Generalized Additive Models on Estimation of Newbuilding Prices and Lead Times for Bulk Carriers

Author:	Chao Yan
Student Number:	4645294
E-mail:	yanchao941102@outlook.com
Faculty:	3ME, Marine Technology
Department:	Ship Production
Report Number:	SDPO.19.026.m
Graduation Committee:	Dr. ir. J. F. J. Pruyn
	Dr. ir. RG Hekkenberg
	Dr. W.W.A. Beelaerts van Blokland
	Dr. R. Delfos

Contents

List of Tables

List of Figures

1 Introduction

1.1 Background

The shipping industry is one of the oldest industries and is still one of the key factors of worldwide economy. As the world's population continues to grow, especially in developing countries, together with the tide of the economic globalization increasing, the reliable and efficient maritime transport has a significant role to play in growth and sustainable development. Without shipping, the international trade, the import/export of products and the transport of raw materials would be on the verge of collapse. I Therefore, it is no exaggeration to say that the shipping industry is the backbone of the global trade and the global economy [1].

However, despite the importance of the shipping market and its massive demand all over the world, making a profit or even surviving in this market is not that easy. Although the combination of planned industrial shipping and intensively competitive markets makes shipping recognizable to the exponents of the perfect competition market [2], the shipping market is characterized by a high degree of uncertainty due to its cyclicality. Furthermore, the high dependence of the shipping market upon the industrial production and the international trade further increase its uncertainty. The uncertainty makes the market almost unpredictable, and in this case, to achieve success in the market, shipowners need to take an overall consideration of four markets of the shipping industry, namely newbuilding market, freight market, sale & purchase market and demolition market [2]. Among the four markets, the newbuilding market deserves special attention because the vessel, as the only carrier of maritime transportation, can influence the whole shipping industry significantly, and the ordering of new vessels will probably bring shipowners from hell to heaven or vice versa.

When placing an order for a vessel, shipowners are usually most concerned about two things: how much they need to pay and how long they need to wait for their vessels. The vessel newbuilding price is of great importance because it means whether shipowners can afford a new vessel or they can directly purchase a second-hand one

instead. With regard to the latter concern, if shipowners can get reliable information on when to get the new vessel, or in other words, the lead time, they can adjust their operational strategies correspondingly, react flexibly to the fast-changing shipping market and have much more chance of making a profit.

1.2 Research Purpose

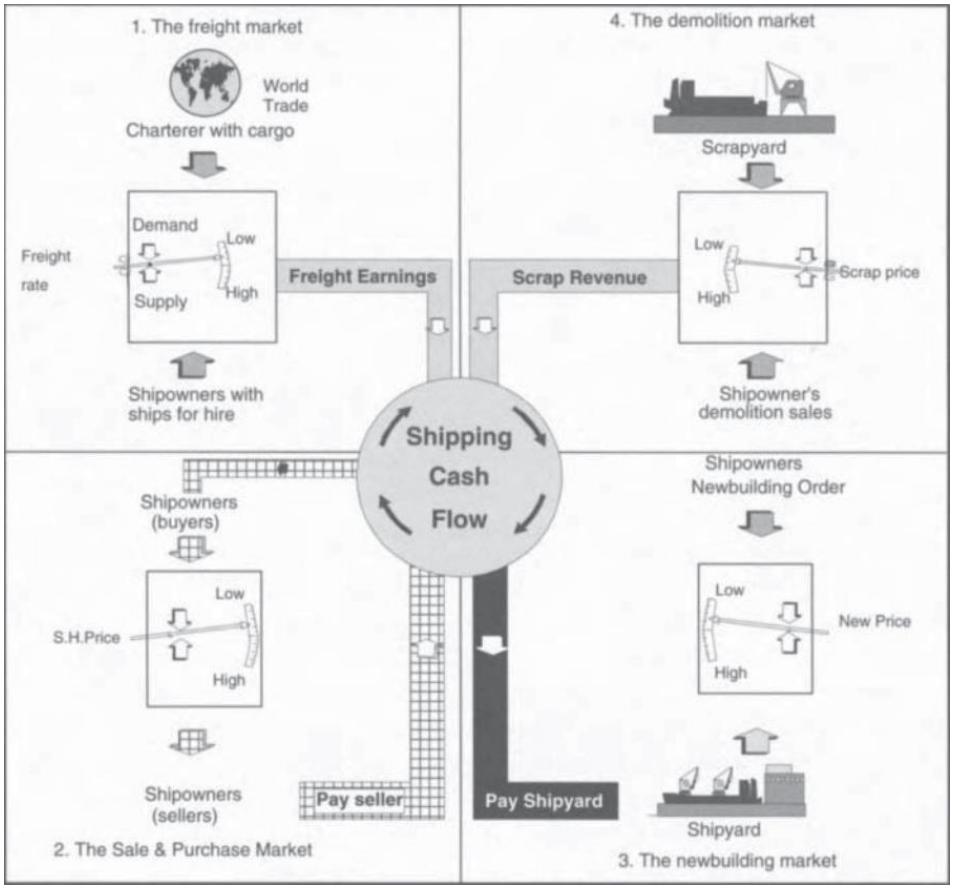

Figure 1.1: The Four Markets Which Controls Shipping

The plot [2] above presents the four submarkets of shipping and the cash flows between them. The freight market is where the shipowners hire vessels to charters with cargo and get freight earnings. The sale & purchase market is where second-hand vessels are traded. The newbuilding market is for shipowners to order new vessels from shipyards and in the demolition market too old vessels are scraped.

The Maritime Business Game (MBG) within the course Maritime Finance Business &

Law models the whole shipping market, where the game participants serve as shipowners to personally experience the correlations between different markets. The MBG is currently solely focused on second-hand vessel trading and it is planned to extend this game with playable shipyards. To allow both the shipowners and shipyards to be played independently and together, it is important to organize a newbuilding market between the two sides. In other words, when the shipowners request an offer with a certain number of input variables, the newbuilding market should be able to generate a number of proposals for this tender. As talked above, the newbuilding prices and lead times are two major concerns between shipowners and shipyards, and, in order to establish a good newbuilding market, it is necessary to accurately estimate the two items with given input variables. Thus, the purpose of this thesis is to find out what variables are influential to a vessel's newbuilding price and lead time, and how to establish the estimation models.

In addition, given that dry bulk commodities account for nearly half of the global maritime trade [3] and most vessels traded in the MBG are bulk carriers, only bulk carriers will be researched in this thesis.

1.3 Research Questions

The discussion in the last section clearly points out that the research purpose of the thesis is to accurately estimate a bulk carrier's newbuilding price and its lead time. However, shipbuilding is a heavy engineering industry which is capital-intensive and labor-intensive. Besides, vessel is also a kind of complex and sophisticated product whose building process is normally very time-consuming. The complexity of the shipbuilding industry and its products makes it very hard to create a definite and precise estimation method of newbuilding prices and lead times. Fortunately, as the rapid development of data collection and analysis in the shipping market, the detailed information of a built vessel as well as lots of data related to the four shipping markets can be found in maritime databases such as Clarksons. With the quick and easy access

to enormously useful data, the estimation of a vessel's new building price and lead time is possible. And this leads to the research question of this thesis:

"Is it possible to estimate a bulk carrier's newbuilding price and its lead time, using open access data and empirical information?"

Clearly, to get the answer to this research question, we need to first figure out the determinants of bulk carrier's newbuilding prices and lead times, and find a way to establish a reasonable estimation model. Therefore, the research question can be further divided into three sub questions:

1. *What is the appropriate method to estimate bulk carrier's newbuilding prices and lead times? And why?*
2. *Where can we obtain the necessary data for research and how should we process the data?*
3. *What variables are influential to bulk carrier's newbuilding prices and how to establish a reasonable estimation model for newbuilding prices?*
4. *What variables are influential to bulk carrier's lead times and how to establish a reasonable estimation model for lead times?*

1.4 Research Structure

Chapter 2 introduces the general requirements for the modelling technique and the regression analysis as one statistic method to estimate relationships among variables, with several types of regression illustrated. Furthermore, the generalized additive model (GAM), as the selected modelling method, is primarily introduced, including the theoretical basis, evaluation methods, some important points and the reasons to choose this. This chapter exactly provides the answer to the sub question 1.

Chapter 3 firstly introduces a series of previous researches on newbuilding prices, based

on which the variables identification is done. Next, the collection and analysis of the data for this thesis are illustrated, followed by the specific processes of using GAM to estimate the newbuliding prices of bulk carriers and the analysis of the fitting results. The sub questions 2 and 3 could be answered by this chapter.

Chapter 4 has a similar structure to that of Chapter 3, except that the research object is lead times of bulk carriers and several tests are made to explain the causes of bad performance of the estimation model. This chapter answers the sub question 4 and part of the sub question 2.

Finally, Chapter 5 summarizes the answers to the research questions, some general remarks and several recommendations about making improvements.

2 Generalized Additive Models

2.1 Introduction

This chapter first illustrates some general requirements for the modelling technique in this thesis. Then, regression analysis is introduced with several common types of this, and the generalized additive model (GAM) is selected as the specific regression method in this thesis for its great flexibility and ability to estimate complicated relationships among different types of variables. Furthermore, the theoretical basis and attention points of GAM are introduced. Finally, the whole chapter is summarized generally.

2.2 Requirements for Modelling Technique

When estimating the newbuilding prices and lead times of bulk carriers, we need to deal with quite complicated relationships among various types of variables within the shipping market. For instance, there are numerical variables versus categorical variables, like vessel speed versus Ice Class, and continuous variables versus discrete variables, like vessel size (Dwt) versus the number of hatches on board. Furthermore, these variables almost have totally different units and magnitudes. What is more, the relationships among those variables could be very complex that cannot be simply defined. For example, high freights rates indicating a prosperous shipping market could possibly increase the bulk carrier fleet size since more shipping capacity is required. However, considering the time to build vessels, this effect is not instantaneous but time lagged. This process might also be influenced by a series of external variables such as the expectations for future market and the turbulence of other industries, and the cyclicality of shipping market makes it more complicated.

To sum up, what we generally need is a modelling technique that could estimate the relationship between newbuilding prices or lead times of bulk carriers and a series of maritime variables. Specifically, this technique should be able to deal with different types of variables and analyze rather complicated relationships.

2.3 Regression Analysis

In statistical modelling, regression analysis is a set of statistical processes to describe and estimate the relationship between a given variable (usually known as the dependent or response variable) and one or more variables (usually known as the independent or explanatory variables). With respect to the structure and the assumptions, there are various types of regression methods.

2.3.1 Linear Regression

Among all the regression methods, linear regression is the most fundamental and common one for finding linear relationship between the dependent variable and independent variables. Generally, given *n* observations, the linear regression describing the relationship between the dependent variable and *k* independent variables has the following form:

$$y_i = \beta_0 + \beta_1 \cdot x_{i1} + \beta_2 \cdot x_{i2} + \cdots + \beta_k \cdot x_{ik} + \varepsilon_i, \qquad i = 1, \ldots, n, \tag{2.1}$$

where y_i is the dependent variable, x_{i1}, x_{i2},…,x_{ik} are the independent variables, β_0, β_1,…,β_k are unknown parameters and ε_i is the independent random variable with zero mean and the same variance.

2.3.2 Additive Model

Additive model (AM) is a non-parametric regression method suggested by Breiman and Friedman [4], representing a generalization of multiple regression, which analyzes the relationship between several independent variables and a dependent variable. Compared to linear regression, AM would be to maintain the additive nature of the model, but to replace the simple terms $\beta_j x_j$ with $f_j(x_j)$, where $f_j()$ is a non-parametric function of the independent variable x_j. Therefore, AM takes the form as follows:

$$E\left[y_i \middle| x_{i1}, \ldots, x_{ik} \right] = \beta_0 + \sum_{j=1}^{k} f_i\left(x_{ij} \right) \tag{2.2}$$

where E[y_i] is the expectation of y_i, β_0 is the unknown parameter, and $f_j(x_{ij})$ are unknown

smooth functions fit from the data. In other words, instead of a single parameter for each independent variable, in the additive models a non-parametric function is estimated for each independent variable, to achieve the best prediction of the dependent variable values.

2.3.3 Generalized Linear Model

The generalized linear model (GLM), as an extension of linear regression, is a framework for modelling the dependent variables that are bounded or discrete, formulated by Nelder and Wedderburn [5]. The differences between GLM and the general linear regression are in two major aspects: first, unlike in linear regression, the distribution of the dependent variable does not have to be continuous, and can be non-normal as long as it follows the exponential family of distributions such as the Poisson, binominal, gamma and normal distributions; second, the dependent variable values are predicted by a linear combination of independent variables, which are connected by to the dependent variable via a link function. A GLM has the basic structure as follows:

$$g(\mu_i) = \boldsymbol{X}_i\boldsymbol{\beta} \tag{2.3}$$

where $\mu_i \equiv \mathrm{E}(Y_i)$ is the expectation of random dependent variable Y_i, g is a smooth monotonic "link function", $\boldsymbol{X}_i$ is the i_{th} row of a model matrix, $\boldsymbol{X}$, and $\boldsymbol{\beta}$ is a vector of unknown parameters.

2.3.4 Generalized Additive Model

The generalized additive model (GAM) was originally developed by Hastie and Tibshirani [6], derived by the combination of AM and GLM. In general, GAM has a structure as follows:

$$g(\mu_i) = \beta_0 + f_1(x_1) + f_2(x_2) + \cdots + f_k(x_k) \tag{2.4}$$

In other words, the purpose of GAM is to maximize the quality of prediction of a

dependent variable from various exponential distributions, by estimating non-parametric functions of the independent variables which are connected to the dependent variable via a link function.

2.3.5 Reasons for GAM

Considering the explanations above, GAM is superior to other regression methods for these reasons: first, unlike linear regression, the dependent variable of GAM does not have to be normally distributed but can follow any distributions of exponential family; second, there are no limitations about variable types that any variables including numerical or categorical variables and continuous or discrete variables could be imported into the estimation models; third, with the application of both link functions and smooth functions, GAM can deal with complicated nonlinear relationships among various variables. In contrast, linear regression can only assume linear relationships between the dependent variable and independent variables and GLM needs to assume linear relationships between the transformed dependent variable (by link function) and independent variables.

Jones [7] and Chambers [8] argued that GAM could relax the universal statistical assumption of linearity, and thereby potentially allowing the discovery of something important missed by traditional parametric regression, especially when the number of potential explanatory variables is large.

The superiority of GAM are combined with some downsides. With the application of both link function and smooth function, the results of GAM are generally in mapping, but not in a distinct continuous function description, which might be hard to interpret. However, this disadvantage is not that significant, considering the main purpose of the accurate estimation. In addition, like other nonparametric regression methods, GAM has a high propensity for over-fitting [8]. This downside could be somehow avoided by observing the plot of each smooth term and adjusting the smoothness.

Therefore, GAM is selected as the specific modelling method in this thesis.

2.4 Generalized Additive Model

GAM is introduced elaborately by Wood [9], with a general structure as follows:

$$g(\mu_i) = \boldsymbol{A}_i\boldsymbol{\theta} + f_1(x_{1i}) + f_2(x_{2i}) + f_3(x_{3i}, x_{4i}) + \cdots \tag{2.5}$$

where $\mu_i \equiv \mathrm{E}(Y_i)$ and $Y_i \sim \mathrm{EF}(\mu_i, \varphi)$. Y_i is a response (dependent) variable and $\mathrm{EF}(\mu_i, \varphi)$ denotes an exponential family of distribution with mean μ_i and scale parameter, φ, $\boldsymbol{A}_i$ is a row of the model matrix for any strictly parametric model components, $\boldsymbol{\theta}$ is the corresponding parameter vector, and the f_j are smooth (non-parametric) functions of the covariates, x_k.

The subsequent subsections introduce the major points requiring special attention when using GAM.

2.4.1 Assumptions

- The data of the dependent variable are independently distributed, i.e., cases are independent. The dependent variable does not need to be normally distributed, but it typically assumes a distribution from an exponential family (e.g. Poisson, binominal, gamma, normal,…).
- Not like linear regression nor GLM, GAM does not assume a linear relationship between the dependent variable and the independent variables, nor a linear relationship between the transformed responses in terms of the link function and the explanatory variables, but it assumes a linear relationship between the independent variable transformed by the link function and the dependent variables transformed by the smooth functions.
- The residuals need to be independent but do not have to be normally distributed, instead they could also follow a distribution of exponential family.
- The residuals should have a mean of zero and their variance should keep constant. In other words, the residuals plots should have the same values for all values of the linear predictors (fitted values).

2.4.2 Presenting Functions with Basis Expansions

The presentation and estimation of component functions of a GAM model is best introduced by considering a model containing one function and one covariate, just as follows:

$$y_i = f(x_i) + \varepsilon_i \tag{2.6}$$

where y_i is a response variable, x_i is a covariate, f is a smooth function and the ε_i are independent $N(0, \sigma^2)$ random variables.

To estimate f requires that f be represented in such a way that (2.6) becomes a linear model. This can be done by choosing some basis functions, defining the space of functions of which f is an element. If $b_j(x)$ is the j^{th} such basis function, for some values of the unknown parameters, f is assumed to have a representation as follows:

$$f(x_i) = \sum_{j=1}^{k} b_j(x)\beta_j \tag{2.7}$$

where k is the basis dimension, which controls the degree of model smoothness.

2.4.3 Controlling Smoothness by Penalizing Wiggliness

One possibility for choosing the degree of model smoothness is to use backwards selection to select k. However, such an approach is problematic. A model based on k-1 evenly spaced knots will not generally be nested within a model based on k evenly spaced knots. It is possible to start with a fine grid of k knots and simply drop knots sequentially, as part of backwards selection, but the resulting uneven knots spacing can itself lead to poor model performance. Furthermore, the fit of such regression models tends to depend quite strongly on the locations chosen for knots.

An alternative is to keep the basis dimension fixed at a size a little larger than it is believed could reasonably be necessary, but to control the model's smoothness by adding a "wiggliness" penalty to the least squares fitting objective. Therefore, rather than fitting the model by minimizing

$$\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|^2 \tag{2.8}$$

it could be fitted by minimizing

$$\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|^2+\lambda\sum_{j=2}^{k-1}\left\{f\left(x_{j-1}^*\right)-2f\left(x_j^*\right)+f\left(x_{j+1}^*\right)\right\}^2 \tag{2.9}$$

where the summation term measures wiggliness as a sum of squared second differences of the function at the knots (where * notes that even knot spacing has been assumed). The smoothing parameter, λ, controls the trade-off between smoothness of the estimated *f* and fidelity to the data. $\lambda \to \infty$ leads to a straight line estimation for *f*, while $\lambda = 0$ results in an un-penalized piecewise linear regression estimate.

2.4.4 Splines

According to [9], representing the smooth model terms using spline basis is likely to obtain substantially reduced function approximation errors for a given dimension of smoothing basis. There are various types of splines in use, and the most common ones are introduced by the following paragraphs.

- **Cubic Regression Splines**

Cubic spline is a piecewise polynomial with a set of extra constraints (continuity, continuity of the first derivative, and continuity of the second derivative). There are many equivalent bases that can be used to represent cubic splines. Cubic regression splines are the approach to parameterize the spline in terms of its values at the knots. In this case, apart from the continuity constraints, the spline should have zero second derivative at the start and end knot.

Unlike other traditional methods such as polynomial regression or broken stick regression, cubic regression spline takes both smoothness and local influence into consideration [10]. In addition to having directly interpretable parameters, this basis does not require any re-scaling of the independent variables before it can be used to construct a GAM, although we do have to choose the locations of the knots.

- **P-splines**

Yet another way to represent cubic splines (and splines of higher or lower order) is by use of the B-spline basis, which is appealing because the basis functions are strictly local: each basis function is only non-zero over the intervals between m+3 adjacent knots, where m+1 is the order of the basis. P-splines are low rank smoothers using a B-spline basis, usually defined on evenly spaced knots, with a difference penalty applied directly to the parameters, β_i, to control function wiggliness.

P-splines are extremely easy to set up and use and allow a good deal of flexibility, in that any order of penalty can be combined with any order of B-spline basis. Their disadvantage is that the simplicity is somewhat diminished if uneven knot spacing is required, and that, relative to the more usual spline penalties, the discrete penalties are less easy to interpret in terms of the properties of the fitted smooth.

- **Thin Plate Regression Splines**

Thin plate splines [11] are an elegant and general solution to the problem of estimating a smooth function of multiple explanatory variables, from noisy observations of the function at particular values of those predictors. Thin plate regression splines (TPRS) are constructed by starting with the basis and penalty for a full thin plate spline and then truncating this basis in an optimal manner, to obtain a low rank smoother.

One key advantage of the TPRS is that it avoids the knot placement problems of conventional regression spline modelling, but it also has the advantage that smooths of lower rank are nested within smooths of higher rank, so that it is legitimate to use conventional hypothesis testing methods to compare models based on pure regression splines. Compared to this plate splines, TPRS are reasonably computationally efficient and it also retain the rotational invariance (isotropy) of full thin plate spline.

Summing up, for a given basis dimension, cubic regression splines typically perform a little less well than TPRS, but a little better than P-splines. Thus, TPRS is selected as the splines for constructing GAMs in this thesis, though it is slower to set up than others.

2.4.5 Exponential Family of Distributions

As mentioned before, we need to determine the distribution of the dependent variable values before constructing a GAM. Table 2.1 presents the most common members of exponential family.

