|^2$$
$$\Leftrightarrow \text{argmin } \| \boldsymbol{y - X\beta_\lambda} \|^2 \text{ s.t. } \sum \beta_{\lambda,j}^2 \leq S_\lambda \tag{5}$$

Solving the above equation yields the ridge estimate of $\boldsymbol{\beta_\lambda}$:

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} + \boldsymbol{\lambda I})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} \tag{6}$$

Similar to ridge regression, Lasso regression involves adding an L^1 norm to the objective function $\boldsymbol{Q^b(\beta_L)}$:

$$\min \boldsymbol{Q^b(\beta_L)} = \| \boldsymbol{\varepsilon} \|^2 = \| \boldsymbol{y - X\beta_L} \|^2 + \boldsymbol{\lambda} \| \boldsymbol{\beta_L} \|^2$$
$$\Leftrightarrow \text{argmin } \| \boldsymbol{y - X\beta_L} \|^2 \text{ s.t. } \sum \beta_{L,j}^2 \leq S_L \tag{7}$$

At this time, the coefficient vector usually does not have a simple explicit expression, because the presence of the L^1 regularization term complicates the optimization problem, leaving it without an analytical solution. However, various numerical optimization algorithms (such as coordinate descent, gradient descent, etc.) can be used to approximate the solution of the Lasso regression problem and obtain a numerical solution for the coefficient vector.

3.2. Implicit Multiple Linear Regression

Gradient Boosting Decision Tree (GBDT) is a boosting algorithm whose base classifier usually uses Classification and Regression Tree (CART tree) for fitting. Its model $\boldsymbol{F}$ is defined as an addition model:

$$\boldsymbol{F(x,\omega)} = \sum_{\boldsymbol{t=0}}^{\boldsymbol{T}} \boldsymbol{\alpha_t\, h_t(x,\omega_t)} = \sum_{\boldsymbol{t=0}}^{\boldsymbol{T}} \boldsymbol{f_t(x,\omega_t)} \tag{8}$$

where $\boldsymbol{x}$ represents the input sample, $\boldsymbol{h}$ stands for the CART tree, $\boldsymbol{\omega}$ denotes the parameters of the tree, and $\boldsymbol{\alpha}$ is the weight of each tree. The iterative formula for solving the parameters of multiple linear regression based on gradient descent is as follows:

$$\boldsymbol{y = \beta^{\mathrm{T}}X \Rightarrow J(\beta)} = \frac{1}{2m}(\boldsymbol{X\beta - y})^{\mathrm{T}}(\boldsymbol{X\beta - y}) \tag{9}$$

and $\widehat{\boldsymbol{\beta}}$ can be calculated as

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta} - \boldsymbol{\alpha_l}\frac{1}{m}\boldsymbol{X}^{\mathbf{T}}(\boldsymbol{X\beta - y}) \tag{10}$$

where $\boldsymbol{\alpha_l}$ is the learning rate. Support Vector Regression (SVR) is an important application branch of Support Vector Machines (SVM). SVR regression aims to find a regression plane that minimizes the distance from all data points in a given set to that plane. The computation process of a Back propagation (BP) neural network consists of a forward computation phase and a backward computation phase. In the forward propagation phase, the input pattern is processed layer by layer from the input layer through the hidden units and transmitted to the output layer. The state of neurons in each layer only affects the state of neurons in the next layer. If the desired output cannot be achieved at the output layer, the error signal is back propagated along the original connection path, and the weights of the neurons are adjusted to minimize the error signal.

$$\boldsymbol{y} = \text{argmax}_{c_j} \sum_{x_i \in N_k(x)} I(y_i = c_j), i = 1,2,\cdots N; j = 1,2,\cdots K \tag{11}$$

DOI: 10.54254/2753-8818/38/20240641

where I is an indicator function, which takes the value of 1 when $y_i = c_j$, and 0 otherwise. N represents the number of samples in the neighborhood of x, c_j denotes a certain class, and K stands for the number of classes.

4. Analysis of Sailboat Pricing

4.1. First-Step Multiple Regression Analysis

Both the demand function and the utility function are first calculated according to the empirical formula, and then the corrected demand function and utility function are obtained according to the tabular data.

4.2. Second-Step Multiple Regression Analysis

Figure 1 then presents the different regression results obtained by using explicit linear regression and implicit linear regression, as well as regression prediction graphs.

