Predicting the "Bet Performer" Ship Valuation: Price Prediction through Regression Model Analysis

ISA 510: Probability and Statistics for Data Analytics
Submitted by: Doha Zaky
Fall 2025

Executive Summary

This report is based on a case study published by Harvard Business Review that is called "Compass Maritime Case Study", the mentioned case study aims to estimates the market value and the optimal bidding price for a capsize bulk carrier Bet Performer using regression analysis on recent comparable ship transactions of previous sales (was provided as part of the case study tables and index sheets). The analysis uses a market-based approach to predict the ship valuation.

While doing the analysis, we cleaned the dataset of the ship sales, explored key relationships between sales price and other related variables, estimated a simple regression model between the sales price and the age to capture depreciation, and estimated a multiple regression model between the sales price and all other variables such as (Age, DWT, BDI) to capture the vessel size and the market conditions in our sales price prediction as part of the ship valuation required by this case study. Finally, generated predictions with two scenarios based on the market: conservative market, or hot market situations.

Case Study Background and Prediction Objectives

Compass Maritime Services advised a client considering buying the "Bet Performer", an 11- year-old capsize ship. The case provides a historical ship sales data and market indicators, enabling valuation through their comparable analysis using regression-based models. The goal of this analysis is to predict and estimate a fair market value for this ship based on the given observations or potential price drivers, such as the following:

- Age: this variable is a representative of the depreciation value of the ship or its potential remaining life.
- DWT: this variable is for potential earnings capacity or return on investment (ROI) for the ship.
- BDI: is the market index for the capsize ships, this variable is a representative of the market index.

Data Preparation, Cleaning, and Scaled Variables

The case study provides the dataset, along with tables consisting of capsize ship historical sales transactions from January 2007 through May 2008. The dataset includes ship characteristics and market index values at sale. The analysis uses the following key variables:

- Price: sale prices (in millions).
- Age: age at scale (in years).
- DWT: deadweight tonnage.
- BDI: capsize index scaled to thousands.

Scaling the above numeric variables (BDI) does not change model fit or future predictions. It was used to improve the regression model interpretability and reduces numeric imbalance across other variables in the estimated regression equation.

An additional cleaning step was done in the given excel sheet that had muti- line headers and inconsistent spacing. The sheet cleaning included removing unnecessary tabs, renaming the headers to ensure they refer reliably to the variables without key error (e.g. Price, Age, DWT, BDI). The features scaling of the the BDI/1000 was done to make the coefficients easier to interpret and improves numeric stability in the regression equation. The final data cleaning step was done in python to drop any missing values to ensure consistent estimations.

Exploratory Data Analysis (EDA)

A correlation matrix and a heatmap were generated for all variables for the: Price, Age, DWT, BDI. The key interpretations generally in the analysis are:

- Price Vs. Age: there is a strong negative relationship, this supports the depreciation model, the older the ship the less the price, by the simple regression.
- Price Vs. DWT: there is a positive relationship (the larger the ship, the higher the price).
- Price Vs. BDI: there is a positive relationship (the stronger the market, the higher the price).