Distrubution	**Range**	**Probability Function**	**Link Function**
Normal	real: $(-\infty, +\infty)$	$f(y) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{\frac{-(y-\mu)^2}{2\sigma^2}\right\}$	μ
Poisson	inegrer: 0, 1, 2, …	$f(y) = \frac{\mu^y \exp(-\mu)}{y!}$	$\log(\mu)$
Binomial	integer: 0, 1, 2, … , n	$f(y) = \binom{n}{y}\left(\frac{\mu}{n}\right)^y \left(1-\frac{\mu}{n}\right)^{n-y}$	$\log\left(\frac{\mu}{n-\mu}\right)$
Gamma	real: $(0, +\infty)$	$f(y) = \frac{1}{\Gamma(\upsilon)}\left(\frac{\upsilon}{n}\right)^y y^{\upsilon-1} \exp\left(-\frac{\upsilon y}{\mu}\right)$	$\frac{1}{\mu}$
Inverse Gaussian	real: $(0, +\infty)$	$f(y) = \sqrt{\frac{\gamma}{2\pi y^3}} \exp\left\{\frac{-\gamma(y-\mu)^2}{2\mu^2 y}\right\}$	$\frac{1}{\mu^2}$

Table 2.1: Common Distributions of Exponential Family

Considering that the values of the datasets in this thesis are almost continuous and positive, Gamma distribution seems a good choice. Meanwhile, during the later research, it is found that the Gamma distribution combined with the link function of *log*() indicates a best performance for constructing GAMs. Therefore, all the regressions in this thesis are carried out using *log*() as the link function and assumes that the dependent variables follow the Gamma distribution.

2.4.6 Software

We can use many kinds of software to construct GAMs, such as S language, R language, SAS and Python, among which the package “mgcv” of the R, developed by Wood [9], is the most common and mature one. In this thesis, all the GAMs are created using “mgcv” in the R.

2.4.7 Results Evaluation

There are various methods and tests for evaluating the regression results, and generally applying only one of them is not enough to judge the model performance.

- **Adjusted R-square**

In statistics, the adjusted R-square, also known as the coefficient of determination, is the proportion of the variance in the dependent variable that is predicted from the independent variable(s). R^2 normally ranges from 0 to 1, and the bigger, the better. The most general definition of R^2 is as follows:

$$SS_{tot} = \sum_i \left(y_i - \bar{y}\right)^2,\ SS_{reg} = \sum_i \left(f_i - \bar{y}\right)^2,\ SS_{res} = \sum_i \left(y_i - f_i\right)^2,\ R^2 \equiv 1 - \frac{SS_{res}}{SS_{tot}} \qquad (2.10)$$

where y_i are the observed values of the dependent variable, f_i are the predicted values fitted by regression, $\bar{y}$ is the mean of the observed data, SS_{tot} is the total sum of squares (proportional to the variance of the data), SS_{reg} is the regression sum of squares and SS_{res} is the residual sum of squares.

- **Generalized Cross Validation**

As mentioned in Subsection 2.3.3, the smoothness of models are usually controlled by the smoothing parameter, λ, and if λ is too high then the data will be over-smoothed while if it is too low then the data will be under-smoothed: in both cases this will mean that the estimated smooth function will not be close to the true one. Therefore, when evaluating a GAM, the choice of the smoothing parameter requires attention, which could be measured by the generalized cross validation (GCV) score. The computation of GCV score could be carried out by "mgcv" and the lowest GCV score indicates the optimal smoothing parameter.

- **Akaike Information Criterion**

The Akaike information criterion (AIC) is an estimator of the relative quality of statistical models for a given dataset. Given a collection of models for the same data, AIC estimates the quality of each model, relative to each of the other models. Thus, AIC provides a direct means for model selection. The criterion is as follows:

$$\text{AIC} = -2l + 2p \qquad (2.11)$$

where l is the maximized log likelihood for the model and p the number of model

parameters that have to be estimated. By using the R language, it is easy to calculate AIC and the model with the lowest AIC is selected.

- **Hypothesis Tests for Significance**

It is necessary to examine whether there is enough evidence to prove that the dependent variable is related to each independent variable. Put another way, we need to test whether each independent variable is significant to the dependent variable and if it is reasonable to include it in the final model. This can be done by applying the hypothesis test as follows:

$$H_0 : f\left(x_i\right) = 0;\ H_1 : f\left(x_i\right) \neq 0;\ \text{for } i = 1, 2, 3, \ldots n \tag{2.12}$$

The significance level of the hypothesis test usually equals to 5%. Therefore, for the smooth terms whose p-value is larger than 5%, the null hypothesis is satisfied, indicating that this term is not significant to the dependent variable and should be discarded. In contrast, for the smooth terms whose p-value is smaller than 5%, the alternative hypothesis is satisfied, meaning that this term is significant to the dependent variable and should be included in the model. The similar hypothesis test can be done with respect to the parametric terms.

- **Plots of Smooth Terms**

Apart from the statistic evaluation methods, observing the plots of smooth terms are also of great necessity. By doing so, we could judge whether the smooth functions are over-fitted or visually weird. In addition, it is also helpful for dealing with the multicollinearity issues, which will be discussed subsequently.

2.4.8 Model Selection Approach

In statistics, stepwise regression is a method of fitting regression models in which the choice of independent variables is carried out based on some pre-specified criterion, such as adjusted R^2, AIC, GCV score and etc. In each step, a variable is considered for addition to or subtraction from the set of independent variables. The main approaches

are as follows [12]:

- Forward selection (FS), which involves starting with no independent variables in the model, testing the addition of each variable using a chosen model fit criterion, adding the variable (if any) whose inclusion gives the most statistically significant improvement of the fit, and repeating this process until none improves the model to a statistically significant extent.
- Backward elimination (BE), which involves starting with all candidate variables, testing the deletion of each variable using a chosen model fit criterion, deleting the variable (if any) whose loss gives the most statistically insignificant deterioration of the model fit, and repeating this process until no further variables can be deleted without a statistically significant loss of fit.
- Bidirectional elimination, a combination of the above, testing at each step for variables to be included or excluded.

As mentioned before, the evaluation of model should be done comprehensively, using multiple evaluation methods together. Therefore, the direct use of the above selection approaches is not advisable and we should apply the model selection approaches in accordance with the actual circumstances.

2.4.9 Multicollinearity

Multicollinearity means that there are strong correlations among independent variables, indicated as higher correlation coefficients close to 1. It increases dimensionality of models while adding no new information, increases noise and leads to bias predictions. If multicollinearity is ignored, R^2 will be high but the individual variable will have high standard errors, and the regression becomes very sensitive to small changes in the specification. There are many ways to detect multicollinearity including correlation matrix, tolerance measures, variance inflation factor and condition index. In this thesis, we would pick out the strongly correlated independent variables by the correlation matrix and then construct GAMs with these variables in pairs to judge whether there are multicollinearity issues.

In addition, according to [13], if the model is rather statistically adequate in terms of each independent variable being of plausible magnitude and having an appropriate sign, multicollinearity could be ignored.

2.5 Summary

Regression analysis is a common statistical method to estimate relationships among variables and in this thesis, to estimate bulk carrier's newbuilding prices and lead times, GAM appears to be the optimal specific regression method for its extreme flexibility and ability to estimate complex nonlinear relationships between the dependent variable and independent variables, which are of various types with different units and magnitudes.

All the GAMs in thesis are based on thin plate regression splines, with the dependent variable following the Gamma distribution and *log*() as the link function. The construction of GAMs will be carried out by the package "mgcv" in the R language. The fitting results are evaluated by a series methods, including R^2, GCV score, AIC, hypothesis test for significance and observing plots of smooth terms. Furthermore, as the case may be, we might use FS or BE to select the optimal model. Besides, the multicollinearity issues are also dealt with through a reasonable approach.

3 Newbuilding Prices of Bulk Carriers

3.1 Introduction

The newbuilding market brings new ships into the shipping industry and sends cash out of the market as materials, labor and profits, and the variables within it are subject to distinct cyclical fluctuations [2]. Considering that the shipbuilding requires substantial capital investment and a high level of technological expertise, the high volatility of the newbuilding prices can sometimes result in severe financial and economic problems for both shipowners and shipbuilders.

Just like the commodities in a general market, new vessels are also highly supply-demand related, so figuring out the supply and demand of new ships, as well as, the determinants of newbuilding prices is of great importance to all the stakeholders, including shipowners, shipbuilders, banks, local governments and etc.

This chapter is organized as follows. Firstly, some previous researches related to newbuilding prices are introduced. Secondly, the factors that are influential to newbuilding prices are identified. Next, the collection, pre-processing and analysis of data used for this research are explained, followed by the method of how to establish a model for newbuilding prices and its results. At last, a summary and some concluding remarks are made.

3.2 Literature Review

The intensive competition of the shipping market sets the scene for a harsh commercial climate, which is intensified by the cyclical nature of the shipbuilding demand [2]. And the cyclicality of the shipbuilding market was firstly analyzed by Tinbergen [14], who argued that the shipbuilding market is of great dependence on the freight rate and the freight rate, in turn, also depends on the present shipping capacity. The correlation results in the shipbuilding cycle because of the time lag between the supply and the demand of the shipping capacity. The phenomenon clearly conforms to the cobweb

theorem, where low total shipping capacity leads to the increase of freight rates and the ships ordered during the prosperous market are delivered several years later, thus arising the shipping capacity.

Hawdon built a Tank Ship Building Model, where a tanker's newbuilding price is related to both cost (steel price) and asset pricing (freight rates) [15]. He also found that it is the current levels of freight rates that significantly influence newbuilding prices, not the lagged freight rates. The negative impact of shipping overcapacity on newbuilding prices is shown as well.

The correlation between newbuilding prices and second-hand prices were researched in [16], where it's observed that second-hand prices are volatile while newbuiliding prices are relatively stable. Beenstock [17]built an asset pricing model of newbuilding prices and assumed that newbuilding vessels and second-hand vessels are exactly perfect substitutes, though later Beenstock observed that this condition is unlikely to hold because newbuilding prices are too sticky compared with second-hand prices. So the more reasonable statement should be that newbuilding prices and second-hand prices are good close substitutes [18]. In 1986, Strandenes [19]defined the long-run earnings of a vessel based on its newbuilding price and assumed that the second-hand value of a vessel is a weighted average of short- and long-run earnings.

Jin [20]took a supply and demand method to analyze the newbuilding market, where various elements are incorporated, such as shipyard labor costs, orderbooks, and other exogenous factors including the improvement of shipbuilding technology. She also pointed out that the overall shipbuilding capacity is of the utmost importance of the supply in the newbuilding market and the orderbook is a significant demand variable.

Volk [21]combined the asset pricing model and the supply-demand method to explain the shipping and newbuilding cycles through the influences of freight rates, shipbuilding innovation, psychological and speculative behaviors of shipowners.

Tsolakis [14]used an Error Erection Model to investigate second-hand prices and found that second-hand prices are mainly driven by newbuilding prices and time charter rates.

Adland [22]tested the instantaneous equilibrium relationship between second-hand prices, newbuilding prices and freight rates in a Vector Error Erection Model and accounted for the time-varying delivery lag.

Based on the supply-demand analysis of the shipbuilding market, Stopford categorized the variables of newbuilding prices into two groups: the demand influences including shipping freight rates, market expectations and credit availability versus shipbuilding capacity, shipyard unit costs and production subsidies. Besides, he also mentioned that exchange rates and inflation may affect the prices in the long term [2].

According to the considerations above, the major variables of newbuilding prices could be preliminary summarized as follows: [18]

- Shipbuilding costs;
- Shipyard capacity;
- Vessel Orderbook;
- Freight rates;
- Second-hand prices;
- Exchange rates;
- Inflation.

3.3 Variables Identification

In Section 3.2, the major influential variables of newbuilding prices are listed, and this section will further analyze the effects of those variables and how they are identified. To not miss anything important, all the potential influential variables based on references and reasonable assumptions will be discussed. In general, the variables can be categorized into three groups: cost related variables, asset pricing related variables, and supply-demand related variables.

3.3.1 Cost Related Variables

- **Shipbuilding Costs**

With large and sophisticated products, shipbuilding industry is a heavy engineering business requiring substantial capital investment and much labor. Hence, the shipbuilding cost is commonly a considerable number and, due to the fierce shipyard competition, is the most influential factor determining newbuilding prices. However, the shipbuilding costs are also determined by many aspects and deserve the detailed analysis.

Above all, the building cost of a vessel is greatly dependent on its specifications and Shetelig defined four subsystems of a vessel to estimate the shipbuilding costs as follows [23].

1) Hull. The size of a vessel is important and this could be expressed in DWT, GT, L*B*D or a number of other measures. As the main interest of shipowners is the cargo carrying capacity, DWT is selected as the sole variable characterizing a vessel's size. Besides, the hull type, including D/Bottom, D/Hull, D/Sides and S/Skin, and coating, including epoxy, zinc and ploy, should be considered as well. In addition, whether the vessel has Ice Class and is strengthened for heavy cargo can also affect the shipbuilding costs.

2) Machinery and Propulsion. As the significant attributes of a vessel, the horsepower and speed deserve much attention, since the higher-speed vessels may have more slender hull or larger engine installation. Main engines, as the most critical parts of vessels, are very expensive and can account for a great portion of the shipbuilding costs, so the types of engines should be taken into account.

3) Cargo Containment and Handling Equipment. Grain capacity, measuring the actual volumes available for cargo, is chosen as the explanatory variable to present a vessel's cargo carrying ability. Besides, whether the ship has built-in cranes is influential to the shipbuilding costs, as well as the number of holds and hatches.

4) Common Systems and Others. Nowadays everyone is concerned about the

environment, so the emission standards of vessels are noteworthy, and it will take shipbuilders more money to build a vessel with a better emission standard. Therefore, the IMO Tiers of vessels are included.

We have to acknowledge that these measures are imperfect, but they are indeed commonly available and comparable by shipowners across contracts for different vessel types.

However, even given detailed specifications of vessels, the shipbuilding costs are still difficult to measure. Hence, considering that steel plates account for 30% of newbuilding prices [13], the Japan steel plate commodity price is taken as a reliable newbuilding cost indicator in this thesis.

Apart from vessel specifications and steel prices, most production for shipbuilding is labor-intensive, and nearly half of the costs incurred during construction of a standardized bulk carrier are related to overheads and labor [2]. Therefore, labor costs can be a good proxy of shipbuilding costs.

- **Government Subsidies**

Shipbuilding is a labor-intensive and export-oriented industry with spill-over effects on the domestic economy, making it the repeated beneficiary of various forms of government aids [24]. To support the shipbuilding industry, some local governments will provide subsidies for their qualified shipyards, and it determines whether the shipyard is able to sell at prices below an acceptable return on capital, thereby affecting the newbuilding prices. Therefore, shipyards receiving more construction subsidies are more competitive in price setting compared to others.

- **Exchange Rates**

As the shipbuilding industry is a worldwide business, most contracts are made between different countries, so exchange rates will play a role in the determination of newbuilding prices [2, 25]. Since most shipbuilding costs are paid by local currencies while, finally, vessels are quoted in US dollars, a higher exchange rate would mean less USD for the same local currencies, indicating a lowering of the price and more

temptation for shipowners to order.

- **Inflation**

In economics, inflation is a sustained increase in the general price level of goods and services in an economy over a period of time. When the general price level rises, each unit of currency buys fewer goods and services; consequently, inflation reflects a reduction in the purchasing power per unit of money. The newbuilding market is no exception for this phenomenon, where a higher inflation indicates a higher nominal vessel price. This argument is also supported by Stopford [2].

- **LIBOR**

As the building costs of vessels are normally extremely high, at least millions of dollars, shipbuilders might not have enough circulating funds to pay for raw materials and labor costs, so they need to apply for loans from banks, which involves a cost of financing [2, 25]. For this cost, LIBOR (London Inter-Bank Offered Rate), the interest rate at which banks lend money to each other, is considered as an indicator.

- **Shipyard Size**

Porter suggests that larger producers are able to influence the prices, particularly if producers are more concentrated than buyers, which is the case of shipbuilding [26]. Thanks to the flexibility and redundancy coming with size, larger shipyards can potentially get higher prices than smaller shipyards. In general, larger shipyards are presumably able to build the largest and most sophisticated vessels, while it may be impossible for those smaller ones. Therefore, larger shipbuilders do not encounter much competition in certain segments, as well as having a greater variety of orders to bid on. In addition, larger shipbuilders could take better advantage of the economy of scale, thereby reducing costs.

3.3.2 Asset Pricing Related Variables

- **Freight Rates**

The positive influence of freight rates on newbuilding prices is well understood in that

high freight rates usually represent a prosperous shipping market, where the demand for shipping capacity is substantial and profitability is great, hereby attracting more market entrants and increasing the demand for new vessels, as well as the level of newbuilding prices. There are many types of freight rates, such as trip charter rates and spot freight rates, among which time charter rates are selected. The reason for this is that time charter rates denote the shipowners' and the charterers' expectations for the shipping market [13]. Thus, it can be assumed that the higher the time charter rate, the higher the vessel's profitability, and consequently, the more willing the shipowners to order new vessels are, indicating a higher newbuilding price. Besides, unlike trip charter rates, time charter rates are not route-specific, which means that they are not strictly restricted and can reflect the overall shipping market better. In addition, spot freight rates are the prices of freight service only at a point in time, which are too volatile and cannot reflect the expectations of the future market. For all these reasons, time charter rates, as a market indicator, are much better than others and are chosen in this thesis.

- **Baltic Exchange Dry Index**

The Baltic Dry Index (BDI) is a composite of the Capesize, Panamax and Supramax time charter averages, used as a proxy for dry bulking shipping stocks as well as a general shipping market bellwether. The main feature of BDI is that this index directly measures the demand for shipping capacity versus the supply of dry bulk carriers, thus it is totally devoid of speculative contents. Therefore, BDI is also selected as an indicator to reflect the prosperity of the shipping market.

- **Second-hand Prices**

As discussed before, new vessels and second-hand vessels are good close substitutes that an increase of second-hand prices will result in an increase demand for new vessels and the convenience for shipowners to switch between the two sides makes this substitution more elastic. Therefore, shipowners need to continuously compare second-hand prices with newbuilding prices to decide whether to purchase a second-hand vessel or just to make an order of a newbuilding vessel.

- **Lead Time**

For shipowners, a vessel with shorter lead time will have a higher value in present value terms than an equal vessel with longer lead time [27]. This statement can be partially supported by the finding that shipyards with superior lead times are more attractive to shipowners to make an order [28]. Thus, shipowenrs may also be willing to pay more to make theirs vessels delivered earlier. Actually, this variable is greatly influenced by the negotiations between shipowners and shipbuilders.

3.3.3 Supply-demand Related Variables

- **Shipyard Capacity**

According to Stopford [2], the newbuilding prices are greatly influenced by the number of slots available at shipyards during a given period. This number exactly represents the utilization of shipyard capacity. If a shipyard does not receive many orders and its orderbook is not full enough compared to its building capacity, it is likely to offer relatively low prices to potential buyers. On the contrary, when a shipyard is quite busy with a lot orders, it can afford to miss some orders and will quote higher prices for new vessels. Therefore, the utilization of shipyard capacity is taken as an indicator to represent the supply and demand condition.

- **Orderbook**

From the perspective of the shipbuilding market, the orderbook of vessels can be regarded as a reliable proxy to reflect the supply-demand situation. According to [13] and [25], with an increasing fleet an increase of the amount of replacements and orderbook is required. So the size of orderbook should be taken relatively, as a percentage of the fleet size. However, at this stage, we cannot be sure whether the ordebook is positively or negatively related to newbuilding prices in that there are two opposite statements: on the one hand, the larger the size of orderbook, the more prosperous the shipbuilding and shipping market, and as a result, the higher newbuilding prices are; on the other hand, because of the time lag between the orders and the deliveries of vessels, a large orderbook may also indicate the potential

oversupply of vessels, and in this case shipowners and banks are unwilling to invest or lend money, thereby leading to a slide in newbuilding prices.

3.3.4 Summary of Variables

The discussed variables are summarized in Table 3.1 below.

Variable	Predicted Effect	Supporting Reference
Cost Related		
Shipbuilding Costs	Positive	[2], [13], [15], [20]
Government Subsidies	Negative	[2], [24]
Exchange Rates	Negative	[2], [25]
Inflation	Positive	[2]
LIBOR	Positive	[2], [25]
Shipyard Size	Positive	[26]
Asset Pricing Related		
Freight Rates	Positive	[2], [14], [15], [21]
Baltic Dry Index	Positive	
Second-hand Prices	Positive	[16], [17], [18], [19], [22]
Lead Time	Negative	[27], [28]
Supply-demand Related		
Shipyard Capacity	Positive	[2], [20]
Orderbook	Positive or Negative	[13], [20]

Table 3.1: Summary of Variables

3.4 Data Collection and Analysis

3.4.1 Data Source

The data gathered for this thesis mainly comes from Shipping Intelligence Network and World Fleet Register of Clarksons Sources, the world's leading shipping services provider. Shipping Intelligence Network provides online access to over 100000 pages of data, including the latest information on the shipping markets at glance, easily downloadable versions of its wide range of market reports, extensive fleet and orderbook listings, and thousands of time-series and graphs of key commercial information. The World Fleet Register is also an online data service, comprising details of all merchant vessels over 100 GT and non-merchant vessels. It covers more than

150,000 vessels in service, on order, under conversion or recently removed from the fleet, any of which may be selected by a wide range of criteria. Besides, for some economic data, we also need to turn to other sources such as World Bank Open Data.