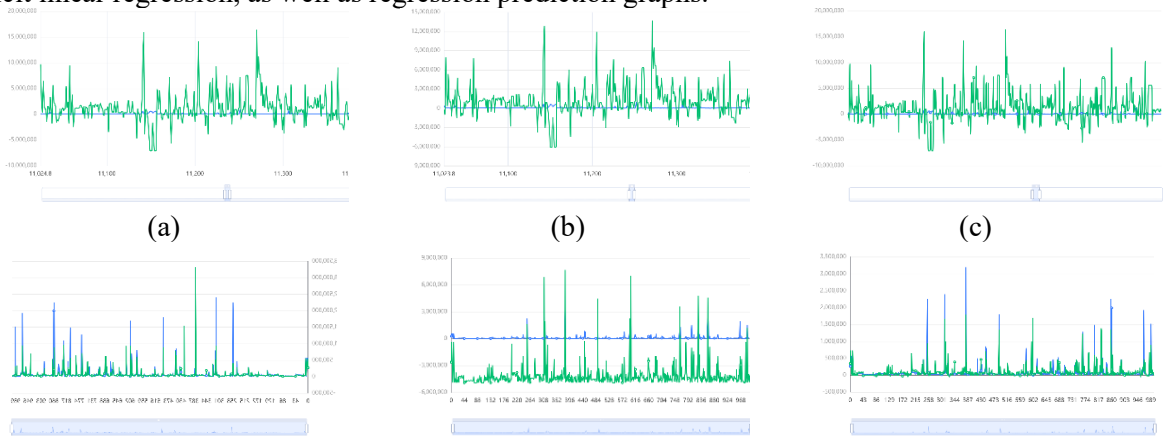

Figure 1. Regression result (a. Least squares regression; b. Ridge regression; c. Lasso regression; d. GBDT regression; e. Support vector regression; f. BPNN regression.)

The least squares regression analysis yielded a result for the F test, with a significance P value of 0.002. This indicates that the null hypothesis that the regression coefficient is 0 is rejected, thereby demonstrating that the model meets the requisite requirements. With regard to the issue of collinearity of variables, the VIF values are all below 10, indicating that the model is free from multicollinearity and that it has been constructed in an appropriate manner. The model is defined by the following formula:

$$\boldsymbol{y} = 7358481.68 - 4148294.79\boldsymbol{\mathcal{C}} + 3997.95\boldsymbol{\mathcal{L}} - 19963.83\boldsymbol{\mathcal{B}} + 1799734.54\boldsymbol{\mathcal{H}} + 1158008.16\boldsymbol{\mathcal{Y}} - 943664.73\boldsymbol{\mathcal{M}} - 630274.11\boldsymbol{\mathcal{E}} \quad (12)$$

The results of ridge regression indicate that, based on the significance of the F-test, the P-value is 0.002, which shows significance at the level, rejecting the null hypothesis. This suggests that there is a regression relationship between the independent variable and the dependent variable. Concurrently, the model's goodness of fit, as indicated by the R^2, is 0.001, which is a relatively poor performance. The formula of the model is as follows:

$$\boldsymbol{y} = 6233218.70 + 3396280.315\boldsymbol{\mathcal{C}} + 3069.62\,\boldsymbol{\mathcal{L}} - 12290.53\boldsymbol{\mathcal{B}} - 1475145.21\boldsymbol{\mathcal{H}} - 951781.43\boldsymbol{\mathcal{Y}} - 775040.69\boldsymbol{\mathcal{M}} - 547873.77\boldsymbol{\mathcal{E}} \quad (13)$$

The ridge trace map visually formalizes the situation when the standardization coefficients of each independent variable of this model tend to be stable, as shown in Figure 2.

DOI: 10.54254/2753-8818/38/20240641

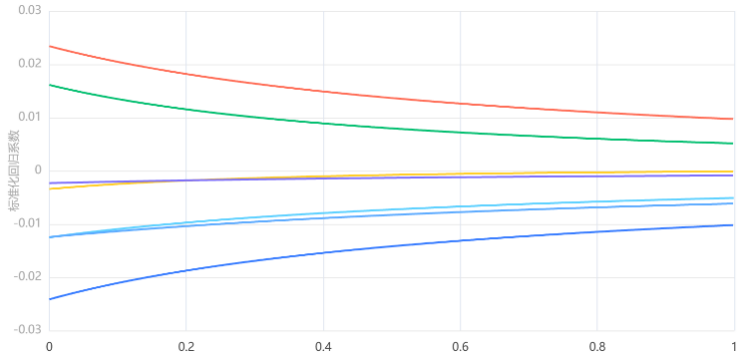

Figure 2. Ridge Regression-Ridge Trace Map

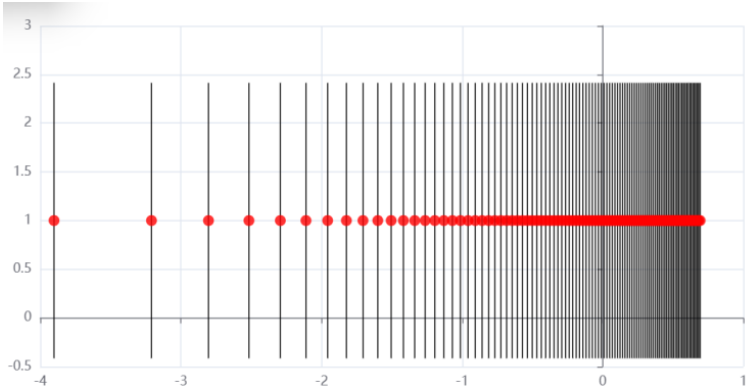

Figure 3. Lasso Regression Cross Validation Graph

Figure 3 visualizes the use of cross-validation to select λ values. Ordinate: model mean square error. Abscissa: the pair value of λ. To minimize the mean square error λ=2.0, log(λ)=0.693. The results of Lasso regression show no variables are deleted. The formula of the model is shown below:

$$\boldsymbol{y} = 7358358.49 - 4148278.58\boldsymbol{C} + 3997.938\,\boldsymbol{L} - 19963.88\boldsymbol{B} + 1799732.62\,\boldsymbol{H} - 1157909.53\,\boldsymbol{Y} - 943662.43\boldsymbol{M} - 630272.60\boldsymbol{E} \quad (14)$$

Table 1 lists the prediction evaluation indicators of cross-validation set, training set and test set, and measures the prediction effect of GBDT, SVR, BPNN through quantitative indicators. Among these, the hyperparameters can be adjusted continuously through the evaluation index of the cross-validation set to obtain a reliable and stable model. The MSE (mean square error): The expected value of the squared difference between the predicted value and the actual value. The smaller the value, the higher the model accuracy. The RMSE (root mean square error): The square root of MSE. The smaller the value, the higher the model accuracy. The MAE (Mean Absolute Error): The mean of the absolute error, which reflects the actual situation of the predicted error. The smaller the value, the higher the model accuracy. The MAPE (Mean Absolute Percentage error): A variation of MAE, which is a percentage value. The smaller the value, the higher the model accuracy. R^2: The closer the predicted value is to 1, the more accurate the model is when compared with the mean only. This table indicates that implicit regression results have a poor predictive effect.

DOI: 10.54254/2753-8818/38/20240641

Table 1. The Prediction Evaluation Indicators of Cross-Validation Set

	MSE	RMSE	MAE	MAPE	R^2
GBDT Test set	2579410506.001	50787.897	21518.705	42.797	0.91
GBDT Training set	17809795232302818	133453344.778	1989323.937	4580.458	0
SVR Test set	23192046569606.145	4815812.14	4784035.514	102.05	-810.133
SVR Training set	17844180351006758	133582110.894	6183044.151	152.412	-0.002
BPNN Test set	18622599709.317	136464.646	43562.113	231.623	0.349
BPNN Training set	17809715848157664	133453047.354	2023866.468	4542.859	0

5. Conclusion

From the above analysis, it can be found that the linear regression results perform well while the implicit regression results have a poor predictive effect. This suggests that there is an important relationship between the ship price and the number of engines, class, length, beam, horsepower, and age of the boat. By analyzing these variables with multiple regression analysis, this paper provides the insights into the dynamics that govern boat prices, offers valuable guidance for stakeholders in the maritime market. In future work, more complex utility functions such as cobb-Douglas function or CES (constant elastic substitution) function or even Leontief utility function can be considered for more complex estimation to make more in-depth analysis of ship prices.

References

[1] Chen, Y., Li, Z., Jia, D. (2024) The Sailing Boat Price Study Based on Principal Component Regression Analysis. Highlights in Business, Economics and Management, 25, 189-196.

[2] Zhou, H., Qian, J., Sun, F. (2024) A Study on the Pricing of Used Sailboats: Utilizing Random Forest Models to Analyze the Impact of Multiple Factors. In 2024 IEEE 3rd International Conference on Electrical Engineering, Big Data and Algorithms (EEBDA). 1708-1713.

[3] Zhu, H., Cui, J. (2023) Secondhand Sailboat Listing Price Prediction Model Based on GBDT Algorithm. In Proceedings of the 2023 4th International Conference on Machine Learning and Computer Application. 536-540.

[4] Tsolakis, S.D., Cridland, C., Haralambides, H.E. (2003). Econometric Modelling of Second-Hand Ship prices. Maritime Economics & Logistics, 5(4), 347-377.

[5] Engström, P., Hagen, J., Johansson, E. (2023). Estimating tax noncompliance among the self-employed—evidence from pleasure boat registers. Small Business Economics, 61(4), 1747-1771.

[6] Jiao, Y., Lau, Y.Y., Gao, J. (2024). Exploring the Factors Affecting Cruise Passengers' Perceptions of Value for Money Expressed in Online Reviews. Humanities and Social Sciences Communications, 11(1), 1-11.

[7] Win, Z.M., Cook, D., Davíðsdóttir, B. (2023). A Comparison of the Economic Value of Fuel Externalities from Whale Watching Vessels: Electric and Diesel Fueled Boats in Iceland. Ocean & Coastal Management, 239, 106588.

[8] Göransson, J., Andersson, H. (2023). Factors that Make Public Transport Systems Attractive: A Review of Travel Preferences and Travel Mode Choices. European Transport Research Review, 15(1), 32.