*Table (1). The Correlation Matrix of Ship Valuation Variables*
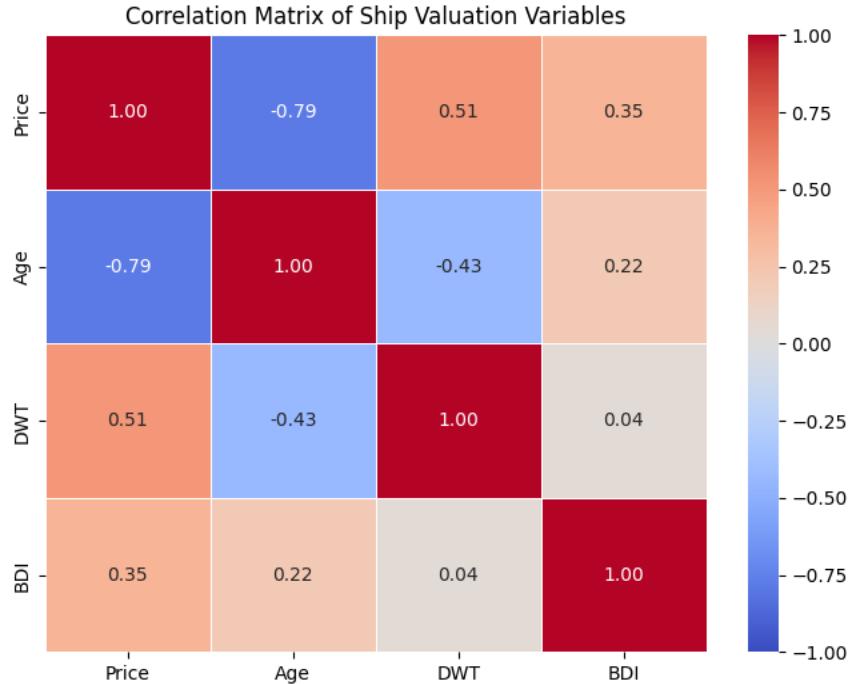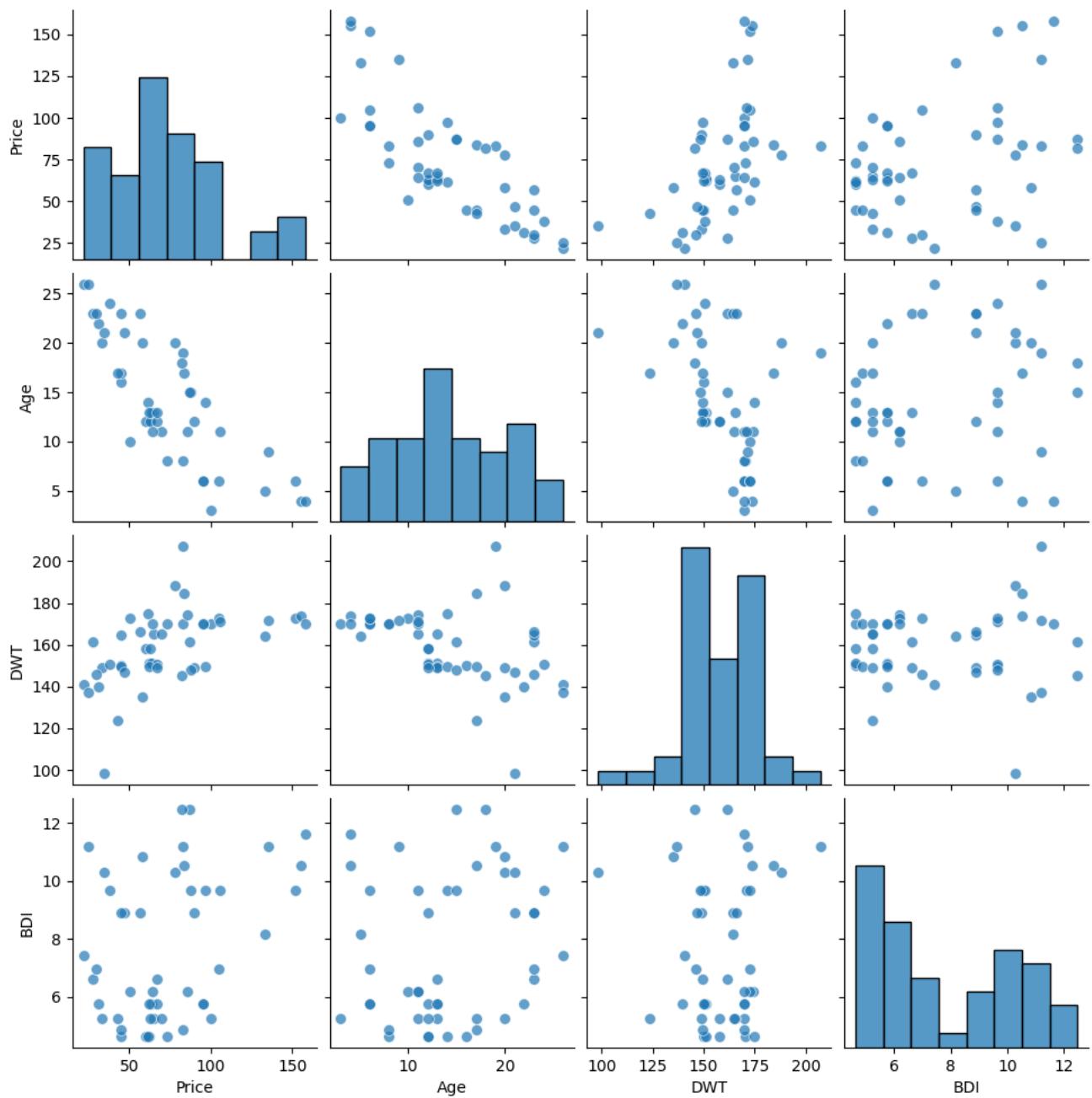