3.4.2 Data Scope

Newbuilding prices play a significant role in the competition among shipyards, thus for some shipyards, the data of newbuilding prices is private and not open to public. After searching the World Fleet Register, it is found that most bulk carriers with a newbuilding price are ordered after 2000. Besides, for some variables like BDI, the time series data is not complete before 2000. In addition, in the database the latest contract date of bulk carriers with a newbuilding price is 20/12/2017. For these reasons, in this chapter we will research all the bulk carriers registered in Clarksons with a newbuilding price ordered between 01/01/2000 and 20/12/2017, and there are 1780 vessels in total. Considering that whether the bulk carrier has a newbuilding price seems random, this division of bulk carriers could somehow represent all the bulk carriers during that period.

3.4.3 Problems and Options Associated with Data

During the process of data collection, a series of problems occurred and we need to make choice between several options to deal with data. These problems and options will be explained in the subsequent paragraphs.

- **Size Group Classification**

Generally, according to the size, bulk carriers are categorized into several groups including Handysize, Handymax, Supramax, Panamax and Capesize. However, the classification method varies between different data sources, and even in the same source the specific classification is confusing. World Fleet Register is taken to exemplify this problem, where Panamax can refer to bulk carriers with 65,000-79,999 dwt or 65,000-99,999 dwt and the range of Supramax is overlapping with that of Handysize and Handymax. To make the whole research consistent, based on World Fleet Register, the

size groups of bulk carriers in this thesis are categorized as follows: Handysize (10,000-39,999 dwt), Handymax (40,000-64,999 dwt), Panamax (65,000-99,999 dwt) and Capesize (100,000+ dwt).

- **Data Frequency**

When colleting time series data, it is found that we need to determine the data frequency, annual, quarterly, monthly or daily. Considering that the negotiation processes between shipowners and shipbuilders always last for months, taking the data monthly seems the best choice. This data frequency is applied to all the collected time series data if possible, including but not limited to time charter rates, orderbook size, second-hand prices, LIBOR, and etc.

- **Shipbuilding Costs**

In section 3.3.1, a series of variables are mentioned to measure the shipbuilding costs. However, after searching all the data sources, it is found that the data of some variables is incomplete or that there is no access to the data. For instance, of all the 1780 bulk carriers there are only 48 vessels having a record of Hull Type, and obviously this variable has to be excluded from the research. The same problem occurs when it comes to coating, engine type and IMO tiers. Nevertheless, as the vessels with Ice Class are rare and important for the trading areas, it would generally be noted. We could safely assume that the bulk carriers without the records of Ice Class do not have Ice Class instead removing them from the dataset.

As for the labor costs, obtaining reliable time series of manufacturing wages is difficult, especially for China which is a dominant player in shipbuilding industry. Hence, the gross domestic product (GDP) per capita is introduced as a proxy for wages, as recommended by International Labor Organization [29]. Besides, this variable can also to some extent capture other cost elements specific to the country of manufacture that are hard to include by other means. In addition, it should be noted that the data of this variable can only be collected annually.

- **Government Subsidies**

Given the available data sources, this variable is hard to quantify, let alone some government aids are not given in the form of direct money support. Therefore, this variable can only be excluded.

- **Exchange Rates, Inflation and LIBOR**

The values of exchange rates vary much between different countries, and to make the rates comparable, they are all indexed at 2000 to equal 1 [25]. When colleting the data from Clarksons, it is found that the monthly exchange rates for China before November 2005 are missing, so we use the quarterly values instead for the time period before that.

As for inflation, there several options including the inflation of the country, the inflation of OECD, and the inflation of USA, among which we take the last one since most traded vessels are finally quoted in US dollars.

For interest rates, we use the 3-Month LIBOR, based on USD, which can be obtained in Shipping Intelligence Network.

- **Time Charter Rates and Second-hand Prices**

In Shipping Intelligence Network, the time charter rates are recorded in terms of "6 Month", "1 Year", "3 Year" and "5 Year" for different vessel sizes. To make the thesis in consistence, we select "1 Year Time Charter Rate in Long Run Historical Series" for Handysize, Handymax, Panamax and Capesize.

A similar problem occurs for second-hand prices that there are lots of time series including Bulk Carrier Secondhand Price Index and second-hand prices for vessels of different ages and sizes. Because at this stage we cannot determine which one is better, we take Bulk Carrier Secondhand Price Index and second-hand prices for 10-year-old bulk carriers of Handysize 32K, Handymax 56K, Panamax 76K and Capesize 180K.

- **Shipyard Capacity and Orderbook**

Given the available data sources and information, it is hard to obtain the data that can quantify the utilization of shipyard capacity. Fortunately, orderbook can take the place of that. We have already determined that the size of orderbook should be taken relatively,

as the percentage of fleet size. However, there still remains a choice to be made between a representation of the different size groups or one value over all the bulk carriers, and we will check both of them in the later research. Moreover, if taking account of the time lag between the demand for vessels and the market response, we should also consider how the orderbook size is changing during the given period. Therefore, we might also take the difference between the demand of current month and the demand several months ago as a proxy, where the time steps are 1 month, 3 months and 6 months [25].

- **Lead Time and Shipyard Size**

In this thesis, lead times are set to the duration between the contract date and the built date of a vessel, with a unit of week. It is easy to get the data for this from World Fleet Register.

Clarksons has a method to classify shipyard size based on the size of orderbook in millions of CGT: very small (<0.049), small (0.049~0.1), medium (0.1~0.49), large (0.49~1) and mega (>1). In this thesis, we use these groups to make up shipyard size dummy variables.

3.4.4 Data Pre-processing

Before the further data analysis, the dataset requires necessary pretreatment in case there are too many outliers, which can have very bad influences on the model results. Table 3.2 and 3.3 shows the summary statistics and the summary of categorical variables of the initial dataset respectively. Figure 3.1 presents the overview of some numerical variables, where it is obviously noticed that there are some outliers for Price and Price/Dwt, which should be dropped.

Variable	Unit	N	Mean	Std.Dev.	Min	Max	Expected Effect
Price	mln USD	1780	44.73	28.53	2.10	240.00	-
Price/Dwt	USD/ton	1780	585.70	394.49	30.00	4,698.88	-
Dwt	ton	1780	95,522	78,530	10,202	403,919	Positive
Grain Capacity	cu m	1735	102,983	58,537	9,346	389,700	Positive
Horsepower	-	1780	15,856	7,128	3,600	84,967	Positive
Speed	knot	1780	14.24	0.67	10.50	17.20	Positive
No.Holds	-	1775	6.32	1.57	3.00	11.00	Positive
No.Hatches	-	1773	6.32	1.57	3.00	11.00	Positive
Lead Time	week	1780	147.23	61.67	18.00	628.00	Negative
TC_Rate	USD/day	1780	36,914	35,671	6,525	161,600	Positive
BDI	-	1780	4,117	2,868	383	10,844	Positive
SH Price	mln USD	1780	37.71	25.40	8.00	116.00	Positive
SH Index	-	1780	252.18	121.13	71.15	499.66	Positive
Exchange Rate	-	1680	0.88	0.10	0.73	1.27	Negative
Inflation	%	1780	2.40	1.41	-2.10	5.60	Positive
LIBOR	%	1780	2.68	2.17	0.32	6.92	Positive
GDP Per Capita	USD	1680	8,751	9,869	959	48,603	Positive
GDP Per Capita Annual Growth	%	1680	8.79	3.82	-1.14	13.64	Positive
Steel Price	USD/ton	1780	675.32	185.81	300.00	1,350.00	Positive
Orderbook%Total	%	1780	43.03	20.58	7.23	79.99	P/N
Orderbook%Group	%	1780	43.08	24.71	3.08	117.76	P/N
Orderbook%Total_1	%	1780	0.57	1.56	-2.77	4.57	P/N
Orderbook%Group_1	%	1780	0.65	0.65	-5.32	7.66	P/N
Orderbook%Total_3	%	1780	1.78	4.42	-6.22	9.44	P/N
Orderbook%Group_3	%	1780	2.00	4.98	-9.05	15.13	P/N
Orderbook%Total_6	%	1780	3.35	8.59	-11.54	18.36	P/N
Orderbook%Group_6	%	1780	3.80	9.32	-16.76	29.49	P/N

Table 3.2: Summary Statistics before Pre-processing

Builder Country	%	**Gear**	%	**Ice Class**	%	**Strengthened**	%	**Shipyard Size**	%
China P.R.	74.04	Yes	47.36	Yes	10.06	Yes	49.44	Very small	2.42
India	0.79	No	52.64	No	89.94	No	50.56	Small	0.62
Indonesia	0.06							Medium	22.53
Japan	6.24							Large	16.18
Philippines	1.35							Mega	16.85
Poland	0.11							N/A	41.40
Romania	0.39								
Singapore	0.11								
South Korea	14.10								
Spain	0.06								
Taiwan	0.84								
Turkey	0.06								
Vietnam	1.85								

Table 3.3: Summary of Categorical Variables before Pre-processing

According to these, the dataset can be processed by two methods:

1）Low data-density in the boundaries of the sample may lead to broad confidence intervals and inconclusive results in regression [30], so to avoid this, 5% of the observations with extreme values for each variable should be dropped.

2）The outliers can be distinguished and deleted visually by observing Figure 3.1 and considering actual situations. The integrity of the dataset can be maintained well by this method.

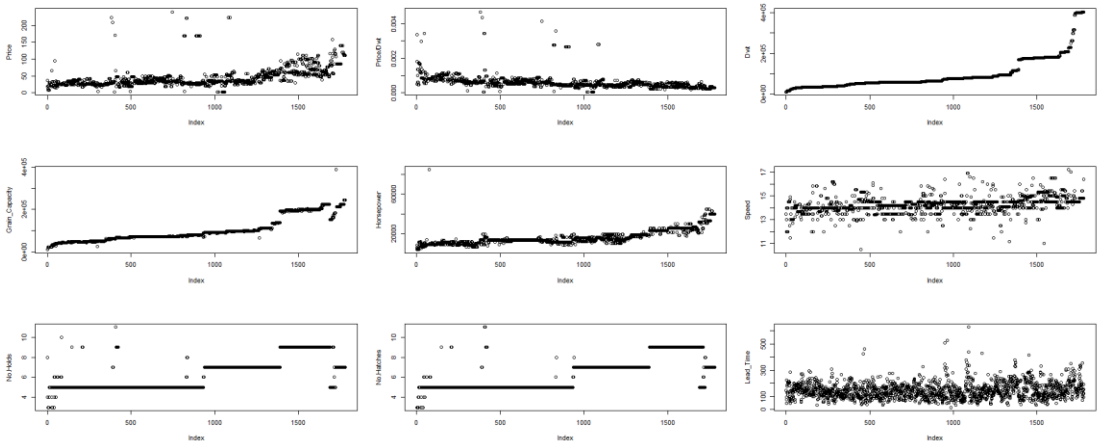

Figure 3.1: Overview of Some Numerical Variables

The two methods both have advantages and disadvantages: the former one can eliminate the influences of outliers to the greatest extent but might exclude too many observations; the latter one keeps the data integrity well but may overlook some outliers. Considering the size of the initial dataset, losing too many observations will have a bad influence on the model results. Therefore, in this research, the method 2 is applied. The summary statistics and the summary of categorical variables of the dataset after pre-processing are shown in Table 3.4 and Table 3.5 respectively.

Variable	Unit	N	Mean	Std.Dev.	Min	Max	Expected Effect
Price	mln USD	1625	43.46	23.26	6.60	157.50	-
Price/Dwt	USD/ton	1625	544.60	205.19	166.36	1,353.07	-
Dwt	ton	1625	97,162	80,179	13,533	403,919	Positive
Grain Capacity	cu m	1594	103,563	58,637	15,718	389,700	Positive
Horsepower	-	1625	15,977	7,057	3,600	44,826	Positive
Speed	knot	1625	14.24	0.67	10.50	17.20	Positive
No.Holds	-	1622	6.33	1.57	3.00	11.00	Positive
No.Hatches	-	1622	6.33	1.57	3.00	11.00	Positive
Lead Time	week	1625	146.81	57.78	18.00	441.00	Negative
TC_Rate	USD/day	1625	37,731	36,299	6,525	161,600	Positive
BDI	-	1625	4,136	2,855	383	10,844	Positive
SH Price	mln USD	1625	38.36	25.64	8.00	116.00	Positive
SH Index	-	1625	254.41	119.63	71.15	499.66	Positive
Exchange Rate	-	1625	0.88	0.10	0.73	1.24	Negative
Inflation	%	1625	2.39	1.41	-2.10	5.60	Positive
LIBOR	%	1625	2.71	2.18	0.32	6.92	Positive
GDP Per Capita	USD	1625	8,464	9,534	959	48,603	Positive
GDP Per Capita Annual Growth	%	1625	8.87	3.78	-1.14	13.64	Positive
Steel Price	USD/ton	1625	679.36	183.35	300.00	1,350.00	Positive
Orderbook%Total	%	1625	43.55	20.20	7.23	79.99	P/N
Orderbook%Group	%	1625	43.81	24.53	3.08	117.76	P/N
Orderbook%Total_1	%	1625	0.59	1.57	-2.77	4.57	P/N
Orderbook%Group_1	%	1625	0.69	0.69	-5.32	7.66	P/N
Orderbook%Total_3	%	1625	1.80	4.47	-6.22	9.44	P/N
Orderbook%Group_3	%	1625	2.10	5.04	-9.05	15.13	P/N
Orderbook%Total_6	%	1625	3.38	8.66	-11.54	18.36	P/N
Orderbook%Group_6	%	1625	3.98	9.43	-16.76	29.49	P/N

Table 3.4: Summary Statistics after Pre-processing

Builder Country	%	Gear	%	Ice Class	%	Strengthened	%	Shipyard Size	%
China P.R.	79.02	Yes	47.02	Yes	9.60	Yes	48.86	Very small	2.28
Japan	5.54	No	52.98	No	90.40	No	51.14	Small	0.55
South Korea	15.45							Medium	21.35
								Large	15.26
								Mega	17.85
								N/A	42.71

Table 3.5: Summary of Categorical Variables after Pre-processing

3.4.5 Data Analysis

After the pretreatment, there are totally 1625 bulk carriers in the dataset, including 304 Handysize, 537 Handymax, 375 Panamax and 409 Capesize. It should be noticed that Table 3.3 shows that all the countries except China, Japan and South Korea only account for 5.62% of the bulk carriers and this is why only these three countries have remained.

From Table 3.4, we can see that, even after the pretreatment, for different bulk carriers the price gap is still very huge; the cheapest vessel only costs 6,600,000 dollars whereas it takes almost 24 times of that price to build the most expensive one. Similarly, Dwt ranges from 13,533 tons to 403,919 tons, Grain Capacity from 15,178 cu m to 389,700 cu m and Horsepower from 3,600 to 44,826. Besides, the numbers of holds and hatches both range from 3 to 11. Due to the boom and the bust of the world economy during the period 2006-2009, the proxies, like time charter rates and BDI, reflecting the prosperity of the shipping and shipbuilding market also have a large range of values. In addition, the data of the orderbook size suggests that the demand for bulk carriers might fluctuate much within a relatively short time period.

Table 3.5 indicates that China, as the dominant player in the dry bulk shipbuilding market, produces most of the bulk carriers, nearly 80%. Besides, about half of the bulk carriers are equipped with gears and half of the vessels are strengthened for heavy cargo as well. In addition, the number of bulk carriers without Ice Class is about nine times of that of bulk carriers with Ice Class, which is logical since most bulk carriers do not have to voyage in ice zones. As for shipyard size, except the shipyards of unknown size, most shipyards are classified as medium, large and mega, suggesting the dominance of

large shipyards in the newbuilding market.

Table 3.6 and Table 3.7 below respectively present the correlations of the categorical and numerical variables in the dataset, where the correlation coefficients larger than 0.6 are shown in bold numbers. The correlation measures the degree of any linear relation between two variables. There are two purposes of this correlation analysis in this chapter: firstly, although the correlations between the response variable (Price or Price/Dwt) and the explanatory variables are essentially nonlinear, the linear correlation analysis could also suggest how the explanatory variables affect newbuilding prices to some degree, which could provide some constructive suggestions for the following models; secondly, the correlations between explanatory variables is of great help to dealing with the multicollinearity issues, which will be explained further in the following section.

	Price	**Price/Dwt**	**Gear_Yes**	**Ice Class_Yes**	**Strengthened _Yes**	**Yard Size_ Mega**	**Yard Size_ Large**	**Yard Size_ Medium**	**Yard Size _Small**	**Yard Size _Very Small**
Price	1									
Price/Dwt	-0.234	1								
Gear_Yes	-0.520	0.512	1							
Ice Class_Yes	-0.207	0.178	0.053	1						
Strengthened_Yes	-0.013	0.006	0.046	-0.093	1					
Yard Size_Mega	0.217	-	-	-	-	1				
Yard Size_Large	0.116	-	-	-	-	-	1			
Yard Size_Medium	-0.251	-	-	-	-	-	-	1		
Yard Size_Small	-0.099	-	-	-	-	-	-	-	1	
Yard Size_Very Small	-0.106	-	-	-	-	-	-	-	-	1

Table 3.6: Correlations between Categorical Variables

Table 3.6 suggests that the introduction of Gear or Ice Class appears to have a negative effect on newbuilding prices. This here can be explained as follows. Gear is a feature more available in smaller bulk carriers than larger ones in that smaller vessels are usually used to deliver goods to more secluded places without proper unloading equipment available on the ports. Thus, the bulk carriers with Gear usually have smaller sizes, and of course with lower prices. Similarly, most Ice Class vessels are Handysize and Handymax. The result of Yard Size seems logical that large and mega shipyards can obtain higher newbuilding prices while smaller yards accept lower quotes. However, this result is not very reliable since 42% of the observations are not recorded with Yard Size and smaller shipyards only account for a very low proportion of all the fixtures.

	Price	Price /Dwt	Dwt	Grain Capacity	Horsepower	Speed	No.Holds	No. Hatches	Lead Time	TC Rate	BDI	SH Price	SH Index	Exchange Rate	Inflation	LIBOR	GDP Per Capita	GDP Per Capita Growth	Steel Price	Orderbook %Total	Orderbook %Group
Price	1																				
Price/Dwt	-0.234	1																			
Dwt	**0.845**	-0.575	1																		
Grain Capacity	**0.802**	**-0.627**	**0.897**	1																	
Horsepower	**0.872**	-0.528	**0.938**	**0.901**	1																
Speed	0.357	-0.272	0.376	0.421	0.414	1															
No.Holds	0.577	-0.474	0.581	**0.794**	**0.644**	0.378	1														
No.Hatches	0.592	-0.488	**0.601**	**0.814**	**0.662**	0.386	**0.984**	1													
Lead Time	0.293	0.071	0.191	0.181	0.205	0.071	0.116	0.120	1												
TC Rate	**0.761**	-0.016	-	-	-	-	-	-	0.335	1											
BDI	0.439	0.410	-	-	-	-	-	-	0.334	**0.766**	1										
SH Price	**0.767**	0.016	-	-	-	-	-	-	0.357	**0.972**	**0.795**	1									
SH Index	0.462	0.442	-	-	-	-	-	-	0.336	**0.773**	**0.951**	-	1								
Exchange Rate	0.026	0.025	-	-	-	-	-	-	0.051	0.084	0.173	0.067	0.064	1							
Inflation	0.276	0.257	-	-	-	-	-	-	0.184	0.522	0.588	0.543	**0.628**	0.090	1						
LIBOR	0.226	0.308	-	-	-	-	-	-	0.399	0.415	**0.605**	0.448	0.555	0.426	0.462	1					
GDP Per Capita	0.048	0.019	-	-	-	-	-	-	-0.131	-0.069	-0.122	-0.088	-0.116	0.290	-0.052	-0.179	1				
GDP Per Capita Growth	-0.026	0.106	-	-	-	-	-	-	0.280	0.169	0.312	0.202	0.293	-0.039	0.135	0.497	-	1			
Steel Price	0.304	0.289	-	-	-	-	-	-	0.070	0.435	0.437	0.517	**0.612**	-0.253	0.540	0.003	-0.057	-0.012	1		
Orderbook%Total	0.319	0.310	-	-	-	-	-	-	0.018	0.443	0.527	0.516	**0.624**	-0.112	0.172	-0.075	-0.069	0.055	**0.656**	1	
Orderbook%Group	**0.615**	-0.023	-	-	-	-	-	-	0.088	**0.695**	0.482	**0.723**	0.576	-0.076	0.215	-0.021	-0.038	0.014	0.595	-	1

Table 3.7: Correlations between Numerical Variables

Table 3.7 indicates that the newbuilding prices have relatively strong and positive linear correlations with Dwt, Grain Capacity, Horsepower, TC Rate, SH Price and Orderbook%Group, suggesting that these variables deserve special attention when establishing the models later. Besides, those variables with bold correlation coefficients, such as Dwt versus Horsepower, TC Rate versus BDI and SH Price versus Orderbook%Group, presumably incur the multicollinearity issues when included in the model together. Therefore, they need to be checked in advance, which will be explained further in the follow paragraphs. Moreover, it is noticed that the correlation between No. Holds and No.Hatches is close to 1, thus we will discard one of them, keeping No.Hatches.

3.5 Methods of Establishing Models

3.5.1 Single Tests of Numerical Variables

Before establishing the estimation model, single tests of all the numerical variables of nonlinear estimation should be made for the following two purposes: first, it can basically suggest the nonlinear relationship between the response variable and tested variables, helping to determine the selected variables in the model; second, the plots given by those tests help to judge whether the multicollinearity issues could be ignored. The statistic results of the single tests are presented in Table 3.8 and Figure 3.2 shows the plots of smooth terms for all the tested variables.

