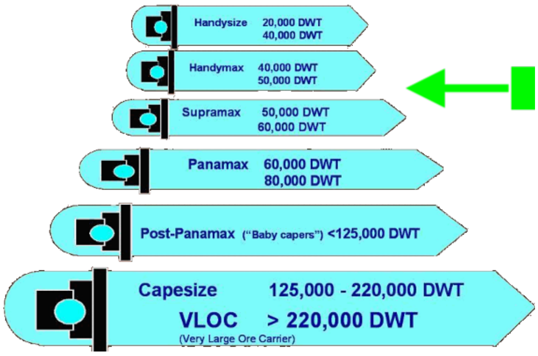
QMM: Bulk Carriers Data Case

F. Baeriswyl, J. Reboulleau

Objective: Build a model that allows to evaluate the price of a shipping vessel.

Type of ship: Bulk carriers of Handymax and Supramax sizes (for transport of cement, grain, coal, sugar), with a capacity between 40,000 and 65,000 tons: medium size, multipurpose, all kind of sea routes, not limited by accesses and/or infrastructures.


Handysize 20,000 DWT 40,000 DWT
Handymax 40,000 DWT 50,000 DWT
Supramax 50,000 DWT 60,000 DWT
Panamax 60,000 DWT 80,000 DWT
Post-Panamax ("Baby capers") <125,000 DWT
Capesize 125,000 - 220,000 DWT
VLOC > 220,000 DWT
(Very Large Ore Carrier)

Available data: Sales from January, 2004 to March, 2014, with information extracted from market reports available online:

Date: month and year of the sale;
Selling Price: selling price in million USD;
Vessel Age: age of the ship in years (possibly negative if under construction);
Dwt: deadweight: capacity of the ship in tons;
Freight (BSI): revenue per day in USD at the time of the sale (baseline revenue based on an index representing the classical revenue of a Supramax ship, the "Baltic Supramax Index" `https://www.balticexchange.com/`)

Steps:

1. Exploratory analysis
2. Simple regression: sale price as a function of age of the ship
3. Multiple regression and estimation error

Step 1: Exploratory analysis

a) Using the `ship1.csv` file, compute both the mean and the standard error of each respective variables (except for Date).

b) Plot the evolution of the revenue (Freight) over time. What can you see?

c) Study the evolution of the annual means of each variable (except for the Date).

d) Give the correlation matrix of the variables.

Step 2: Simple regression: selling price as a function of age

a) Plot the age of the ship against the selling price.

b) Fit a third-order linear model to explain the price with the age of the ship. Report the regression equation and the coefficient of determination R^2. Add the resulting regression line on the plot from 2a).

c) Compute the estimated price for this type of ship for vessels aged from 0 to 35 years old. Report the relative annual variation of the estimated price. What do you observe?

d) Fit an exponential model, that is a model of the form

$$y = \beta_0 e^{\beta_1 x}.$$

Note that taking the natural logarithm on both side leads to an equation of the form

$$\ln(y) = \ln(\beta_0) + \beta_1 x$$

by the properties of the logarithm. Denoting $y' := \ln(y)$ and $\beta_0' := \ln(\beta_0)$, one has

$$y' = \beta_0' + \beta_1 x,$$

i.e. the form of a simple linear regression model. Is this model better than the previous one?

e) Sales taking place in 2007 and 2008 should be considered as outliers, due to the perturbed economic cycle. Use the updated `ship2.csv` file to fit a second exponential model. Compute the estimated price under this model according to the age of the ship and report the relative annual variation of this estimated price.

f) Compute the log of the price and then the log of the estimations. Compute the errors and plot them. What can you say about them?

Step 3: Multiple regression and estimation error

a) Fit a multiple linear model of the form

$$sp_i = b_0 + b_1 va_i + b_2 dwt_i + b_3 freight_i + \epsilon_i$$

for $i \in \{1, ..., 881\}$, where sp_i is the selling price, va_i is the vessel age, dwt_i the deadweight and $freight_i$ the revenue of ship i, using the `ship1` data file. Report the regression equation, the coefficient of determination and the adjusted R^2 (what is the difference between the two?). What can you say about the regression coefficients?

b) Give the correlation matrix of the variables in 3a).

c) Test the overall significance of the model in 3a).

d) Plot the residuals against each of the dependent variables. What can you say about them?

e) For this question, download the code in moodle named Q3E.R. This is an R Script in which a new column is first created to give an age category to the vessel: category 1 for vessel aged less than 5; category 2 for vessels between 6 and 10 years old; category 3 for vessels between 11 and 15 years old; category 4 for vessels between 16 and 20 years old and finally category 5 for vessels aged more than 20 years old.

   Then, we fit moving linear regressions taking only 30 observations to train the linear model, moving over time. We end up with more than 850 models. To analyse them, we study the mean and standard error in the relative and absolute errors in the price according to the model. Finally, we distinguish for the 5 age categories to see where the model is performing better, respectively worse. Review the code and analyse the results.

f) How could you improve the model in 3a) ?