Table (2). The Scatterplot Matrix of Ship Valuation Variables
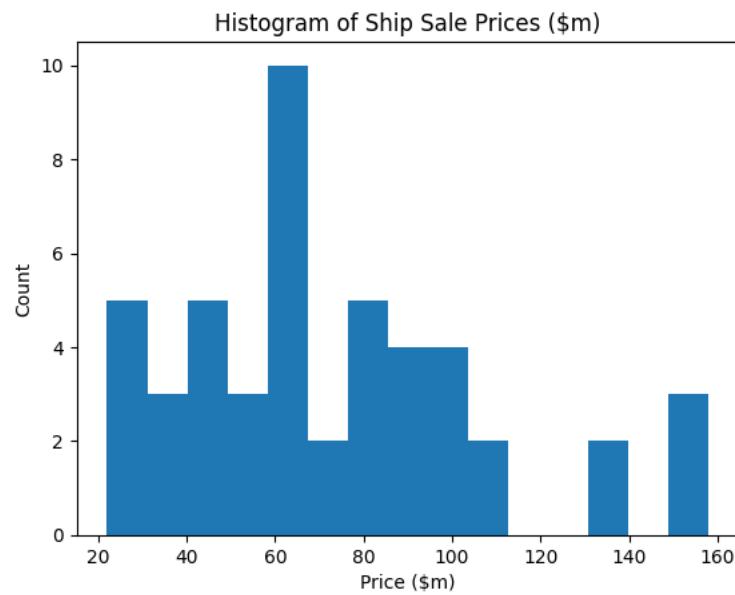
Scatterplot Matrix of Ship Valuation Variables