Combining Table 3.8 and Figure 3.3, we can preliminarily judge that there are several variables obviously having positive influences on newbuilding prices, including Dwt, Grain_Capacity, Horsepower, No.Hatches, TC Rate, SH Price, Orderbook%Total and Orderbook%Group. As for Lead Time, BDI, SH Index and Inflation, they also present positive effects except the beginning or end of the curves. The abnormality of Lead Time and Inflation could be explained by the low density of data. For BDI and SH Index, we assume it is because that when BDI or SH Index is at low levels, shipowners prefer to buy bulk carriers with large sizes, which could be verified by the plots below.

These two plots come from the models where Dwt is regarded as the dependent variable and BDI or SH Index is regarded as independent variable, showing the nonlinear relationship between them. It is clear that when BDI or SH Index is at low levels, Dwt is at high levels, verifying the assumption.

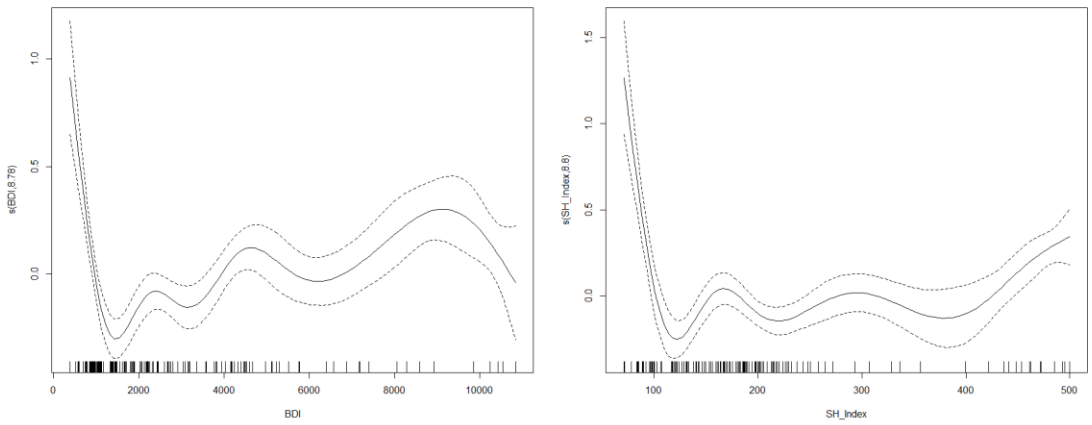

Figure 3.2: Smooth Terms of Models for Dwt versus BDI or SH Index

When Steel Price is at low levels, it shows a substantial positive influence on newbuilding prices and after that the curve fluctuates to some degree. The other variables like Speed appear to have no direct correlations with newbuilding prices at this stage.

In addition, we also make single tests for the orderbook terms, and the results are shown in Table 3.9.

Variable	**Dwt**	**Grain Capacity**	**Horsepower**	**Speed**	**No.Hatches**	**Lead Time**	**Steel Price**	**TC Rate**	**BDI**
R-sq.(adj)	0.742	0.678	0.782	0.286	0.352	0.103	0.134	0.586	0.240
P-value	< 2E-16	< 2E-16	< 2E-16	< 2E-16	< 2E-16	< 2E-16	< 2E-16	< 2E-16	< 2E-16
Significance	***	***	***	***	***	***	***	***	***
Variable	**SH Price**	**SH Index**	**Exchange Rate**	**Inflation**	**LIBOR**	**GDP Per Capita**	**GDP Per Capita Growth**	**Orderbook %Total**	**Orderbook %Group**
R-sq.(adj)	0.609	0.275	0.131	0.159	0.212	0.088	-0.034	0.139	0.499
P-value	< 2E-16	< 2E-16	< 2E-16	< 2E-16	< 2E-16	< 2E-16	< 2E-16	< 2E-16	< 2E-16
Significance	***	***	***	***	***	***	***	***	***

Table 3.8: Statistic Results of Single Tests

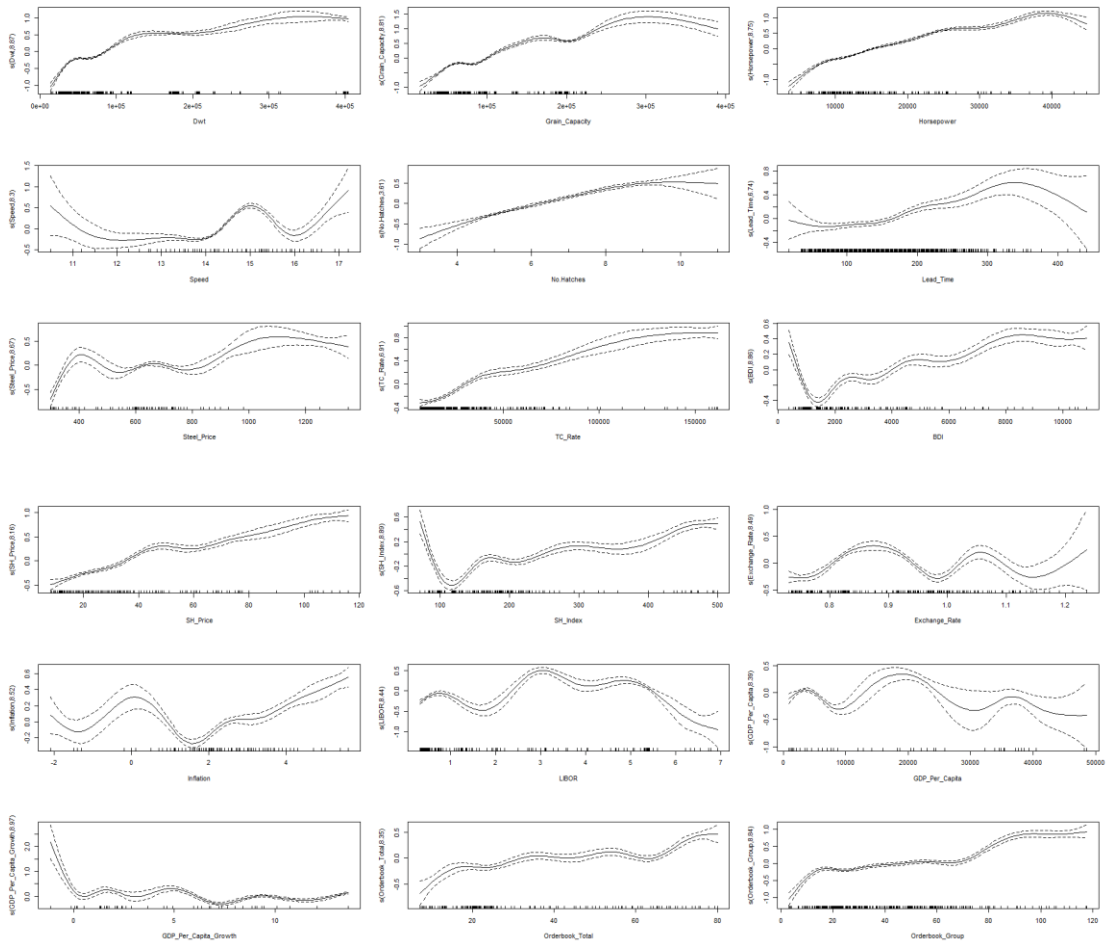

Figure 3.3: Smooth Terms of Single Tests

Variable	**Orderbook %Total**	**Orderbook %Group**	**Orderbook %Total_1**	**Orderbook %Group_1**
R-sq.(adj)	0.139	0.499	0.097	0.220
P-value	< 2E-16	< 2E-16	< 2E-16	< 2E-16
Significance	***	***	***	***
Variable	**Orderbook %Total_3**	**Orderbook %Group_3**	**Orderbook %Total_6**	**Orderbook %Group_6**
R-sq.(adj)	0.157	0.408	0.193	0.439
P-value	< 2E-16	< 2E-16	< 2E-16	< 2E-16
Significance	***	***	***	***

Table 3.9: Single Tests for Orderbook Terms

Obviously, the size of orderbook in the current month gives a higher adjusted R-square, so in this Chapter we do not take the differences between demands into account.

3.5.2 Tests for Multicollinearity

Although the singles tests give many variables with positive effects on newbuilding prices, it is not reasonable to import them into the estimation model directly since there are also strong correlations between them, which is supported by the correlation analysis in Table 3.6. Hence, we need to deal with the multicollinearity issue first. As mentioned in Chapter 2, the easiest way to judge multicollinearity is to see if the variables have appropriate signs. In this subsection, we import the variables with high correlation coefficients in pairs into the model and check if there is multicollinearity.

To exemplify this, we use Dwt versus Grain Capacity, Horsepower and No.Hatches to set up three models and their smooth terms are shown in Figure 3.4. It can be seen that when imported into the model together with Grain Capacity or Horsepower, the curve of Dwt changes a lot compared to that in Figure 3.3, indicating a multicollinearity effect. In contrast, when using Dwt and No.Hatches to establish a model, the curves for both are similar to those given by single tests. Therefore, we need to make a choice between Dwt, Grain Capacity and Horsepower, and considering Dwt as the most representative variable to describe bulk carriers, it is maintained, together with No.Hatches, while Crain Capacity and Horsepower are discarded from the model.

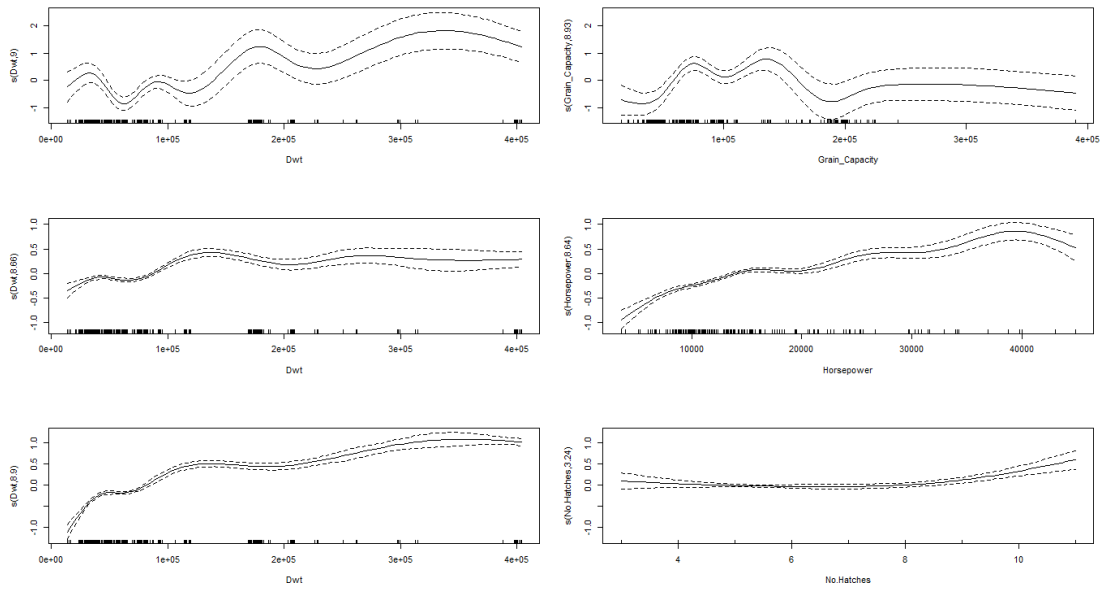

Figure 3.4: Smooth Terms of Multicollinearity Tests for Dwt

The similar tests are taken continually until all the variables with high correlation coefficients are tested and a summary of the multicollineary tests is shown in Table 3.10, where Y indicates multicollinearity, N for non-multicollinearity and Y/N suggests that we cannot judge definitely at this stage.

Dwt	**Grain Capacity**	**Horsepower**	**No.Hatches**		
	Y	Y	N		
TC Rate	**BDI**	**SH Price**	**SH Index**	**Orderbook%Total**	**Orderbook%Group**
	Y	Y	Y	N	N
BDI	**SH Price**	**SH Index**	**Inflation**	**LIBOR**	
	Y	Y	N	N	
SH Price	**Orderbook%Total**	**Orderbook%Group**	**Steel Price**		
	Y/N	N	Y/N		
SH Index	**Orderbook%Total**	**Orderbook%Group**	**Steel Price**		
	N	N	N		
Steel Price	**Orderbook%Total**	**Orderbook%Group**			
	Y/N	N			

Table 3.10: Summary of Multicollinearity Tests

Because TC Rate and BDI are both very good proxies for the shipping market, combined with the fact that the newbuilding market and the sale/purchase market are closely related, it is hard to judge which one is better for the final estimation model. As a result, all four variables are kept for now, as well as the variables with "Y/N".

3.5.3 Model Specifications

This subsection will describe how the estimation model for newbuilding prices is established. For model specifications, we adopt a different grouping compared to the one from Section 3.3 and classify all the remaining variables into two groups: contract-specific variables and macro variables. The former group contains all the variables that are specific to certain contracts, including Dwt, Speed, No.Hatches, Lead Time, Gear, Ice Class Strengthened and Yard Size. The latter group consists of the remaining variables which are not confined to specific vessels or contracts but reflect the current status of the shipping and shipbuilding market, with Exchange Rate, Inflation, LIBOR, GDP Per Capita and Steel Price included as well. By adopting this new variable grouping, the multicollinearity issues are divided into two sides and it is clearer to construct models. In addition, to not miss anything, we will start with the most

influential variables, but generally all the discussed variables will be included in the model and the significance tests will show which variables are not significant and should be discarded.

- **Models with Contract-specific Variables**

From the results of single tests, we can see that, when imported into the model alone, Dwt has an extremely high R^2. Hence, we apply the FS approach here to first set up a model only with this most influential variable, expressed by Equation 3.1. Then other contract-specific numerical variables are also included, expressed by Equation 3.2. In addition to the numerical variables, categorical variables should not be ignored and they are imported in the Model (3.3). However, as shown in Table 3.5, quite a bit of bulk carriers do not have the data for Yard Size, which means including this variable will reduce the number of observations dramatically. Thus, Model (3.4) is introduced which excludes Shipyard Size. The results for all four models are presented and compared in Table 3.11.

$$g\left(E\left(NB_i\right)\right) = \beta_0 + s\left(Dwt\right) \tag{3.1}$$

$$g\left(E\left(NB_i\right)\right) = \beta_0 + s\left(Dwt_i\right) + s\left(Speed_i\right) + s\left(No.Hatches_i\right) + s\left(Lead\ Time_i\right) \tag{3.2}$$

$$\begin{aligned} g\left(E\left(NB_i\right)\right) = {} & \beta_0 + s\left(Dwt_i\right) + s\left(Speed_i\right) + s\left(No.Hatches_i\right) + s\left(Lead\ Time_i\right) \\ & + \beta_1 \cdot Gear + \beta_2 \cdot Ice\ Class + \beta_3 \cdot Strengthened + \beta_4 \cdot Shipyard\ Size \end{aligned} \tag{3.3}$$

$$\begin{aligned} g\left(E\left(NB_i\right)\right) = {} & \beta_0 + s\left(Dwt_i\right) + s\left(Speed_i\right) + s\left(No.Hatches_i\right) + s\left(Lead\ Time_i\right) \\ & + \beta_1 \cdot Gear + \beta_2 \cdot Ice\ Class + \beta_3 \cdot Strengthened \end{aligned} \tag{3.4}$$

Model	(3.1)			(3.2)			(3.3)			(3.4)		
Parametric Terms	Estimate	P-value	Significance	Estimate	P-value	Significance	Estimate	P-value	Significance	Estimate	P-value	Significance
Intercept	3.6874	<2E-16	***	3.6847	<2E-16	***	3.7281	<2E-16	***	3.6999	<2E-16	***
Gear	-	-	-	-	-	-	-0.1325	7.15E-03	**	-0.0365	0.205	
Ice Class_Yes	-	-	-	-	-	-	-0.1044	3.04E-03	**	-0.0914	5.13E-05	***
Strengthened_Yes	-	-	-	-	-	-	0.0016	0.923		0.0215	0.077	.
Yard Size_Mega	-	-	-	-	-	-	0.0435	0.050	.	-	-	-
Yard Size_Medium	-	-	-	-	-	-	0.0757	6.46E-04	***	-	-	-
Yard Size_Small	-	-	-	-	-	-	-0.0374	0.735		-	-	-
Yard Size_Very Small	-	-	-	-	-	-	0.0531	0.247		-	-	-
Smooth Terms	EDF	P-value	Significance	EDF	P-value	Significance	EDF	P-value	Significance	EDF	P-value	Significance
s(Dwt)	8.870	<2E-16	***	8.970	<2E-16	***	8.746	<2E-16	***	8.969	<2E-16	***
s(Speed)	-	-	-	8.295	4.86E-03	**	7.045	0.003	**	8.373	2.06E-03	**
s(No.Hatches)	-	-	-	3.462	4.40E-05	***	3.703	4.01E-08	***	3.611	6.50E-06	***
s(Lead Time)	-	-	-	5.447	<2E-16	***	6.520	1.65E-11	***	5.441	<2E-16	***
R-sq.(adj)	0.742			0.768			0.755			0.767		
GCV	0.05914			0.05317			0.05719			0.05266		
N	1625			1622			928			1622		

Signifi. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 3.11: Statistic Results of Models with Contract-specific Variables

From the comparison between Model (3.1) and (3.2), it can be seen that although not substantial, the model quality is indeed improved after including all the numerical variables. By contrasting Model (3.3) to (3.4), the result of (3.3) indicates that Strengthened is not significant to the model and should be excluded whereas for (3.4) Strengthened has an acceptable significance level. This difference may be caused by the fact that Model (3.3) has too few fixtures and could incur severe random effects. Considering that the estimated value and the significance level of Yard Size is quite low, this variable should be excluded. For these reasons, we can judge that Model (3.2) and (3.4) are superior to (3.1) and (3.3) respectively. Nevertheless, for Model (3.2) and (3.4) we cannot judge which one is better, even considering the statistics results including AIC and the plots of smooth terms (see Appendix A). Consequently, they are both maintained to undergo further checks. Besides, Speed, No.Hatches and Lead Time might be regarded as parametric terms for their relatively steady curves of corresponding smooth terms.

- **Models with Macro Variables**

The models with macro variables are more complex than those with contract-specific variables in that macro variables might be correlated with each other and incur multicollinearity, such as TC Rate versus BDI, and we have to make choices among them. Taking account of the multicollinearity tests above, we set up a series of models, expressed by the equations below, making sure no multicollinear variables are in the same model. The statistic results of all these models are presented in Table 3.12.

$$g\left(E\left(NB_i\right)\right)=\beta_0+s\left(Orderbook\%Total_i\right)+s\left(TC\ Rate_i\right)+s\left(GDP\ Per\ Capita_i\right)+s\left(Steel\ \mathrm{Pr}ice_i\right)+s\left(Exchange\ Rate_i\right)+s\left(Inflation_i\right)+s\left(LIBOR_i\right) \quad (3.5)$$

$$g\left(E\left(NB_i\right)\right)=\beta_0+s\left(Orderbook\%Total_i\right)+s\left(BDI_i\right)+s\left(GDP\ Per\ Capita_i\right)+s\left(Steel\ \mathrm{Pr}ice_i\right)+s\left(Exchange\ Rate_i\right)+s\left(Inflation_i\right)+s\left(LIBOR_i\right) \quad (3.6)$$

$$g\left(E\left(NB_i\right)\right)=\beta_0+s\left(Orderbook\%Total_i\right)+s\left(SH\ \mathrm{Pr}ice_i\right)+s\left(GDP\ Per\ Capita_i\right)+s\left(Steel\ \mathrm{Pr}ice_i\right)+s\left(Exchange\ Rate_i\right)+s\left(Inflation_i\right)+s\left(LIBOR_i\right) \quad (3.7)$$

$$g\left(E\left(NB_i\right)\right)=\beta_0+s\left(Orderbook\%Total_i\right)+s\left(SH\ Index_i\right)+s\left(GDP\ Per\ Capita_i\right)+s\left(Steel\ \mathrm{Pr}ice_i\right)+s\left(Exchange\ Rate_i\right)+s\left(Inflation_i\right)+s\left(LIBOR_i\right) \quad (3.8)$$

$$g\left(E\left(NB_i\right)\right)=\beta_0+s\left(Orderbook\%Group_i\right)+s\left(TC\ Rate_i\right)+s\left(GDP\ Per\ Capita_i\right)+s\left(Steel\ \ \mathrm{Pr}ice_i\right)+s\left(Exchange\ Rate_i\right)+s\left(Inflation_i\right)+s\left(LIBOR_i\right) \quad (3.9)$$

$$g\left(E\left(NB_i\right)\right)=\beta_0+s\left(Orderbook\%Group_i\right)+s\left(BDI_i\right)+s\left(GDP\ Per\ Capita_i\right)+s\left(Steel\ \ \mathrm{Pr}ice_i\right)+s\left(Exchange\ Rate_i\right)+s\left(Inflation_i\right)+s\left(LIBOR_i\right) \quad (3.10)$$

$$g\left(E\left(NB_i\right)\right)=\beta_0+s\left(Orderbook\%Group_i\right)+s\left(SH\ \ \mathrm{Pr}ice_i\right)+s\left(GDP\ Per\ Capita_i\right)+s\left(Steel\ \ \mathrm{Pr}ice_i\right)+s\left(Exchange\ Rate_i\right)+s\left(Inflation_i\right)+s\left(LIBOR_i\right) \quad (3.11)$$

$$g\left(E\left(NB_i\right)\right)=\beta_0+s\left(Orderbook\%Group_i\right)+s\left(SH\ Index_i\right)+s\left(GDP\ Per\ Capita_i\right)+s\left(Steel\ \ \mathrm{Pr}ice_i\right)+s\left(Exchange\ Rate_i\right)+s\left(Inflation_i\right)+s\left(LIBOR_i\right) \quad (3.12)$$