Data IID Check

Before conducting the regression analysis, a data sample IID check was performed to ensure the data is IID, to ensure the integrity of the observations. A histogram was built to represent the historical sales Price confirming that the data is IID. This step helps preventing any future bias and makes the prediction price more reliable and accurate, as shown in the below table.

Table (3). The Histogram of Ship Sal Prices ($m)

Histogram of Ship Sale Prices ($m)

Co-linearity Check

Table (4). Price Vs. Age

Price vs Age
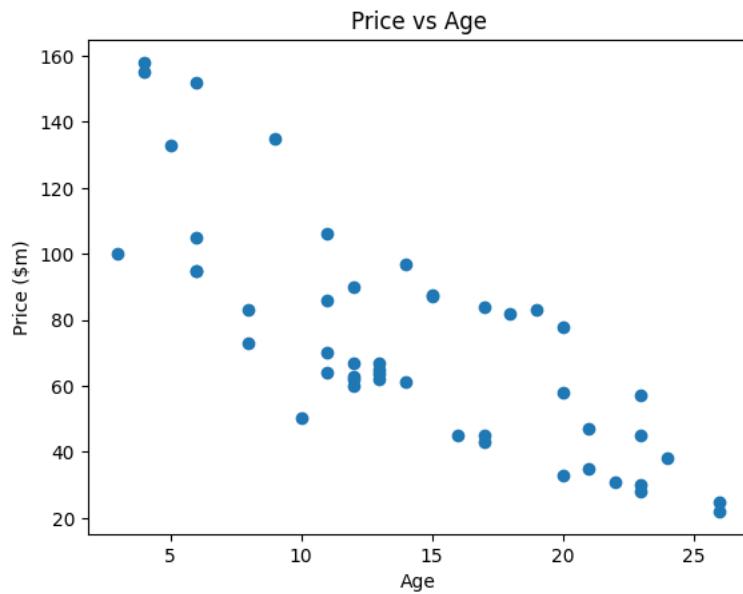
Table (5). Price Vs. DWT

Table (6). Price Vs. BDI
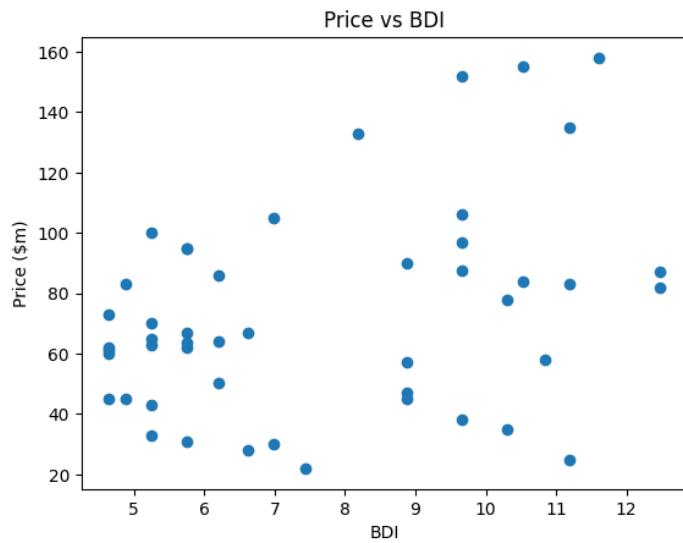
Table (7). Comparison Summary Table

	Price	Age	DWT	BDI
Price	1.000000	-0.787491	0.514805	0.352348
Age	-0.787491	1.000000	-0.431264	0.217360
DWT	0.514805	-0.431264	1.000000	0.042766
BDI	0.352348	0.217360	0.042766	1.000000

Regression Modeling Approach

1- Simple Regression: Price = f (Age)

A simple regression model was initially used to capture the relationship between the Price and the Age. The purpose of this model was to isolate depreciation. In the shipping markets, Age is a powerful single predictor

for the ship price. So that the simple regression model will help us understand and provide an intuitive baseline for the model and the future predication. The model prediction summary generates the regression line, with the confidence and prediction intervals, as shown below:

Table (8). Simple Regression Results Summary

```
                        OLS Regression Results
==============================================================================
Dep. Variable:                  Price   R-squared:                       0.620
Model:                            OLS   Adj. R-squared:                  0.612
Method:                 Least Squares   F-statistic:                     75.10
Date:                Tue, 23 Dec 2025   Prob (F-statistic):           3.15e-11
Time:                        22:47:46   Log-Likelihood:                -213.49
No. Observations:                  48   AIC:                             431.0
Df Residuals:                      46   BIC:                             434.7
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         133.1295      7.583     17.556      0.000     117.865     148.394
Age            -4.2165      0.487     -8.666      0.000      -5.196      -3.237
==============================================================================
Omnibus:                        3.980   Durbin-Watson:                   0.585
Prob(Omnibus):                  0.137   Jarque-Bera (JB):                3.225
Skew:                           0.518   Prob(JB):                        0.199
Kurtosis:                       2.265   Cond. No.                         38.9
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
        mean    mean_se  mean_ci_lower  mean_ci_upper  obs_ci_lower  \
0  86.747791   3.438383      79.826685      93.668896     43.682758

    obs_ci_upper
0     129.812823
Estimated Regression Equation:
Price = 133.13 + (-4.22 × Age)
```

SR Model Diagnostics

A model diagnostic was used to support whether the regression model is a defensible or not based on the residuals' behavior and leverage points. Both diagnostics supported the simple regression model is not reliable, as shown in the below tables:

Table (9). Residuals Vs. Fitted Plot
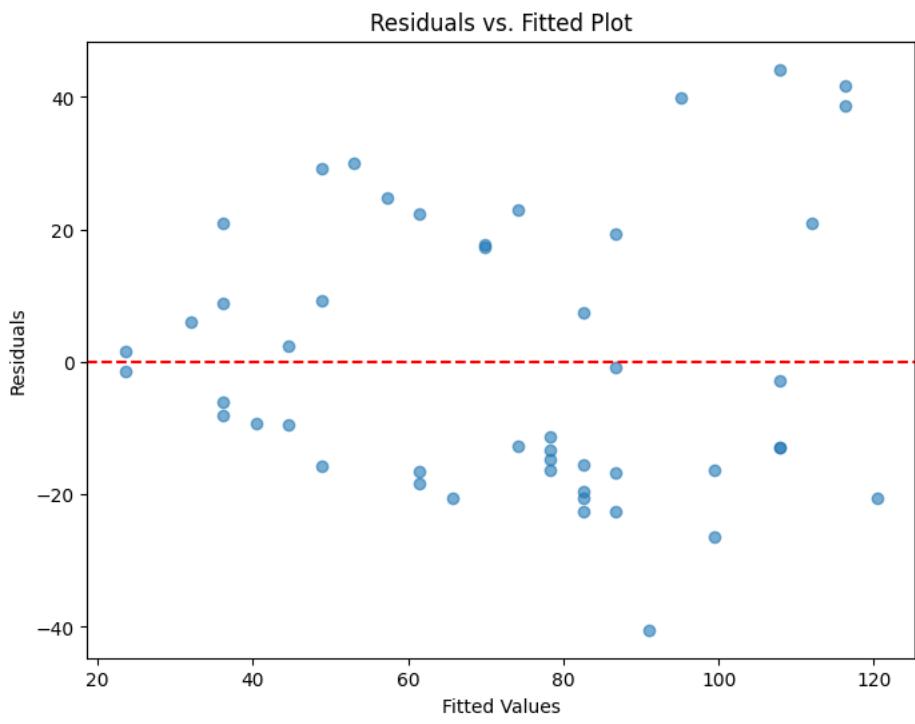
Table (10). Histogram of Residuals
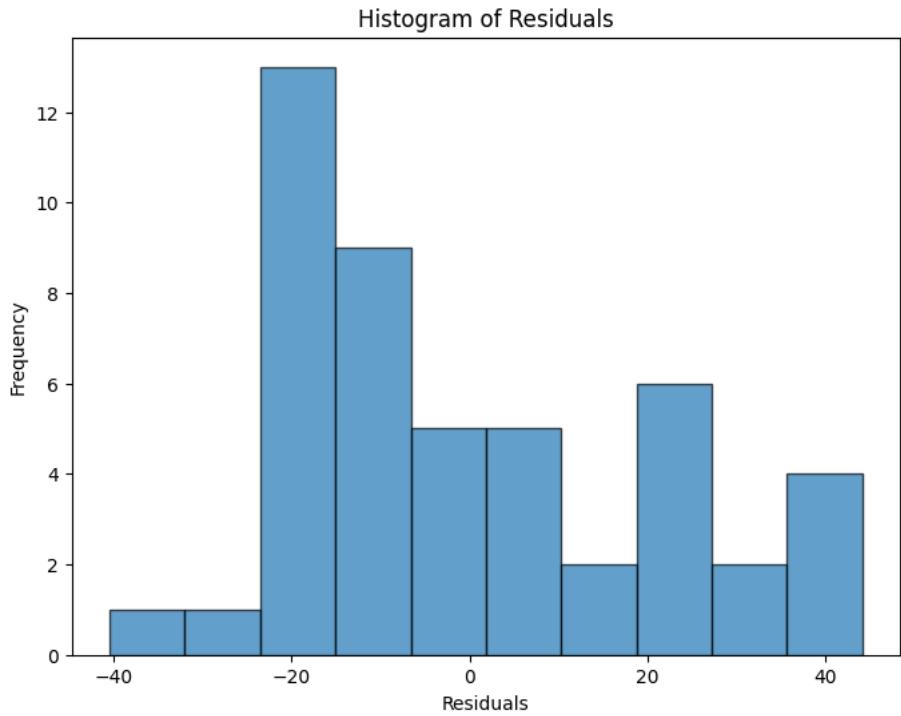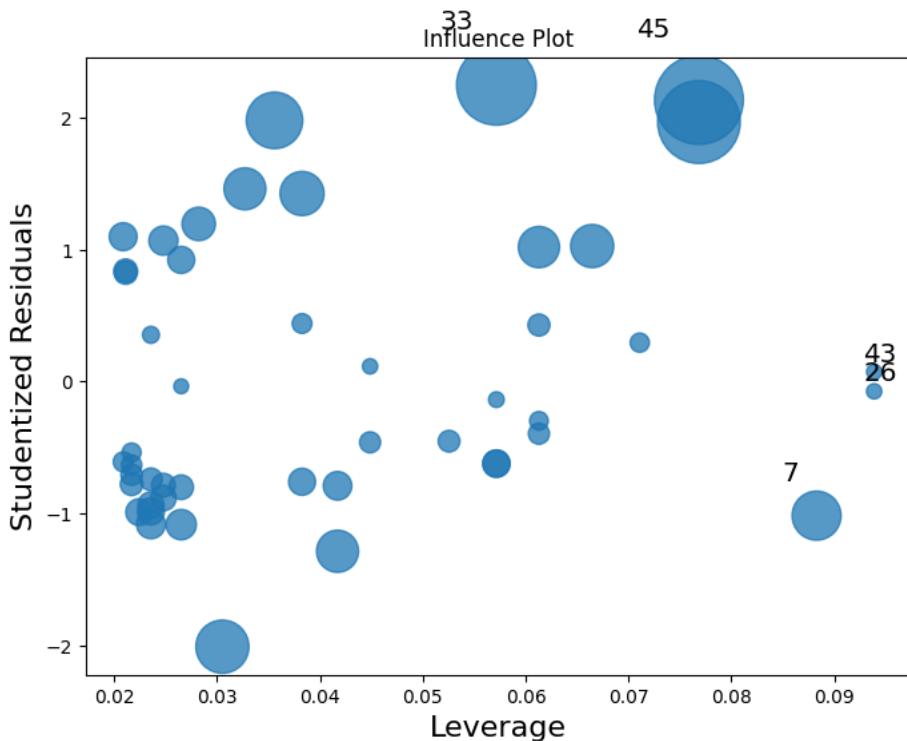
Table (11). Influence Plot