Model	(3.5)			(3.6)			(3.7)			(3.8)		
Parametric Terms	Estimate	P-value	Significance	Estimate	P-value	Significance	Estimate	P-value	Significance	Estimate	P-value	Significance
Intercept	3.6797	<2E-16	***	3.7211	<2E-16	***	3.6743	<2E-16	***	3.7196	<2E-16	***
Smooth Terms	EDF	P-value	Significance	EDF	P-value	Significance	EDF	P-value	Significance	EDF	P-value	Significance
s(Orderbook%Total)	9.000	<2E-16	***	8.673	4.03E-10	***	8.989	<2E-16	***	8.998	2.65E-10	***
s(Orderbook%Group)	-	-	-	-	-	-	-	-	-	-	-	-
s(TC Rate)	8.879	<2E-16	***	-	-	-	-	-	-	-	-	-
s(BDI)	-	-	-	8.462	1.03E-04	***	-	-	-	-	-	-
s(SH Price)	-	-	-	-	-	-	8.345	<2E-16	***	-	-	-
s(SH Index)	-	-	-	-	-	-	-	-	-	8.779	1.55E-13	***
s(GDP Per Capita)	8.999	<2E-16	***	8.650	2.76E-08	***	8.924	<2E-16	***	7.209	7.50E-06	***
s(Steel Price)	8.317	2.43E-13	***	5.507	5.18E-05	***	8.273	<2E-16	***	4.222	5.75E-05	***
s(Exchange Rate)	8.011	<2E-16	***	8.624	9.25E-04	***	8.050	<2E-16	***	8.609	4.74E-05	***
s(Inflation)	8.688	<2E-16	***	8.653	6.60E-11	***	8.600	<2E-16	***	8.766	2.39E-07	***
s(LIBOR)	8.924	<2E-16	***	5.600	4.48E-09	***	8.916	<2E-16	***	4.025	4.55E-05	***
R-sq.(adj)		0.767			0.366			0.809			0.377	
GCV		0.04644			0.13493			0.03475			0.13095	
N		1625			1625			1625			1625	
Model	(3.9)			(3.10)			(3.11)			(3.12)		
Parametric Terms	Estimate	P-value	Significance	Estimate	P-value	Significance	Estimate	P-value	Significance	Estimate	P-value	Significance
Intercept	3.6820	<2E-16	***	3.6944	<2E-16	***	3.6766	<2E-16	***	3.6925	<2E-16	***
Smooth Terms	EDF	P-value	Significance	EDF	P-value	Significance	EDF	P-value	Significance	EDF	P-value	Significance
s(Orderbook%Total)	-	-	-	-	-	-	-	-	-	-	-	-
s(Orderbook%Group)	7.623	<2E-16	***	8.162	<2E-16	***	7.384	1.19E-14	***	8.155	<2E-16	***
s(TC Rate)	8.809	<2E-16	***	-	-	-	-	-	-	-	-	-
s(BDI)	-	-	-	8.680	1.48E-07	***	-	-	-	-	-	-
s(SH Price)	-	-	-	-	-	-	8.019	<2E-16	***	-	-	-
s(SH Index)	-	-	-	-	-	-	-	-	-	8.885	<2E-16	***
s(GDP Per Capita)	8.862	<2E-16	***	8.672	<2E-16	***	8.917	<2E-16	***	8.958	<2E-16	***
s(Steel Price)	3.959	1.39E-13	***	7.256	7.25E-05	***	7.215	<2E-16	***	7.703	2.51E-08	***
s(Exchange Rate)	7.797	<2E-16	***	1.001	0.035	*	7.916	<2E-16	***	7.657	7.62E-06	***
s(Inflation)	7.756	<2E-16	***	8.249	4.75E-10	***	8.716	<2E-16	***	8.613	2.18E-08	***
s(LIBOR)	8.892	<2E-16	***	8.247	<2E-16	***	8.889	<2E-16	***	7.267	3.31E-09	***
R-sq.(adj)		0.752			0.675			0.790			0.696	
GCV		0.05103			0.07724			0.03968			0.07389	
N		1625			1625			1625			1625	

Signifi. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 3.12: Statistic Results of Models with Macro Variables

Table 3.12 indicates that Model (3.7) and (3.11) are statistically superior because of their high R^2 and low GCV scores, while, in contrast, Model (3.6) and (3.8) only obtain bad results, suggesting inappropriate models. Nevertheless, only based on the statistic results, we cannot determine which of two models (3.7) and (3.11) is definitely better than the other, since there might be multicollinearity issues which we have not noticed in the former research. Therefore, it is also necessary to check the plots of smooth terms for both those models.

After checking, it is found that for Model (3.5) and (3.7), the curve of Orderbook%Total changes unexplainably compared to those in Figure 3.3 (see Appendix A). Meanwhile, the same problem occurs to BDI for Model (3.10) and SH Index for Model (3.12) (see Appendix A). Thus, we could conclude that there are multicollinearity issues for these four models. For Model (3.9) and (3.11), the curve of Steel Price also has changed, but at acceptable levels (see Appendix A). However, in order to establish a reliable and robust estimation model, we still exclude Steel Price from the models. In addition, by observing the smooth terms of GDP Per Capita, Exchange Rate, Inflation and LIBOR, we can find that the curves for them are relatively steady, suggesting that we might import them into the model as parametric terms. The new models are expressed by the equations below, with the statistic results given by Table 3.13.

$$g\left(E\left(NB_i\right)\right)=\beta_0+s\left(Orderbook\%Group_i\right)+s\left(TC\ Rate_i\right)+s\left(GDP\ Per\ Capita_i\right)+s\left(Exchange\ Rate_i\right)+s\left(Inflation_i\right)+s\left(LIBOR_i\right) \quad (3.13)$$

$$g\left(E\left(NB_i\right)\right)=\beta_0+s\left(Orderbook\%Group_i\right)+s\left(SH\ \ \mathrm{Pr}ice_i\right)+s\left(GDP\ Per\ Capita_i\right)+s\left(Exchange\ Rate_i\right)+s\left(Inflation_i\right)+s\left(LIBOR_i\right) \quad (3.14)$$

$$g\left(E\left(NB_i\right)\right)=\beta_0+s\left(Orderbook\%Group_i\right)+s\left(TC\ Rate_i\right)+\beta_1\cdot GDP\ Per\ Capita_i+\beta_2\cdot Exchange\ Rate_i+\beta_3\cdot Inflation_i+\beta_4\cdot LIBOR_i \quad (3.15)$$

$$g\left(E\left(NB_i\right)\right)=\beta_0+s\left(Orderbook\%Group_i\right)+s\left(SH\ \ \mathrm{Pr}ice_i\right)+\beta_1\cdot GDP\ Per\ Capita_i+\beta_2\cdot Exchange\ Rate_i+\beta_3\cdot Inflation_i+\beta_4\cdot LIBOR_i \quad (3.16)$$

Model	(3.13)			(3.14)			(3.15)			(3.16)		
Parametric Terms	Estimate	P-value	Significance	Estimate	P-value	Significance	Estimate	P-value	Significance	Estimate	P-value	Significance
Intercept	3.6832	<2E-16	***	3.6789	<2E-16	***	4.3490	<2E-16	***	4.2070	<2E-16	***
GDP Per Capita	-	-	-	-	-	-	6.16E-06	1.24E-10	***	5.85E-06	1.58E-10	***
Exchange Rate	-	-	-	-	-	-	-0.6682	1.91E-09	***	-0.4544	9.98E-06	***
Inflation	-	-	-	-	-	-	-0.0050	9.87E-12	***	-0.0542	6.65E-14	***
LIBOR	-	-	-	-	-	-	0.0015	0.808		-0.0128	0.0363	*
Smooth Terms	EDF	P-value	Significance	EDF	P-value	Significance	EDF	P-value	Significance	EDF	P-value	Significance
s(Orderbook%Group)	8.432	<2E-16	***	8.578	<2E-16	***	8.944	<2E-16	***	8.936	<2E-16	***
s(TC Rate)	8.759	<2E-16	***	-	-	-	7.642	<2E-16	***	-	-	-
s(SH Price)	-	-	-	7.953	<2E-16	***	-	-	-	6.764	<2E-16	***
s(GDP Per Capita)	8.918	<2E-16	***	8.975	<2E-16	***	-	-	-	-	-	-
s(Exchange Rate)	7.683	<2E-16	***	7.953	<2E-16	***	-	-	-	-	-	-
s(Inflation)	7.770	4.31E-16	***	8.142	<2E-16	***	-	-	-	-	-	-
s(LIBOR)	8.769	<2E-16	***	8.755	<2E-16	***	-	-	-	-	-	-
R-sq.(adj)	0.747			0.778			0.675			0.697		
GCV	0.05340			0.04425			0.07758			0.06919		
N	1625			1625			1625			1625		
AIC	11753.33			11445.74			12368.47			12180.39		

Signifi. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 3.13: Statistic Results of Model (3.13), (3.14), (3.15) and (3.16)

By comparing the plots of smooth terms and the statistic results, Model (3.16) appears the best one although its statistic result is not the best, with SH Price as the proxy for

the market status and GDP Per Capita, Exchange Rate, Inflation as parametric terms (see Appendix A).

- **Integrated Models**

After obtaining the suitable models of contract-specific and macro variables, next step is to integrate them to get the final model. There are still two choices to be made: which one of Model (3.2) and (3.4) is better, and whether to import Speed, No.Hatches and Lead Time as parametric terms. Therefore, four integrated models could be established, with their statistic results shown in Table 3.14 below.

Model	(3.17)			(3.18)			(3.19)			(3.20)		
Parametric Terms	Estimate	P-value	Significance	Estimate	P-value	Significance	Estimate	P-value	Significance	Estimate	P-value	Significance
Intercept	3.8540	<2E-16	***	3.8240	<2E-16	***	3.7800	<2E-16	***	3.4290	<2E-16	***
Gear_Yes	0.0780	1.70E-05	***	-	-	-	0.0791	6.38E-05	***	-	-	-
Ice Class_Yes	-0.0139	0.3317		-	-	-	-0.0112	0.4726		-	-	-
Strengthened_Yes	-0.0141	0.0874	.	-	-	-	-0.0205	0.0224		-	-	-
Speed	-	-	-	-	-	-	-7.04E-03	0.3145		6.54E-03	0.426	
No.Hatches	-	-	-	-	-	-	0.0324	8.50E-06	***	0.0559	<2E-16	***
Lead Time	-	-	-	-	-	-	-8.14E-05	0.2864		-9.50E-05	0.1520	
GDP Per Capita	3.01E-06	1.46E-11	***	4.45E-06	<2E-16	***	3.53E-06	5.41E-13	***	4.73E-06	<2E-16	***
Exchange Rate	-0.3374	1.54E-08	***	-0.3135	1.08E-11	***	-0.3799	5.81E-09	***	-0.3680	1.74E-15	***
Inflation	1.42E-02	4.03E-05	***	3.77E-03	0.2440		0.0158	3.50E-05	***	4.68E-03	0.1560	
LIBOR	0.0304	<2E-16	***	0.0266	<2E-16	***	0.0341	<2E-16	***	0.0287	<2E-16	***
Smooth Terms	EDF	P-value	Significance	EDF	P-value	Significance	EDF	P-value	Significance	EDF	P-value	Significance
s(Dwt)	8.958	<2E-16	***	8.981	<2E-16	***	8.948	<2E-16	***	8.919	<2E-16	***
s(Speed)	1.001	0.1913		8.410	4.32E-03	**	-	-	-	-	-	-
s(No.Hatches)	3.975	<2E-16	***	3.255	<2E-16	***	-	-	-	-	-	-
s(Lead Time)	5.950	2.68E-02	*	5.295	0.0159	*	-	-	-	-	-	-
s(Orderbook%Group)	6.839	<2E-16	***	8.094	6.01E-15	***	7.681	<2E-16	***	8.097	<2E-16	***
s(SH Price)	7.332	<2E-16	***	8.616	<2E-16	***	7.697	<2E-16	***	8.738	<2E-16	***
R-sq.(adj)	0.953			0.940			0.946			0.935		
GCV	0.01249			0.14587			0.01455			0.01527		
N	1622			1622			1622			1622		
AIC	9571.38			9620.05			9596.60			9695.06		

Signifi. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 3.14: Statistic Results of Integrated Models

Model (3.20) is selected as the most appropriate estimation model for newbuilding prices by the comprehensive consideration of statistic results, plots of smooth terms and model structure (see Appendix A).From Table 3.14, Speed, Lead Time and Inflation are insignificant to the model, thus should be discarded. Besides, the default value of the basis dimension is 10 and we should double it to see if it is large enough. After trying that, there is indeed a statistically improvement. However, based on the plots of smooth terms, it is found that all three smooth terms are over-fitted. Thus, the basis dimension still remains the same. Eventually, the estimation model for newbuilding prices could be expressed by the following equation, with the statistic result shown in Table 3.15.

$$g\left(E\left(NB_i\right)\right) = \beta_0 + s\left(Dwt_i\right) + s\left(Orderbook\%Group_i\right) + s\left(SH\ \ \mathrm{Pr}ice_i\right) + \beta_1 \cdot No.Hatches + \beta_2 \cdot GDP\ Per\ Capita_i + \beta_3 \cdot Exchange\ Rate_i + \beta_4 \cdot LIBOR_i \quad (3.17)$$

Parametric Terms	Estimate	P-value	Significance	**Smooth Terms**	EDF	P-value	Significance
Intercept	3.5130	<2E-16	***	s(Dwt)	8.919	<2E-16	***
No.Hatches	0.0563	<2E-16	***	s(Orderbook%Group)	8.081	<2E-16	***
GDP Per Capita	4.86E-06	<2E-16	***	s(SH Price)	8.739	<2E-16	***
Exchange Rate	-0.3643	2.23E-15	***	R-sq.(adj)		0.934	
LIBOR	0.0284	<2E-16	***	GCV		0.01527	
				N		1622	

Signifi. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 3.15: Statistic Result of Estimation Model for Newbuilding Prices

- **Estimation Model in Terms of Price/Dwt**

The newbuilding prices could be expressed in two way: in a sales price with a unit of mln USD and in a price/dwt with a unit of USD/ton, which reflects the unit price of all the bulk carriers. The estimation model for the former one has been established, and it is worthwhile to use the latter one instead as a comparison. By applying the similar method from start to finish (see Appendix B), the estimation model for newbuilding prices in terms of Price/Dwt is set up, expressed by the following equation and with the statistic results presented in Table 3.16.

$$g\left(E\left(NB_i\right)\right) = \beta_0 + s\left(Dwt_i\right) + s\left(Orderbook\%Total_i\right) + s\left(BDI_i\right) + \beta_1 \cdot No.Hatches + \beta_2 \cdot GDP\ Per\ Capita_i + \beta_3 \cdot Exchange\ Rate_i + \beta_4 \cdot LIBOR_i \tag{3.18}$$

Parametric Terms	Estimate	P-value	Significance	**Smooth Terms**	EDF	P-value	Significance
Intercept	-7.9020	<2E-16	***	s(Dwt)	8.551	<2E-16	***
No.Hatches	0.0569	<2E-16	***	s(Orderbook%Total)	8.767	<2E-16	***
GDP Per Capita	4.46E-06	<2E-16	***	s(BDI)	8.892	<2E-16	***
Exchange Rate	-0.2073	5.09E-05	***	R-sq.(adj)		0.875	
LIBOR	3.97E-02	<2E-16	***	GCV		0.01416	
				N		1622	

Signifi. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 3.16: Statistic Results of Estimation Model for Price/Dwt

In contrast, the two models have similar structure, except that in the model with Price/Dwt, Orderbook%Group is replaced by Orderbook%Total and SH Price replaced by BDI. This difference here could be explained that Orderbook%Group and SH Price distinguish size groups whereas the other two represent the whole bulk carrier fleet of all sizes. When researching on Price/Dwt, the size of vessels is not important, thus Orderbook%Total and BDI suit the model better.

3.6 Results Analysis

Combing Table 3.15 and 3.16, the following conclusions about parametric terms could

be made. First, bulk carriers with more hatches deserve higher newbuilding prices while whether the vessels have Ice Class or are strengthened for heavy cargo appear not influential. Second, from the perspective of macro variables, GDP Per Capital, namely wage levels and LIBOR have a positive effect on newbuilding prices while Exchange Rate has a negative influence, all of which correspond to the initial expectations mentioned in Subsection 3.4.4. When it comes to smooth terms of the models, it is advisable to refer to the models for Price and Price/Dwt respectively, with the plots shown below, where the upper three panels are for Price and the lower three are for Price/Dwt.

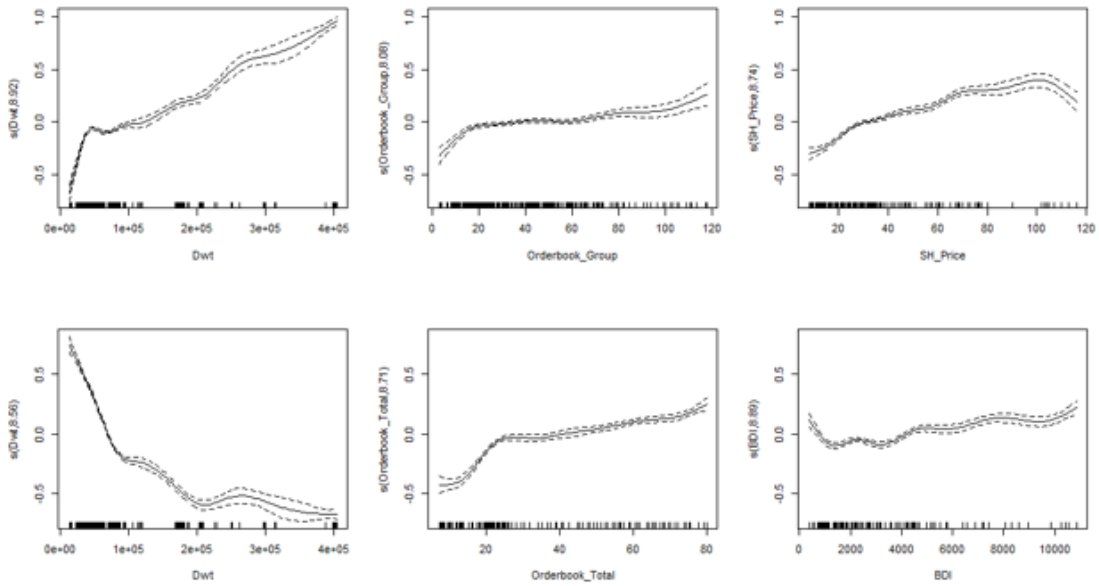

Figure 3.5: Smooth Terms of Models for Price and Price/Dwt

The upper left plot shows the relationship between Dwt and Price that the sale prices of new bulk carriers are positively related to Dwt, especially true for Handysize and Handymax, where the prices arise dramatically with the increase of Dwt. In contrast, the lower left plot indicates a totally opposite trend that, as Dwt increases, Price/Dwt declines abruptly at first and then drops relatively steady. This difference is logical that larger bulk carriers generally have higher sale prices for their higher shipbuilding costs and the economy of scale makes vessels cheaper to build with respect to Price per Dwt.

The upper middle plot and lower middle plot both indicate the positive influence of demand for new bulk carriers on newbuilding prices with similar curve trends. When

the orderbook size is at low levels, newbuilding prices arise with that drastically. After the orderbook size becomes large, it still has a positive effect on newbuilidng prices though not as considerable as before. From these results, we could assume that the expectations of oversupply for bulk carriers are not influential to newbuilding prices, at least for this dataset.

The upper right plot suggests how second-hand prices affect newbuilding prices. The sale prices of vessels generally increase as second-hand prices grow, except at the end of the curve there is a decline of newbuilding prices. This exception might result from the low density of data within that range. The lower right plot indicates the relationship between Price/Dwt and another shipping market proxy, BDI. This plot also suggests a generally positive relationship between BDI and newbuilding prices, except the start of the curve. By comparing the lower right plot to the plot of BDI in single tests for Price/Dwt, we could conclude that the abnormality of the curve start results from the presence of the multicollinearity. However, as [13] suggests, this level of multicollinearity could be ignored.

3.7 Summary

This chapter introduces, given data sources, how to use GAM to develop a reliable estimation model for bulk carrier newbuilding prices.

According to previous researches and further analysis, there are many variables theoretically having influences on bulk carrier newbuilding prices, which could be generally classified into three groups: cost-related variables, asset-pricing related variables and supply/demand related variables. However, it turns out that when setting up the estimation model, contract-specific variables and macro variables seem a better grouping since these two groups are completely irrelevant to each other. Besides, some initially determined variables are not included in the final model for low levels of significance.

Among all the remaining variables for estimating newbuilding prices, Dwt it the most

influential one, which could be concluded from the combination of single tests and plots of smooth terms. Other parametric terms including No.Hatches, GDP Per Capita and LIBOR are also positively related to newbuilding prices. Therefore, it could be concluded that newbuilding prices are highly cost related, which is in line with the economic theory stating that newbuiliding prices are principally cost driven [13]. In addition to that, by comparing the models for Price or Price per Dwt, it could be suggested that the economy of scale is also effective for building vessels.

Furthermore, how the second-hand prices affect newbuilding prices are explained. Besides, the positive influences of the orderbook size on newbuilding prices indicate that the urge to earn more overweighs the fear of the potential oversupply of new bulk carriers.

4 Lead Times of Bulk Carriers

4.1 Introduction

Lead time, or more specifically manufacturing lead time, is the time period between the placement of an order and the shipment of the completed order to the customer. There are many factors affecting the lead time, including producer's capacity, scheduling, batching, product's complexity, and etc. [20].

From the standpoint of the customer, lead time can also be called delivery time. The relevance between lead time and customer's satisfaction has been verified and the great importance of accurate estimating this performance indicator is obvious [21].

This chapter will firstly introduce previous researches about lead time estimation. Secondly, how to identify the variables of a bulk carrier's lead time, or more specifically the contract lead time (the latter one to be consistent with Clarksons) will be illustrated. After the identification of the variables, the choice of data will be discussed, followed by the filtering of the data and related analysis. Next, how to use GAM to set up a lead time estimation model will be illustrated, combined with necessary model checks and results analysis.

4.2 Literature Review

Considering that there are few specific studies on the lead time estimation in the shipbuilding industry, the reference to other industries will be necessary. In the shipbuilding industry, the product, namely the ship, is designed, engineered and finished after an order has been received. The product needs to be engineered to meet the specifications desired by the received order. Thus, the shipbuilding industry can be classified as the engineer-to-order (ETO) industry [31], in which field various estimating methods have been proposed.

Up to now, various approaches of estimating lead time can be generally divided into several categories (Figure 4.1): simulation, logistic curves, queuing theory, statistics, stochastic analysis, artificial intelligent analysis and hybrid systems (which combines

two or more of the mentioned methods) [32].

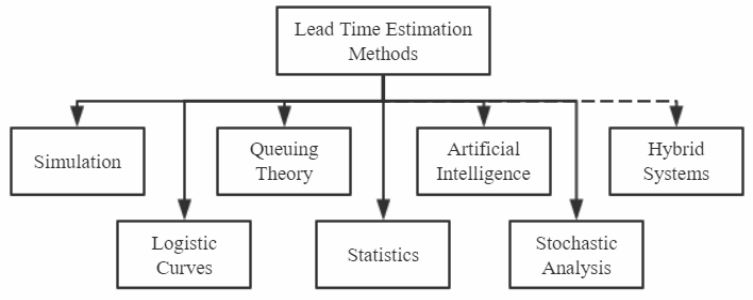

Figure 4.1: Lead Time Estimation Methods

Nyhuis [33]proposed a simulation model using logistic operating curves to predict the logistic performance of a production department in the work-in-process (WIP) level, where it's shown that the throughput time (lead time) is increasing with the higher WIP level of the work system. Solomon [34], based on actual data, used Monte Carlo and Arena simulation to acquire lead time estimation in leather shoe manufacturing. Chryssolouris [35]came up with discrete event simulation models to shorten lead time. Kawasaki [36]modeled a low-carbon supply chain network on a discrete event simulation, evaluated multi criteria decisions for the lead times, and analyzed the effect of the fluctuating lead time.

In order to be able to reduce the effort of extensive simulation experiments at least partly, Nyhuis in [33] also attempted to develop approximation equations to calculate the logistic operating curves and described logistic performance measuring output rate, throughput time (lead time) and WIP level and their interdependencies within a work system. Again in [33], the authors formulated a queuing theory model to predict a work system's logistic performance including lead time. Besides, Parlar [37]created an inventory model combined with queuing theory to consider lead time as a random parameter.

Simulation, logistic curves and queuing theory can build models to predict the logistic performance of a production system, hereby estimating the lead time, but they all have inevitable disadvantages [33]. Simulation requires high efforts in application phase, impossible to be validated generally and hard to draw general conclusions. Logistic

curves theory and queuing theory both need high efforts in definition phase, can only be valid for steady operating states and are limited to resource perspective. Moreover, for queuing theory, the models cannot be adapted very flexibly and the parameters may not conform to reality.

Seyedhosseini [38]proposed a step-by-step stochastic analysis method, and also by using non-parametric Kernel estimation methods, to select the best future state map, thereby evaluating and analyzing the production lead time. Pfeiffer [39]introduced a novel method to select tuning parameters improving accuracy and robustness for multi-model based prediction of manufacturing lead times with the combination of simulation and statistical learning methods.

At the present stage, Artificial Intelligent (AI) Methods are the most reliable and robust methods for estimating lead time, which can be grouped into the following types (Figure 4.2): data mining, expert systems, neural networks, genetic algorithms, fuzzy logic, case based reasoning and hybrid methods [32].

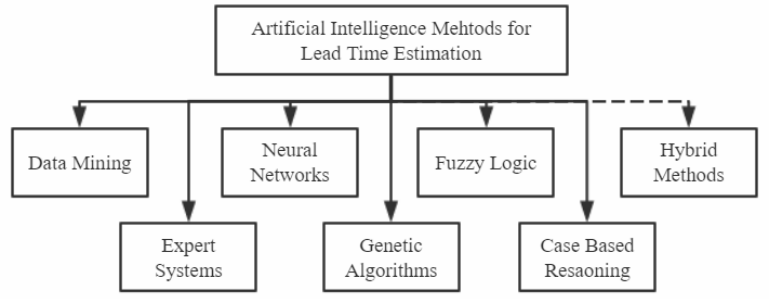

Figure 4.2: Artificial Intelligent Methods for Estimating Lead Time

Ozturk [40]explored the use of data mining for lead time estimation in make-to-order manufacturing and chose the regression tree as the specific data mining method. Okubo [41]showed a basic neural network model for lead time estimation, examined the validity of this model comparing it with the conventional estimation formula, and tried to solve the accuracy of the estimation due to neural network losses by means of an addition of the input parameters. Mourtzis [32]estimated the manufacturing lead time for complex engineered-to-order products by using Cased Based Reasoning (CBR).

Given the easy access to mass related data, the method of data mining is distinguished.

The specific classification of data mining is given in [40] (Figure 4.3).

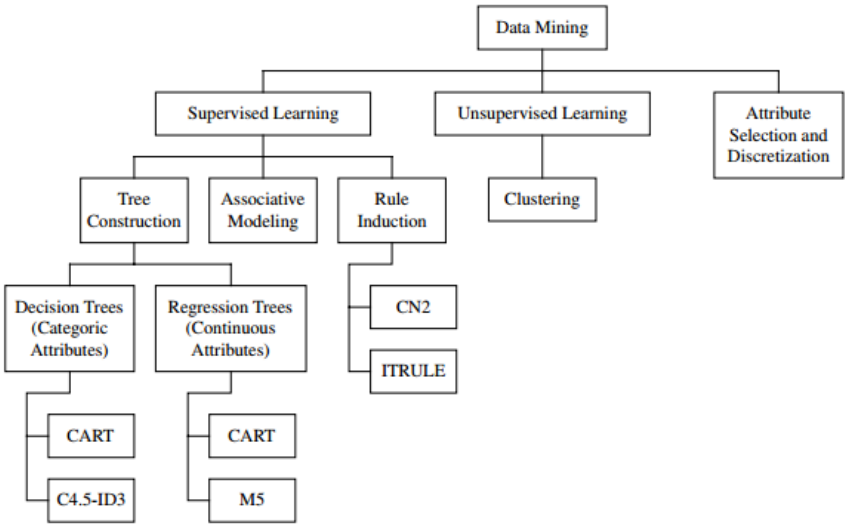

Figure 4.3: Classification of Data Mining

Among all the data mining methods, regression is the one used to predict a range of numeric values (also called continuous values), given a particular dataset, and the GAM we use in this research is a specific type of that.

4.3 Variables Identification

Lead time is a critical measure of manufacturing performance for most industries, but it has not received a great deal of attention in the literature. Generally, it is believed that lead time is mainly affected by the producer's capacity, scheduling, batching, and product's complexity, and etc. Given the available data and information, the variables of lead times for bulk carriers are categorized into three groups: shipyard related, vessel related and market related. In order to not miss anything crucial, all the potential influential variables based on some references and reasonable assumptions are discussed.

4.3.1 Shipyard-related Variables

As one of the major components of lead time, a vessel's building time requires special

attention, which is strictly dependent on the shipyard capacity. Shipyard capacity is mainly in terms of total area, erection area and capacity for moving blocks [42], affecting overall productivity and building time. The shipyard's total area and erection area clearly have positive effects on building time, but since total area and erection area are generally highly correlated, which means a choice is needed between them. Considering that a shipyard normally has a large nonoperational area, erection area is a more robust and representative indicator. However, the data of erection area for different shipyards is much harder to collect than total area, so which one to use can only be determined after further analysis of available data. Moving capacity is critical in the short term, but it is not a long-term or permanent bottleneck [42], and the number of gantry cranes is considered.

When it comes to manufacturing industries, facilities can never be neglected. For shipbuilding, docks and berths for building are the most representative ones, and the number of them can somehow reflect a shipyard's building capacity.

Shipbuilding is a labor-intensive industry, so the workforce conditions should be focused on. Usually, a factory's workforce is evaluated from different aspects, such as worker education level, worker average age, availability of qualified workers and etc. And in this case, the total workforce, including permanent employees and contracted employees, is taken into account.

Intuitively, the maximum annual Dwt output of all the vessels during a certain period can represent a shipyard's capacity over that time, so it is another proxy for lead times.

As mentioned above, lead times are significantly affected by scheduling and batching. Shipbuilding also encounters this situation, because the building process of a new ship must follows a strict sequence, and due to the limitations of equipment and personnel, the building of a new ship will be blocked because of the remained orders. For this, the capacity utilization rate is introduced, which measures the proportion of potential economic output is actually realized, to reflect how busy the shipyards are. In this case, the capacity utilization rate could be expressed by the following equation, where the

yearly Dwt output and Max Dwt output represent the actual total output and potential total output of shipyards respectively.

$$Capacity\ Utilization\ Rate = \frac{DWT\ Output\ in\ Year\ x}{Max\ DWT\ Output\ during\ the\ Period} \times 100$$

Obviously, a shipyard's technological level has a significant influence on its building capability, thereby the performance of lead times. But the evaluation of a shipyard's technological level is complex and hard to be quantified, so this indicator is not considered. The evaluation of a shipyard's management level and operational strategies encounters the same problem.

For a vessel, the time to obtain critical inputs is crucial for its lead time. For instance, in the present newbuilding market, the main determinant of lead times, in many cases, is the time needed to obtain the main engines for a vessel [42]. However, the data of this time for different shipyards, again, is hard to gather in the thesis and it can only be excluded.

Apart from the shipyard's own attributes (like facilities, workforce and technological level), the shipyard's building capability also depends on its country or region. Thanks to the application of GAM in the thesis, the shipyard's country can be regarded as a categorical variable and used in the later analysis. Moreover, the introduction of this indicator can somehow make up for lacking the detailed evaluation of the shipyard's workforce, technological and management level, because the levels of those performance , in general, will not fluctuate much among different shipyards in a certain region .

4.3.2 Vessel-related Variables

Usually, a product's building time is closely related to its own complexity, and this is also the case for a vessel. But the evaluation of a vessel's complexity is generally systematical and complicated and hard to make on the basis of available datasets in this research. Therefore, the main focus is put on the influences of the ship's main attributes

on its building time (lead time), in the expectation that these may reflect complexity to some extent.

Firstly, it is generally believed that the size of a vessel will affect its building time. Similar to the estimation of newbuilding prices, DWT is selected as the sole proxy for a vessel's size. Then, the total horsepower and speed, as important attributes of a ship, are also considered. In addition, hull type and the number of decks can also somehow influence a ship's building difficulty. What's more, whether the ship has Ice Class or is strengthened for heavy cargo should be in consideration as well.

In addition to the vessel's own attributes, the influences of newbuilding prices cannot be missed. Because the newbuilding price of a ship is determined after lots of negotiations and compromises between shipyards and shipowners. And among those negotiations, when to deliver the ship deserves much attention in that almost all the shipowners want ships delivered as soon as possible and some of them are willing to pay more for making their own ships prior to others, thereby reducing the lead times.

4.3.3 Market-related Variables

From the perspective of the market, a prosperous shipping market will stimulate the demand of ships, thereby increasing the orders of corresponding ships. According to [43], high ordering activity is always followed by longer lead times. Therefore, time charter rates of bulk carriers, which can reflect the prosperity of bulk shipping markets, and the orderbook size are taken into account. Besides, interest rates, representing the cost of financing, should also be considered and London Inter-Bank Offered Rate (LIBOR) is selected.

In a word, the potential variables of lead times for bulk carriers can be preliminarily summarized as follows.

- Shipyard-related variables: total area, number of cranes and building facilities, number of permanent and contracted employees, maximum annual Dwt output,

capacity utilization rate, country or region;

- Vessel-related variables: Dwt, horsepower, speed, number of decks, hull type, Ice Class, strengthened for heavy cargo, newbuilding price;
- Market-related variables: time charter rates, vessel orderbook, LIBOR.

4.4 Data Collection and Analysis

4.4.1 Data Source

Just like the estimation for newbuilding prices, the data of this part is mainly from Clarksons Sources as well. Besides, the data from websites of shipyards and other sources are also necessary in case there are contradictions or missing data.

4.4.2 Data Scope

Considering that the analysis is highly dependent on data of shipyards, to make the estimation more robust and reliable, the data scope is constrained to Top 50 bulk carrier builders all over the world, which built majority of bulk carriers so far, including Oshima Shipbuilding, Tsuneishi Zosen and etc. Then, the time range of the research should be determined. By looking through the annual orderbook development of those shipyards, it is found that before Year 2006 and after 2017 much data is unrecorded for many of the shipyards. So the time range is set to the period between 2006 and 2017. Furthermore, the data of four shipyards (Sanoyas Shipbuilding, Sanoyas, Koyo Dock and I.H.I. Yokohama) during that period is incomplete and cannot be taken into account. Eventually, the preliminary data scope is determined: bulk carries of 46 shipyards whose contract dates are from 01/01/2006 to 31/12/2017, and there are 3986 bulk carriers in total, among which there are 748 Handysize, 1526 Handymax, 1081 Panamax and 631 Capesize.

4.4.3 Problems and Options Associated with Data

After the initial organization of the collected data, it is found that there are lots of missing data for certain terms and contradictions between data for from different sources for a same term. The problems and some options to be made are as follows.

1) Shipyard's total area or erection area: The dataset for this term, combined with many contradictions and missing data, is mainly from the shipyard profiles downloaded from Clarksons Sources and the websites of those shipyards. In the shipyard profiles, a shipyard's area information is often given in terms of "Yard Space and Workshops Fabrication" whereas "the shipyard covers an area of " is a more common expression on the shipyard websites. Thus contradictions may occur. For example, it is said that the Beihai Shipyard has a "Yard Space and Workshops Fabrication" of 535353 square meters while this shipyard covers an area of 3300000 square meters according to its website. The reasons for this paradox may be the differences of measuring methods or just the errors of records. After checking all the shipyard profiles and websites, the data of total area for a shipyard seems more practical for this research. However, for some shipyards, like Hakodate Dock and Namura Shipbuilding, no data related to areas can be found. So the bulk carriers produced by these shipyards cannot be researched further.

2) Number of gantry cranes: This indicator is selected to represent the moving capacity of shipyards. However, only a few shipyards can provide the information of this indicator and some of shipyards only give the data of annual crane lifting capacities. Faced with this situation, this indicator can only be excluded.

3) Number of docks and berths for building: There are mainly three types of facilities used for ship construction, including dry docks, floating docks and some berths. The World Fleet Register can provide the data on the number of dry docks, floating docks and berths for different shipyards. But the direct use of this data is unreasonable because of the following situations: in reality, some dry docks are used for shipbuilding while some are specially designed for repairing, but the data from Clarksons will not distinguish this difference; some shipyards use berths for building instead of dry docks and floating docks, while the number of berths may

not be recorded by Clarksons. Therefore, to get the relatively authentic data of building facilities, shipyard profiles and websites are of necessity, and after comparisons and integration, the number of building facilities for different shipyards can be determined roughly.

4) Number of permanent and contracted employees: this indicator cannot be regarded as a robust input for the estimation model because the total workforce of different shipyards generally fluctuates much annually and the exact fluctuations are not available without further investigations. Besides, the records of this indicator from Clarksons and shipyard websites for a same shipyard can be totally different and we are not sure which one is more reliable.

5) Hull type and deck number of bulk carriers: unfortunately, the data of these two indicators are not available in the World Fleet Register and to investigate these one by one is not realistic.

6) Bulk carrier's speed: some bulk carriers do not have a speed in record, so the missing data is completed by referring to the data of its peer group and the bulk carriers with a similar horsepower and dimensions.

7) Bulk carrier's newbuilidng prices: the newbuilding prices of ships for a shipyard is generally private so the record of this indicator is not complete, and there are only 703 bulk carriers recorded with newbuilidng prices among 3986 bulk carriers in total. Therefore, this indicator might not be included in the final estimation model.

8) Time charter rates and vessel orderbook: Similar to Chapter 3, we take "1 Year Time charter Rate in Long Run Historical Series" and the orderbook as a percentage of the total fleet or size group, both monthly.

4.4.4 Data Pre-processing

Variable	Unit	N	Mean	Std.Dev.	Min	Max
Lead Time	week	3986	143.18	58.46	4.00	485.00
Total Area	m^2	3697	1,437,024	1,437,024	50,000	9,300,000
No.Building Facilities	\	3986	2.58	1.21	1.00	9.00
Max DWT Output	\	3986	2,254,000	1,903,187	135,220	12,654,828
Utilization Rate	%	3986	54.85	27.63	0.32	100.00
DWT	ton	3986	79,670	47,447	12,427	216,656
Horsepower	\	3986	14,085	5,008	1,026	34,099
Speed	knot	3986	14.47	0.89	9.60	17.60
Price (NB)	$m	703	45.66	45.66	4.10	240.00
Timecharter Rate	$/day	3986	28,908	27,790	4,875	161,600
Orderbook%Total Fleet	%	3986	41.39	20.61	8.68	79.99
Orderbook%Per Group	%	3986	39.89	21.26	6.47	118.25
LIBOR	%	3986	2.118	2.103	0.323	5.594

Table 4.1: Summary Statistics before Pre-processing

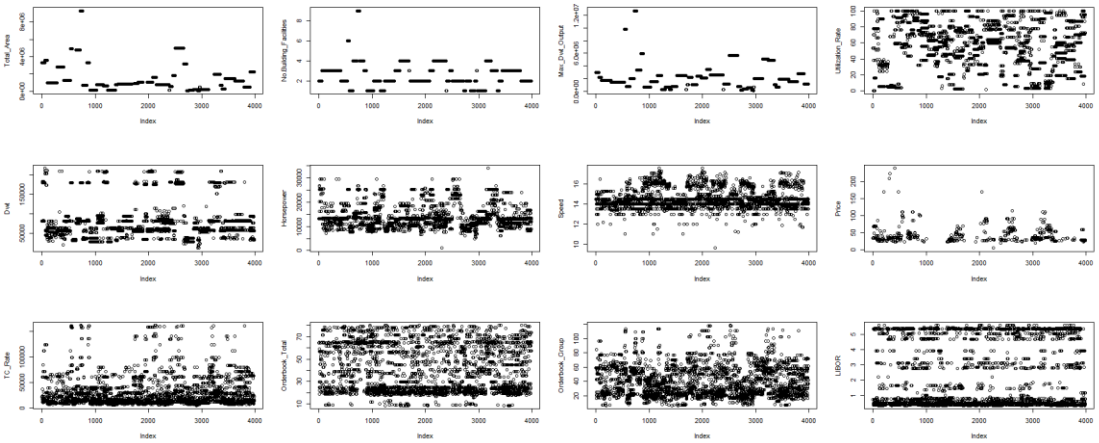

Figure 4.4: Overview of Numerical Variables

Table 4.1 shows the summary statistics of the initial dataset. It must be mentioned here that because Clarksons Sources does not provide the exact lead times, so the lead times in Table 4.1 and following research are actually the duration between contract dates and built dates of vessels. Figure 4.1 presents the distributions of all the numerical variables. Here, the dataset is also pretreated by the same method of Chapter 3 to delete outliers, and the summary statistics of the dataset after pre-processing is shown in Table 4.2.

Variable	Unit	N	Mean	Std.Dev.	Min	Max
Lead Time	week	3801	143.77	58.25	22.00	417.00
Total Area	m^2	3517	1,290,196	1,290,196	50,000	5,000,000
No.Building Facilities	\	3801	2.47	0.95	1.00	4.00
Max DWT Output	\	3801	2,027,682	1,353,889	135,220	5,916,058
Utilization Rate	%	3801	54.01	27.49	1.43	100.00
DWT	ton	3801	77,051	45,386	16,656	211,182
Horsepower	\	3801	13,789	4,762	5,710	29,694
Speed	knot	3801	14.47	0.87	11.50	17.30
Price (NB)	$m	645	40.70	40.70	18.50	115.00
Timecharter Rate	$/day	3801	26,911	23,680	5,200	145,000
Orderbook%Total Fleet	%	3801	40.66	20.49	8.68	79.99
Orderbook%Per Group	%	3801	38.57	20.05	6.47	118.25
LIBOR	%	3801	2.104	2.105	0.323	5.594

Table 4.2: Summary Statistics after Pre-processing

4.4.5 Data Analysis

From Table 4.2, it can be seen that, although the dataset has been pretreated, there is still a huge gap between the sizes of different shipyards. The mega shipyards can cover an area of 5,000,000 square meters and the max annual DWT output can be up to 5,916,058 tons. In contrast, the small shipyards only have a few building facilities and the max DWT output can be very low, 135,220 tons only. Shipyards can be very busy, with a capacity utilization rate up to 100%, while sometimes they can only receive few orders and the minimum utilization rate is 1.43%. The ship size ranges from 16,656 dwt to 211,182 dwt, having a horsepower between 5,710 and 29,694 and a speed between 11.5 knots and 17.3 knots. It is obvious that there is only limited information for newbuilding prices, with 645 observations. So the estimations including this variable will have a much smaller data size. As for the variables related to the market status, the extreme values occur during the boom and bust of the world economy in the period between 2006 and 2009.