Influence Plot

2- Multiple Regression: Price = f (Age+DWT+BDI)

A multiple regression model was then used to capture the relationship between the Price and all other variables. The purpose of this model is to capture how the Price is driven by other variables which improves price prediction. During building the model, it is noticeable that the price is driven by both vessel fundamentals and market cycles strength, so adding the DWT and the BDI typically improves the explanatory power, price predication, and reduce the bias that may occur if the price was predicted using Age alone. This combines approach used the fitted multiple regression model and created a prediction for the "Bet Performer" based on the below ship characteristics:

- Age: 11 years (it was built in 1997; the use limitation is 25)
- DWT: 172 (i.e. 172,000 in Thousands)
- BDI: chosen market scenario (i.e. 11.6 or 12.5)

To decide whether the results derived through the regression is reliable, we checked the statistically significance of the results to validate the significant level of the p-value and R-square. The higher R-square, the better fit the model is for the regression analysis. Based on the conducted calculations of the projected sale Price, it can be determined whether the ship is worth the money, in terms of value, that the buyer is willing to pay.

Table (14). Multiple Regression Results Summary

```
                        OLS Regression Results
==============================================================================
Dep. Variable:                  Price   R-squared:                       0.920
Model:                            OLS   Adj. R-squared:                  0.915
Method:                 Least Squares   F-statistic:                     169.7
Date:                Tue, 23 Dec 2025   Prob (F-statistic):           3.39e-24
Time:                        22:47:46   Log-Likelihood:                -175.97
No. Observations:                  48   AIC:                             359.9
Df Residuals:                      44   BIC:                             367.4
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          44.2255     16.383      2.699      0.010      11.207      77.244
Age            -4.5438      0.261    -17.378      0.000      -5.071      -4.017
DWT             0.2422      0.092      2.643      0.011       0.058       0.427
BDI             7.2069      0.598     12.051      0.000       6.002       8.412
==============================================================================
Omnibus:                       13.373   Durbin-Watson:                   1.749
Prob(Omnibus):                  0.001   Jarque-Bera (JB):               19.393
Skew:                          -0.851   Prob(JB):                     6.15e-05
Kurtosis:                       5.607   Cond. No.                     1.85e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.85e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
        mean    mean_se   mean_ci_lower  mean_ci_upper  obs_ci_lower  \
0  125.829503  3.443368       118.88985     132.769155    104.740145

    obs_ci_upper
0      146.91886
```
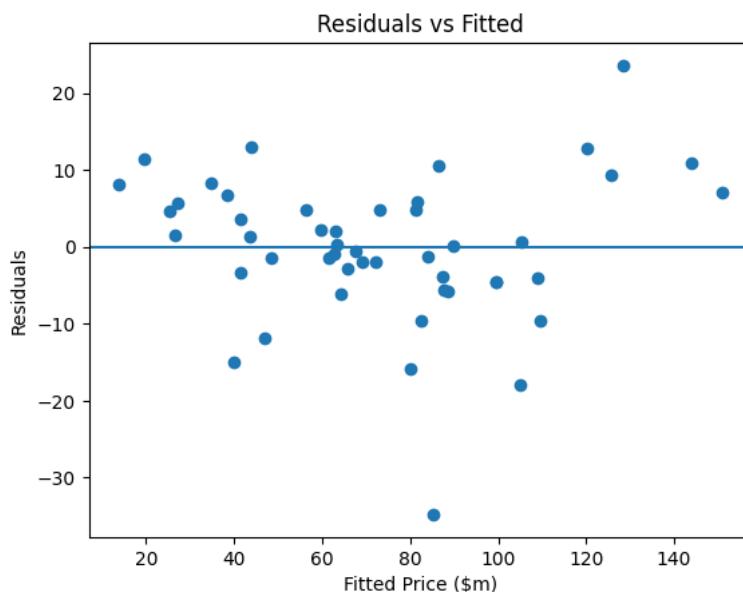
MR Model Diagnostics