In addition to the data presented in Table 4.2, the dataset also includes some categorical variables: builder country, Ice Class and strengthened for heavy cargo, shown in Table 4.3.

Builder Country	%	**Ice Class**	%	**Strengthened for Heavy Cargo**	%
China P.R.	47.70	Yes	5.95	Yes	58.14
Japan	39.31	No	94.05	No	41.86
South Korea	8.18				
Philippines	4.81				

Table 4.3: Summary of Categorical Variables after Pre-processing

It can be seen that all the involved bulk carriers are built in four countries: China, Japan, South Korea and Philippines, among which China and Japan account for majority of the output, up to 87.01%. Most bulk carriers do not have Ice Class, which is reasonable because only a few bulk carriers will voyage in ice zones. More than a half of bulk carriers are strengthened for heavy cargo, suggesting that the demand for carrying heavy cargos is substantial.

To complement the data analysis, the pairwise correlations between all the variables of

lead times are presented in Table 4.4 and relatively strong relations are highlighted by bold numbers. Among the three variable groups, the market-related variables seem to have most substantial relations with lead times, and all the correlation coefficients are positive, indicating that a prosperous shipping market may result in the increase of lead times, which seems logical. As for the vessel-related variables, Dwt and Horsepower are also positively related to lead times. In contrast, variables about shipyards still do not have evident correlations with lead times.

When it comes to the categorical variables, two preliminary judgements can be made: China and South Korea may provide an industrial environment with shorter lead times than that of Japan and Philippines; if a bulk carrier is required to have Ice Class and be strengthened for heavy cargos, it will take a longer lead time to deliver the vessel.

Apart from the correlations with lead times, the correlations between the variables of themselves also deserve much attention, for the input of two strongly correlated variables in a regression model may lead to the multicollinearity issue. For instance, Dwt is strongly related to Horsepower, with a correlation coefficient of more than 0.9, so we should be careful when involving these two variables together. The same issue occurs for Price and TC Rate.

	Lead Time	Total Area	No. Building Facilities	Max Dwt Output	Utilization Rate	Builder Country_ China	Builder Country_ Japan	Builder Country_ South Korea	Builder Country_ Philippines	Dwt	Horsepower	Speed	Ice Class _Yes	Strengthened For Heavy Cargo_Yes	Price	TC Rate	Orderbook %Total	Orderbook %Group	LIBOR
Lead Time	1																		
Total Area	0.0655	1																	
No.Building Facilities	0.0384	**0.2324**	1																
Max Dwt Output	0.0538	**0.5701**	**0.3681**	1															
Utilization Rate	0.0015	0.0492	0.0434	0.0490	1														
Builder Country_China	-0.0800	**0.4149**	0.0992	0.0109	-0.1661	1													
Builder Country_Japan	0.0738	**-0.4910**	**-0.3498**	**-0.3361**	0.1851	-	1												
Builder Country_South Korea	-0.0723	0.0933	**0.3450**	**0.5877**	0.0233	-	-	1											
Builder Country_Phi	**0.1109**	0.0335	**0.1250**	-0.0111	-0.0646	-	-	-	1										
Dwt	**0.1405**	-	-	-	-	-	-	-	-	1									
Horsepower	**0.1504**	-	-	-	-	-	-	-	-	**0.9099**	1								
Speed	-0.0267	-	-	-	-	-	-	-	-	**0.2156**	**0.1669**	1							
Ice Class_Yes	0.0546	-	-	-	-	-	-	-	-	**-0.1730**	**-0.1334**	**-0.1560**	1						
Strengthened For Heavy Cargo_Yes	**0.1173**	-	-	-	-	-	-	-	-	**-0.1137**	-0.0948	-0.0036	**-0.1204**	1					
Price	**0.3361**	-	-	-	-	-	-	-	-	**0.7958**	**0.7796**	**0.2512**	**-0.1729**	-0.0198	1				
TC Rate	**0.4046**	-	-	-	0.0256	-	-	-	-	-	-	-	-	-	**0.8126**	1			
Orderbook%Total	**0.1242**	-	-	-	**0.3657**	-	-	-	-	-	-	-	-	-	**0.2194**	**0.4723**	1		
Orderbook%Group	**0.1366**	-	-	-	**0.2963**	-	-	-	-	-	-	-	-	-	**0.4347**	**0.5873**	**0.8460**	1	
LIBOR	**0.4623**	-	-	-	**-0.2539**	-	-	-	-	-	-	-	-	-	**0.4251**	**0.6025**	**0.1173**	**0.1219**	1

Table 4.4: Correlations between Variables

4.5 Method of Establishing Models

4.5.1 Single Tests of Numerical Variables

Just like what we do to newbuilding prices, single tests of all the numerical variables related to lead times are also necessary, with Table 4.5 showing the statistic results and Figure 4.5 presenting the plots of smooth terms.

Table 4.5 shows a low level of R^2 for all the variables and Figure 4.5 indicates that none obvious relations between a single variable and lead times can be observed. Considering both of them, all the numerical variables appear to not relate to lead times.

Variable	**Total Area**	**No.Building Facilities**	**Max Dwt Output**	**Utilization Rate**	**Dwt**	**Horsepower**
R-sq.(adj)	0.0421	0.00209	0.0345	0.0218	0.0426	0.0335
P-value	< 2E-16	< 2E-16	< 2E-16	< 2E-16	< 2E-16	< 2E-16
Significance	***	***	***	***	***	***
Variable	**Speed**	**Price**	**TC Rate**	**Orderbook %Total**	**Orderbook %Group**	**LIBOR**
R-sq.(adj)	0.0172	0.142	0.196	0.0823	0.036	0.257
P-value	< 2E-16	< 2E-16	< 2E-16	< 2E-16	< 2E-16	< 2E-16
Significance	***	***	***	***	***	***

Table 4.5: Statistic Results of Single Tests

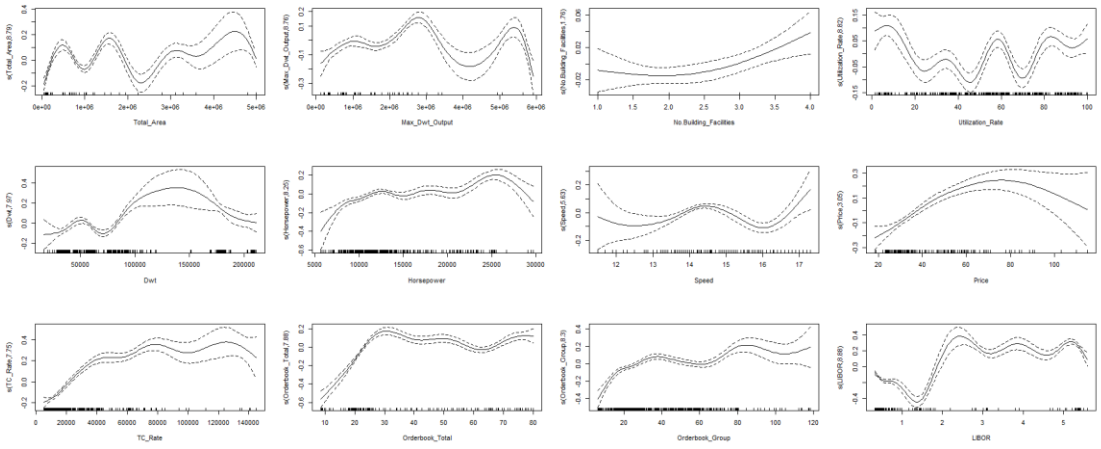

Figure 4.5: Smooth Terms of Single Tests

4.5.2 Tests for Multicollinearity

Similarly, for the variables with relatively high correlation coefficients, tests to check

multicollinearity should also be taken. The test objects include Total Area versus Max Dwt Output, Utilization Rate versus Orderbook%Total, Dwt versus Horsepower, TC Rate versus Orderbook%Total, Orderbook%Group and LIBOR. By checking the results of models where these variables are imported included in pairs and observing the plots of corresponding smooth terms, we assume that the multicollinearity issues for these variables could be ignored for now.

4.5.3 Model Specifications

In Section 4.3, all the variables of lead times has been identified and can be divided into three groups. According to this, we can classify market models, market-vessel models and market-vessel-shipyard models. Considering that the R^2 for every single variable is quite low, the application of FS or BE approach cannot make much difference, so for this part, all the variables will be imported into the model together and the significance tests will judge which one should be discarded or included.

- **Market Models**

Again, for market models, we need to distinguish the variable Orderbook%Total and Orderbook% Group, and the equations below can express models with one of the two variables respectively.

$$g\left(E\left(LT_i\right)\right) = \beta_0 + s\left(TC\ Rate_i\right) + s\left(Orderbook\%Total_i\right) + s\left(LIBOR_i\right) \tag{4.1}$$

$$g\left(E\left(LT_i\right)\right) = \beta_0 + s\left(TC\ Rate_i\right) + s\left(Orderbook\%Group_i\right) + s\left(LIBOR\right) \tag{4.2}$$

The model results are shown in Table 4.6. Although there is differences between the results with respect to R^2 and GCV scores, it is not obvious to judge which model is better, suggesting that both model are remained at this stage.

The smooth terms of two model all have an effective DF of more than 5, indicating a clear non-linearity. Besides all the smooth terms have an extremely small p-value, which means they are all significant to lead times. However, the adjusted R-square shows that the fitting result is not satisfactory that only about 28% of the variances of

lead times can be explained by the market-related variables.

Model	(4.1)			(4.2)		
Parametric Terms	Estimate	P-value	Significance	Estimate	P-value	Significance
Intercept	4.94541	< 2E-16	***	4.94575	< 2E-16	***
Smooth Terms	EDF	P-value	Significance	EDF	P-value	Significance
s(TC Rate)	8.779	9.54E-13	***	8.773	9.54E-13	***
s(Orderbook%Total)	5.743	3.85E-08	***	-	-	-
s(Orderbook%Group)	-	-	-	8.209	3.09E-03	**
s(LIBOR)	8.621	< 2E-16	***	8.810	< 2E-16	***
R-sq.(adj)		0.282			0.274	
GCV		0.12297			0.12383	
N		3801			3801	
AIC		40032.73			40059.76	

Signifi. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 4.6: Statistic Results of Market Models

- **Market-vessel Models**

The single market model cannot provide a good fitting result, so the variables of vessels should also be taken into account, including that whether the bulk carrier has Ice Class and whether it is strengthened for heavy cargo are considered as dummy variables. The market-vessel model can be expressed by the following equations and the statistic results are presented in Table 4.7.

$$g\left(E\left(LT_i\right)\right)=\beta_0+s\left(TC\ Rate_i\right)+s\left(Orderbook\%Total_i\right)+s\left(LIBOR_i\right)+s\left(Dwt_i\right)+s\left(Horsepower_i\right)+s\left(Speed_i\right)+\beta_1\cdot Ice\ Class_i+\beta_2\cdot Strengthened_i \quad (4.3)$$

$$g\left(E\left(LT_i\right)\right)=\beta_0+s\left(TC\ Rate_i\right)+s\left(Orderbook\%Group_i\right)+s\left(LIBOR_i\right)+s\left(Dwt_i\right)+s\left(Horsepower_i\right)+s\left(Speed_i\right)+\beta_1\cdot Ice\ Class_i+\beta_2\cdot Strengthened_i \quad (4.4)$$

Model	(4.3)			(4.4)		
Parametric Terms	Estimate	P-value	Significance	Estimate	P-value	Significance
Intercept	4.49647	< 2E-16	***	4.94817	< 2E-16	***
Ice Class_Yes	0.20745	1.53E-15	***	0.21349	3.65E-16	***
Strengthened_Yes	-0.03478	5.96E-03	**	-0.03223	0.0108	*
Smooth Terms	EDF	P-value	Significance	EDF	P-value	Significance
s(TC Rate)	8.707	1.14E-04	***	8.818	9.78E-06	***
s(Orderbook%Total)	6.045	6.15E-11	***	-	-	-
s(Orderbook%Group)	-	-	-	7.801	3.11E-05	***
s(LIBOR)	8.329	< 2E-16	***	8.625	< 2E-16	***
s(Dwt)	8.715	< 2E-16	***	8.710	< 2E-16	***
s(Horsepower)	8.399	2.46E-12	***	8.367	1.15E-11	***
s(Speed)	4.782	4.06E-03	**	4.673	0.0058	**
R-sq.(adj)		0.325			0.319	
GCV		0.11711			0.11793	
N		3801			3801	
AIC		39842.04			39868.95	

Signifi. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 4.7: Statistic Results of Market-vessel Models

Although the statistic results are indeed improved after including the vessel-related variables, the 0.3 R^2 is still not good enough. Besides, by observing the plots of smooth terms, it is found that the curve of Horsepower for both models changes a lot compared to that in the single test (see Appendix A), indicating there is a multicollinearity issue. Hence, only one of Dwt and Horsepower could be remained, and Dwt is eventually chosen for the further research. Again, the differences between the models with Orderbook%Total or Orderbook%Group are still not obvious.

- **Market-vessel-shipyard Model**

For this section, the variables related to shipyards are also involved. Hence, there are two models in total, expressed as follows, with the results shown in Table 4.8.

$$\begin{aligned} g\left(E\left(LT_i\right)\right) = {} & \beta_0 + s\left(TC\ Rate_i\right) + s\left(Orderbook\%Total_i\right) + s\left(LIBOR_i\right) + s\left(Dwt_i\right) \\ & + s\left(Speed_i\right) + s\left(Total\ Area_i\right) + s\left(Max\ Dwt\ Output_i\right) + s\left(No.Building\ Facilities_i\right) \\ & + s(Utilization\ Rate_i) + \beta_1 \cdot Ice\_Class_i + \beta_2 \cdot Strengthened_i + \beta_3 \cdot Builder\ Country_i \end{aligned} \quad (4.5)$$

$$\begin{aligned} g\left(E\left(LT_i\right)\right) = {} & \beta_0 + s\left(TC\ Rate_i\right) + s\left(Orderbook\%Group_i\right) + s\left(LIBOR_i\right) + s\left(Dwt_i\right) \\ & + s\left(Speed_i\right) + s\left(Total\ Area_i\right) + s\left(Max\ Dwt\ Output_i\right) + s\left(No.Building\ Facilities_i\right) \\ & + s(Utilization\ Rate_i) + \beta_1 \cdot Ice\_Class_i + \beta_2 \cdot Strengthened_i + \beta_3 \cdot Builder\ Country_i \end{aligned} \quad (4.6)$$

Model	(4.5)			(4.6)		
Parametric Terms	Estimate	P-value	Significance	Estimate	P-value	Significance
Intercept	4.91236	< 2E-16	***	4.90785	< 2E-16	***
Ice Class_Yes	0.16313	2.19E-09	***	0.17864	8.30E-11	***
Strengthened_Yes	-0.02135	0.1092		-0.02012	0.1330	
Builder Country_Japan	0.08750	1.08E-04	***	0.08582	7.41E-05	***
Builder Country_Philippines	0.15859	0.0140	*	0.24606	1.17E-06	***
Builder Country_South Korea	-0.37634	1.06E-07	***	-0.38992	1.34E-08	***
Smooth Terms	EDF	P-value	Significance	EDF	P-value	Significance
s(TC Rate)	2.097	0.2911		8.683	3.14E-04	***
s(Orderbook%Total)	7.302	1.54E-11	***	-	-	-
s(Orderbook%Group)	-	-	-	7.251	3.80E-04	***
s(LIBOR)	8.717	< 2E-16	***	8.695	< 2E-16	***
s(Dwt)	8.720	< 2E-16	***	8.745	< 2E-16	***
s(Speed)	4.495	3.70E-06	***	4.360	3.27E-06	***
s(Total_Area)	8.993	3.38E-08	***	8.996	4.82E-08	***
s(Max_Dwt_Output)	8.923	5.49E-12	***	7.269	1.64E-10	***
s(No.Building Facilities)	1.863	0.0363	*	1.465	0.3198	
s(Utilization Rate)	8.284	6.01E-08	***	8.501	2.69E-10	***
R-sq.(adj)		0.394			0.391	
GCV		0.10743			0.10835	
N		3516			3516	
AIC		36485.27			36515.32	

Signifi. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 4.8: Statistic Results of Market-vessel-shipyard Models

Up to now, we could find that all the models with Orderbook%Total have better performance that the corresponding models with Orderbook%Group, thus Orderbook%Total should be selected eventually. According to Table 4.8, "TC Rate" is not significant for lead times in this model, so it should not be taken into account in the final model. Besides, we also double the basis dimensions to check if this indicator is large enough and the result is similar to that for newbuilding prices that the doubling basis dimensions could lead to over fitting.

In addition, by observing the plots of smooth terms, we assume that regarding Speed, No.Building Facilities and Utilization Rate as parametric terms would be a better choice, with Total Area dicarded. Finally, the estimation model for lead times of bulk carriers could be expressed by the equation below, with the statistic result shown in Table 4.9, where Speed and No.Building Facilities are excluded as well for low levels of significance.

$$\begin{aligned} g\left(E\left(LT_i\right)\right) = \beta_0 &+ s\left(Orderbook\%Total_i\right) + s\left(LIBOR_i\right) + s\left(Dwt_i\right) \\ &+ s\left(Max\ Dwt\ Output_i\right) + \beta_1 \cdot Utilization\ Rate_i \\ &+ \beta_2 \cdot Ice\_Class_i + \beta_3 \cdot Builder\ Country_i \end{aligned} \tag{4.7}$$

Parametric Terms	Estimate	P-value	Significance
Intercept	4.85491	< 2E-16	***
Ice Class_Yes	0.17291	1.27E-11	***
Utilization Rate	0.00135	5.36E-08	***
Builder Country_Japan	0.04621	2.21E-03	**
Builder Country_Philippines	0.23855	< 2E-16	***
Builder Country_South Korea	-0.34854	1.99E-13	***

Smooth Terms	EDF	P-value	Significance
s(Orderbook%Total)	8.114	< 2E-16	***
s(LIBOR)	8.730	< 2E-16	***
s(Dwt)	8.778	< 2E-16	***
s(Max_Dwt_Output)	8.832	3.34E-14	***
R-sq.(adj)		0.363	
GCV		0.11149	
N		3801	

Signifi. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 4.9: Statistic Results of Estimation Model for Lead Times

4.6 Results Analysis

To make detailed analysis of the model, it is necessary to analyze the separate influences of every single variable on lead times. As for the categorical variables, Table 4.9 shows that if a bulk carrier has Ice Class, it will take more time to delivery it, which seems logical. In addition, it is indicated that, using China as the reference, being built in South

Korea can shorten lead times while in Japan and Philippines it takes more time to deliver a bulk carrier, which means the shipyards in South Korea might be superior to other yards in terms of lead times. In addition, Capacity Utilization Rate is positively related to lead times, suggesting that busier shipyards have to take more time to deliver a new vessel, which is logical that, with too many orders, the order backlog could occur.

When it comes to the smooth terms, Figure 4.5 presents the effects of the smooth terms separately. It can be seen that lead times increase as the increase of orderbook size when the size of orderbook is below 20%, after which the fluctuations of lead times become relatively steady. The relationship between lead times and Dwt encounters a similar changing trend, except the huge gap around Dwt of 150,000 tons which results in a very broad confidence interval. Besides, Max Dwt Output is generally positively related to lead times though it may not be reliable due to the low density of data points. Similarly, LIBOR also presents an overall positive influence on lead times. On the whole, the current model cannot provide a reliable and accurate estimation of lead times, and this bad fitting result is consistent with the low R^2 of 0.363 shown in Table 4.9.

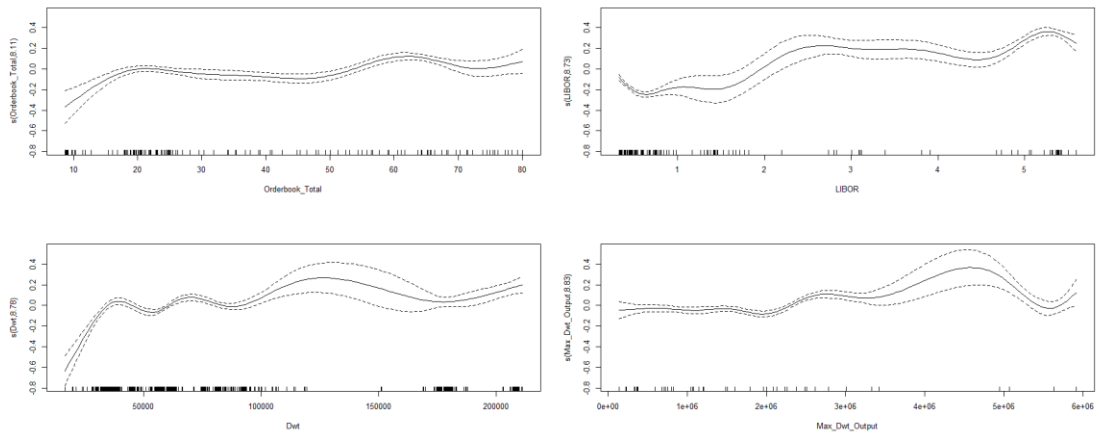

Figure 4.6: Smooth Terms of Estimation Model for Lead Times

The reasons for the bad performance of the estimation model may be multiple and complex, which might be as follows and we need to analyze them one by one.