A model diagnostic was used to support whether the regression model is a defensible or not based on the residuals' behavior and leverage points. Both diagnostics supported the multiple regression model is reliable, as shown in the below table:

Table (15). Residuals Vs. Fitted Price

Price Prediction

To predict the price of the Bet Performer, couple approaches were conducted. These are the Simple Linear Regression and the Multiple Linear Regression.

1. Multiple Linear Regression

Simple Linear Regression is a very straightforward approach predicting a quantitative response Y on the basis of a single predictor variable X. It assumes It assumes that there is approximately a linear relationship between X and Y. Mathematically, we can write this linear relationship as $Y = \beta_0 + \beta_1 X$. We computed a simple linear regression model on Age, DWT, and BDI (independent variables) individually to predict the price of the Bet Performer as shown in the table below:

- Predicted Price: Age $86.74 M, DWT $85.87 M, and BDI $96.06 M

By analyzing the price of the ships against independent variable (Age, DWT and BDI) we found out that:

- Age: Ships value depreciates at a rate of around $4.2 million with every year as it ages.
- DWT: For every one unit increase in DWT, the price of the ship increases by $0.98 million.
- BDI: For every one unit increase in BDI, the price of the ship increases by $0.004 million.

2. Multiple Linear Regression

Instead of fitting a separate simple linear regression model for each predictor, a better approach is to directly accommodate multiple predictors. We can do this by giving each predictor a separate slope coefficient in a single model. We can go for 4 different predictor combinations for our Multiple Regression model.

The RSE estimate gives a measure of the error of prediction. The lower the RSE, the more accurate the model (on the data in hand). Adjusted R^2 estimates the % of the variation in the measure of the sale price of the ships that can be predicted by independent variables.

Based on the Adjusted R Square and Residual Standard Error the combination of age at sale, DWT and Capesize combination emerge out as the best model for multiple linear regression to predict the price of Bet Performer. This resulted in a model where 91.5% of the variation in the sale price of ships could be explained by Age, DWT, and Capesize. We went ahead and used this model to predict the price of the Bet Performer. The predicted price as per multiple linear regression came out to be $125.83 M.

Confidence Interval $ 118.80M $132.85 M Prediction Interval $104.48 M $147.17 M

3. Different Market Scenarios

Predicted Price of Bet Performer under different scenarios 5 years younger $148.54 M 20K DWT lighter $120.98 M 30% lower charter rate $98.84 M

The above inference was considered to adjust the underpriced prediction. The final price predicted for Bet Performer as per our case study will be $ 132.19 M.

Conclusion

The client should consider he economic index (Capsize) for the bidding price of the Bet Performer. Hence, we recommend applying the market approach to evaluate the ship price. It is advised that the optimal price should range from the conservative price and the hot market price. We did a detailed study using the Market Approach as the base and came to a final price of $ 132.19M for the Bet Performer which we think will be the amount to bid and successfully buy the bulk carrier ship which the client had interest in.

Limitations and Recommendations

- The sample size is small, and other variables can be added to improve the model in relevance to the ship characteristics such as the fuel efficiency.
- The analysis requires additional data or information such as ship conditions (good or bad), ship maintenance cost, operation costs, risk or insurance costs, etc. to evaluate the ROI.
- The Capesize index continuously increased from Jan 2000 onwards, but the model and the observations did not consider the factors that are out of the dataset.
- The analysis was considering the sale prices of the previous ships; their prices were originally high. The average time charter was at its highest since 2000, and it has started to decline, and it shows a declining trend in ships available for sale since the first half of 2007. This is a concerning indicator, showing an alarming imbalance in the supply and demand scene in this market, that lead to the ship prices increase.
- Besides the unusual high increase in the prices of the ships, between 1980 and 1996, it is mainly due to many heavy-duty carriers was reported lost in the sea. And this number represents one-fifth of the total bulk trading fleet in the US at that time (the prices were high based on the market conditions at that time.