Firstly, there might be something wrong with respect to the rationality of the model establishment, which could be verified by model checking. In “mgcv”, the model

checking for GAMs can be done by “gam.check”, which produces some diagnostic information about the fitting procedure and results. The default is to produce 4 residual plots, shown in Figure 4.7.

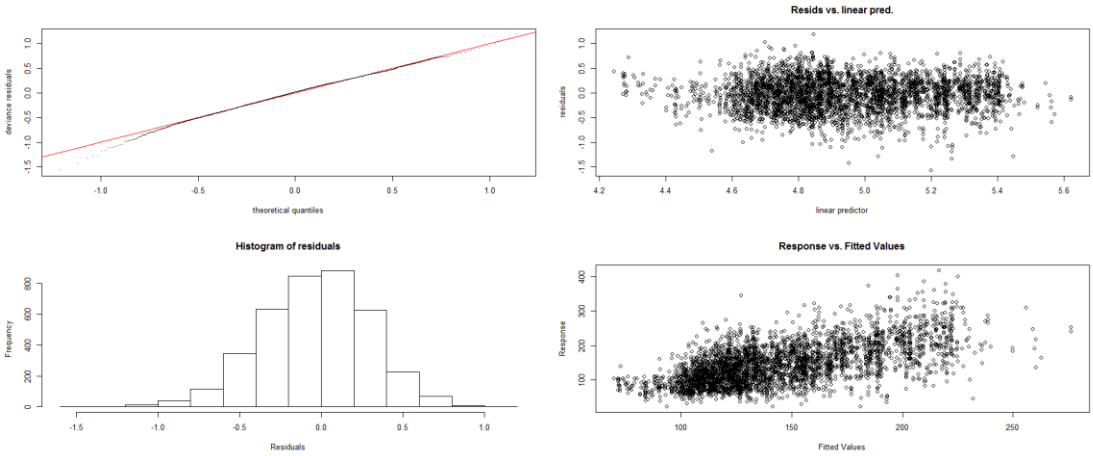

Figure 4.7: Model Checking Results

The upper left QQ (quantile-quantile) plot is very close to a straight line, suggesting that the distributional assumption is reasonable. The upper right panel shows that the variance is approximately constant as the mean increases, suggesting that the constant variance assumption is tenable. The histogram of residuals at lower left seems approximately consistent to a normal distribution. The lower right plot, of response against fitted values, shows a positive linear relationship with a good deal of scatters, except some outliers. To sum up, the model checking is not problematic.

Secondly, the bad fitting result is due to the discontinuity of some variables of the dataset. It can be seen that for “Dwt” there are no values around 150,000 tons and only a few shipyard have a “Max Dwt Output” of more than 4,000,000 dwt. To verify this point, the dataset is processed again by deleting all the fixtures with a “Dwt” of more than 100,000 tons or a “Max Dwt Output” of more than 3,500,000 dwt. With the same modelling method from start to end, a new model using the modified dataset is established, with the statistic results shown in Table 4.9.

Parametric Terms	Estimate	P-value	Significance	**Smooth Terms**	EDF	P-value	Significance
Intercept	4.85491	< 2E-16	***	s(Orderbook%Total)	8.120	< 2E-16	***
Ice Class_Yes	0.17291	9.98E-09	***	s(LIBOR)	8.279	< 2E-16	***
Utilization Rate	0.00135	3.27E-06	***	s(Dwt)	8.799	< 2E-16	***
				s(Max_Dwt_Output)	8.375	< 2E-16	***
				R-sq.(adj)		0.346	
				GCV		0.11439	
				N		3146	

Signifi. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 4.10: Statistic Results of Model with Modified Dataset

From Table 4.10, no improvements can be seen that the adjusted R-square almost remains unchanged. Therefore, the value discontinuity of some variables is not the primary cause of the bad model performance.

Thirdly, as mentioned before, the data related to shipyards are not very reliable. The maximum DWT output and the capacity utilization rate only roughly represent a shipyard's building capacity and occupancy. For this point, with a same modelling method from beginning to end, a model is set up for a single shipyard, in which case the influences of shipyard variables are eliminated. Oshima Shipbuilding, as the shipyard produces most bulk carriers of all the researched shipyards, is selected. The statistic results are shown in Table 4.11.

Parametric Terms	Estimate	P-value	Significance	**Smooth Terms**	EDF	P-value	Significance
Intercept	5.01556	< 2E-16	***	s(Orderbook%Total)	7.441	4.85E-07	***
Ice Class_Yes	0.17844	4.80E-03	**	s(LIBOR)	1.000	1.55E-11	***
Utilization Rate	0.00068	0.8697		s(Dwt)	7.511	< 2E-16	*
				R-sq.(adj)		0.439	
				GCV		0.07937	
				N		229	

Signifi. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 4.11: Statistic Results of Model for Oshima Shipbuilding

Table 4.11 shows that the fitting quality improves with respect to the adjusted R-square, from 0.368 to 0.439. Although it's still not good enough for a reliable model, with half of the observed variation explained by random items, it might be concluded that the unreliable data of shipyards does restrict the fitting quality of the model for lead times, though it is not the primary cause.

In addition, I used Dwt as the proxy for vessel size, while Compensated Gross Tonnage (CGT) might be a better choice as an indicator of the amount of work needed to build

a given vessel since CGT takes the type and size of a particular vessel into account. However, the replacement of Dwt by CGT makes no difference after attempts and this might be because the researched vessels in this thesis are strictly restricted to bulk carriers, whose design requirements, internal structure and required level of details are quite close regardless of the size.

Finally and most importantly, the absence of some crucial variables may also result in the unsatisfactory model performance. For instance, according to [42], the shipyard building capacity is mainly in terms of erection area, which we cannot get access to and use unreliable total area instead. Besides, the time needed to get critical inputs, like main engines, substantially influences building times, thereby affecting lead times greatly. However, it is not possible to get this data in this research. And as discussed in Section 4.2, in most industries, lead times are mainly affected by scheduling, batching and management levels. Unfortunately, these factors are generally very hard to be quantified, let alone that, in this research, there is none access to these data without the detailed on-the-spot investigations.

4.7 Summary

In this chapter, we try to model the estimation of lead times for bulk carriers by using the GAM method, given the corresponding dataset. Although the variable identification, data pre-processing and model building method all appear not problematic, the fitting quality of the final model is not satisfactory at all. The reasons of the unreliable estimation model are multiple, including unreliable data and the lack of some crucial variables.

Nevertheless, even with the bad fitting result, we could still find some useful points about lead times of bulk carriers. First, the shipyards in China and South Korea could be able to deliver a standard bulk carrier sooner than those in Japan and Philippines, which might result from the tremendous manpower of China’s shipyards and the advanced production systems of shipyards in South Korea, but could not be verified in

this thesis. As for the hull of bulk carriers, it turns out that whether being strengthened for heavy cargo seems not significant to bulk carrier's lead times while Ice Class might have a positive influence. Dwt as the most important characteristic parameter of bulk carriers is influential that larger bulk carriers generally require more time to deliver, especially for Handysize and Handymax. From the perspective of shipyards, no obvious relationship between lead times and Total Area or Max Dwt Output could be found whereas we could assume that shipyards with higher capacity utilization need more time to deliver their ordered vessels. When it comes to the external factors from the shipping or shipbuilding market, only the size of orderbook appear to relate to lead times that general lead times arise with the increase of orderbook size at first and then plateaus after the demand for vessels is at relatively high levels. The above points might not be exactly accurate but could provide some insights into lead times of bulk carriers.

Furthermore, as far as I am concerned, in contrast with the estimation of newbuilding prices, establishing a quantitative model for lead times in the shipbuilding industry is far more difficult, since there are so many factors influencing lead times and a considerable part of them are hard to quantify. The lack of references related to shipbuilding lead times also makes it very difficult to theoretically summarize a list of variables that are influential to lead times. Therefore, in order to obtain accurate estimation of lead times, it is necessary to corporate with certain shipyards to carry out the detailed investigations on the spot. Moreover, turning to the practitioners and specialists in other fields who also study lead times might be of great help.

5 Conclusions

5.1 Answers to Research Questions

5.1.1 Sub Questions

1. *What is the appropriate method to estimate bulk carrier's newbuilding prices and lead times? And why?*

Regression analysis, as a mature statistical method, can estimate the relationships among variables with necessary data. Given the fact that the data related to the shipping and shipbuilding market has a great variety of types with different units and magnitudes, GAM is selected as the specific method for its great flexibility and ability to estimate both linear and nonlinear relationships.

2. *Where can we obtain the necessary data for research and how should we process the data?*

Clarksons is the main data source for this thesis, where we can find detailed data of registered bulk carriers and the latest information on the shipping market. Besides, there are other sources such as World Bank Open Data and shipyards' websites. The specific choice of the data for each variable should be based on the actual conditions and the collected dataset needs to be pre-processed at first to avoid the bad influences of outliers.

3. *What variables are influential to bulk carrier's newbuilding prices and how to establish a reasonable estimation model for newbuilding prices?*

According to the previous researches, the factors that are influential to bulk carrier's newbuilding prices can be theoretically summaries as shipbuilding costs, shipyard capacity, vessel orderbook, freight rates, second-hand prices, exchange rates and inflation. After a series of modelling process including single tests, multicollinearity tests and model selection, the final estimation model for newbuilding prices contains these specific variables: vessel size (Dwt), orderbook size (percentage of corresponding size group fleet), second-hand prices, number of hatches, GDP per capita (proxy for wages), exchange rates and LIBOR. Furthermore, both the statistic results and plots

indicate a satisfactory fitting performance of this model.

4. *What variables are influential to bulk carrier's lead times and how to establish a reasonable estimation model for lead times?*

Unlike newbuilding prices, there are not that many literatures to refer to when estimating bulk carrier's lead times. Based on a few references and some reasonable assumptions, the variables having effects on lead times can be preliminarily summarized as shipyard information (including total area, number of building facilities, number of employees, maximum annual Dwt output, capacity utilization rate and builder country), vessel specifications (including Dwt, horsepower, speed, Ice Class and etc.) and market-related proxies (including time charter rate, vessel orderbook and LIBOR). After the modelling processes similar to those for newbuilding prices, the final estimation model for lead times consists of these specific variables: orderbook size (percentage of total bulk carrier fleet), LIBOR, vessel size (Dwt), Ice Class, maximum annual Dwt output, utilization rate, country of shipyards. Nevertheless, this model cannot provide a good fitting result and we assume that the unreliability of data and the lack of crucial variables are probably the main causes.

5.1.2 Main Question

Combining the discussions above, the main research question can be answered.

Is it possible to estimate a bulk carrier's newbuilding price and its lead time, using open access data and empirical information?

Given necessary data and information, it is possible to accurately estimate a bulk carrier's newbuilding price with the generalized additive model (GAM) as the specific modelling method through a series of processes including variables identification, data collection and analysis, model specification and related analysis. However, because of the lack of crucial influential variables and access to necessary data, the lead time cannot be estimated reliably.

5.2 General Remarks

As a general conclusion, it could be stated that, given necessary data and information, GAM is capable of estimating the linear and nonlinear relationships between the dependent variable and various independent variables. Thus, when having access to the data related to the shipping and shipbuilding market, using GAM to estimate bulk carrier's newbuilding prices and lead times is reasonable and practical.

Furthermore, it could be concluded that the newbuilding prices of bulk carriers are mainly affected by vessel size (Dwt), shipping market status (second-hand prices) and the demand for new vessels (orderbook size as the percentage of corresponding size group fleet), all of which have a strong positive influences on the prices, with the vessel size as the most influential one. This also suggests that the newbuilding prices are cost-driven, which conforms to the economic theory. In addition, the positive effects of number of hatches, GDP per capita and LIBOR, together with the negative effect of exchange rates, are also in line with the theoretical expectations. Besides, it is demonstrated that the scale of economy is also applicable to the newbuilding market since the price per Dwt decreases with vessel sizes.

Another conclusion, resulting from the estimation of lead times of bulk carriers, is that lead times of bulk carriers appear to be mainly positively influenced by vessel size (Dwt) and the demand for new vessels (orderbook size as the percentage of total bulk carrier fleet). Similarly, the maximum Dwt output and capacity utilization rate of shipyards, LIBOR and having Ice Class also indicate positive influences. Moreover, with respect to lead times, China and South Korea show superiority to Japan and Philippines. However, with a low R^2 nor more than 0.4, the fitting result is not very reliable. To find the causes, the model is checked and two new models are constructed: one with the modified dataset where some extreme values are discarded, and the other one with the data of single shipyard, Oshima Shipbuilding. Based on these, the unreliability of necessary data is a reason and further the lack of some most influential factors is presumably the major cause.

5.3 Recommendations

During the whole research, the collection, flittering and analysis of necessary data is a difficult and time-consuming process, and there is still room for improvement about these, especially for collecting data. As mentioned, the main data source of this thesis is Clarksons and maybe it is more reasonable to consult other maritime databases such as Lloyd's List and Maritime Traffic Data, in which way we can make comparisons and obtain more reliable useful data. Besides, the data of shipyards only from the shipyard profiles and websites is limited and non-specific, which might be improved by contacting the researched shipyards and obtain more details.

Another point that can be improved is about the choice of basis dimensions when constructing GAMs. In this thesis, all the basis dimensions default to 10 and the results of doubling them are checked. Finally, all the basis dimensions are set to remain the default value. However, if possible, it is more reasonable to find the optimal basis dimension for each smooth term based on the trade-off between smoothness and accuracy.

Third, when faced with the multicollinearity issues, the solution in this thesis is to keep the most appropriate variable and discard others. This approach is easy to apply and somehow plausible, but it is likely to miss some information by simply dropping variables. To improve this point, by using "mgcv" in the R language, the function *ti*() could be introduced, which is used to specify tensor product interactions between smooth terms, reflecting the main and interaction effects.

At last, if we want to reliably and accurately estimate the lead times of bulk carriers, there is a lot else to do. On the one hand, we can contact shipyards and make detailed investigations on the spot there, with recording all types of necessary data and information, which is certainly more reliable than what we can get though the databases or the internet. On the other hand, there is little previous work about lead times of vessels, and at this time, the cooperation with the practitioners from other industries might provide valuable inspirations and suggestions.

Reference

1. Nations, U. *Maritime Transport Is 'Backbone of Global Trade and the Global Economy', Says Secretary-General in Message for International Day*. 2016; Available from: https://www.un.org/press/en/2016/sgsm18129.doc.htm.
2. Stopford, M., *Maritime economics 3e*. 2009: Routledge.
3. UNCTAD, *The Review of Maritime Transport 2018 (UNCTAD/RMT/2018)*. 2018: United Nations Publication.
4. Breiman, L. and J.H. Friedman, *Estimating optimal transformations for multiple regression and correlation.* Journal of the American statistical Association, 1985. **80**(391): p. 580-598.
5. Nelder, J.A. and R.W. Wedderburn, *Generalized linear models.* Journal of the Royal Statistical Society: Series A (General), 1972. **135**(3): p. 370-384.
6. Hastie, T. and R. Tibshirani, *Generalized additive models: some applications.* Journal of the American Statistical Association, 1987. **82**(398): p. 371-386.
7. Jones, K. and S. Almond, *Moving out of the linear rut: the possibilities of generalized additive models.* Transactions of the Institute of British Geographers, 1992: p. 434-447.
8. Chambers, M. and T.W. Dinsmore, *Advanced analytics methodologies: driving business value with analytics*. 2014: Pearson Education.
9. Wood, S.N., *Generalized additive models: an introduction with R*. 2017: Chapman and Hall/CRC.
10. Faraway, J.J., *Linear models with R*. 2016: Chapman and Hall/CRC.
11. Duchon, J., *Splines minimizing rotation-invariant semi-norms in Sobolev spaces*, in *Constructive theory of functions of several variables*. 1977, Springer. p. 85-100.
12. Heinze, G., C. Wallisch, and D. Dunkler, *Variable selection–a review and recommendations for the practicing statistician.* Biometrical Journal, 2018. **60**(3): p. 431-449.
13. Tsolakis, S., *Econometric Analysis of Bulk Shipping: implications for investment strategies and financial decision-making*. 2005.
14. Tsolakis, S., C. Cridland, and H. Haralambides, *Econometric modelling of second-hand ship prices.* Maritime Economics & Logistics, 2003. **5**(4): p. 347-377.
15. Hawdon, D., *Tanker freight rates in the short and long run.* Applied Economics, 1978. **10**(3): p. 203-218.
16. Beenstock, M., *A theory of ship prices.* Maritime Policy and Management, 1985. **12**(3): p. 215-225.
17. Beenstock, M. and A. Vergottis, *An econometric model of the world market for dry cargo freight and shipping.* Applied Economics, 1989. **21**(3): p. 339-356.
18. Haralambides, H., S. Tsolakis, and C. Cridland, *Econometric modelling of newbuilding and secondhand ship prices.* Research in Transportation Economics, 2005. **12**: p. 65-105.
19. Strandenes, S., *Norship: a simulation model of markets in bulk shipping.*

Discussion Paper 11, Norwegian School of Economics and Business Administration, Bergen, Norway, 1986.
20. Jin, D., *Supply and demand of new oil tankers.* Maritime Policy and Management, 1993. **20**(3): p. 215-227.
21. Volk, B., *The shipbuilding cycle-a phenomenon explained?* 1994: Institute of Shipping Economics and Logistics.
22. Adland, R., H. Jia, and S. Strandenes, *Asset bubbles in shipping? An analysis of recent history in the drybulk market.* Maritime Economics & Logistics, 2006. **8**(3): p. 223-233.
23. Shetelig, H., *Shipbuilding Cost Estimation: Parametric Approach*. 2013, Institutt for marin teknikk.
24. Strandenes, S.-P., *Economics of the markets for ships.* The handbook of maritime economics and business, 2002: p. 186-202.
25. Pruyn, J.F.J., *Shipping and shipbuilding scenario evaluations through integration of maritime and macroeconomic models.* 2013.
26. Porter, M.E., *How competitive forces shape strategy*, in *Readings in strategic management*. 1989, Springer. p. 133-143.
27. Adland, R. and H. Jia, *Shipping market integration: The case of sticky newbuilding prices.* Maritime Economics & Logistics, 2015. **17**(4): p. 389-398.
28. Bertram, V., *Strategic control of productivity and other competitiveness parameters.* Proceedings of the Institution of Mechanical Engineers, Part M: Journal of Engineering for the Maritime Environment, 2003. **217**(2): p. 61-70.
29. Adland, R., K. Norland, and E. Sætrevik, *The impact of shipyard and shipowner heterogeneity on contracting prices in the newbuilding market.* Maritime Business Review, 2017. **2**(2): p. 58-78.
30. Koehn, S., *Generalized additive models in the context of shipping economics*. 2008, University of Leicester.
31. Gosling, J., et al., *Principles for the design and operation of engineer-to-order supply chains in the construction sector.* Production Planning & Control, 2015. **26**(3): p. 203-218.
32. Mourtzis, D., et al., *Knowledge-based estimation of manufacturing lead time for complex engineered-to-order products.* Procedia CIRP, 2014. **17**: p. 499-504.
33. Nyhuis, P., et al., *Applying simulation and analytical models for logistic performance prediction.* CIRP annals, 2005. **54**(1): p. 417-422.
34. Solomon, H., K. Jilcha, and E. Berhan. *Lead time prediction using simulation in leather shoe manufacturing*. in *Afro-European Conference for Industrial Advancement*. 2015. Springer.
35. Chryssolouris, G., et al., *Digital manufacturing: history, perspectives, and outlook.* Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture, 2009. **223**(5): p. 451-462.
36. Kawasaki, T., et al., *Multi criteria simulation model for lead times, costs and CO2 emissions in a low-carbon supply chain network.* Procedia Cirp, 2015.

26: p. 329-334.
37. Parlar, M., *Continuous-review inventory problem with random supply interruptions.* European Journal of Operational Research, 1997. **99**(2): p. 366-385.
38. Seyedhosseini, S.M. and A. Ebrahimi-Taleghani, *A stochastic analysis approach on the cost-time profile for selecting the best future state map.* South African Journal of Industrial Engineering, 2015. **26**(1): p. 267-291.
39. Pfeiffer, A., et al., *Manufacturing lead time estimation with the combination of simulation and statistical learning methods.* Procedia CIRP, 2016. **41**: p. 75-80.
40. Öztürk, A., S. Kayalıgil, and N.E. Özdemirel, *Manufacturing lead time estimation using data mining.* European Journal of Operational Research, 2006. **173**(2): p. 683-700.
41. Okubo, H., et al., *Production lead-time estimation system based on neural network.* Proceedings of Asia-Pacific Region of Decision Sciences Institute, 2000.
42. Pires Jr, F., T. Lamb, and C. Souza, *Shipbuilding performance benchmarking.* International journal of business performance management, 2009. **11**(3): p. 216-235.
43. OECD, *Shipbuilding Market Developments Q2 2018*. 2018.

Appendix A: Plots to Evaluate Models

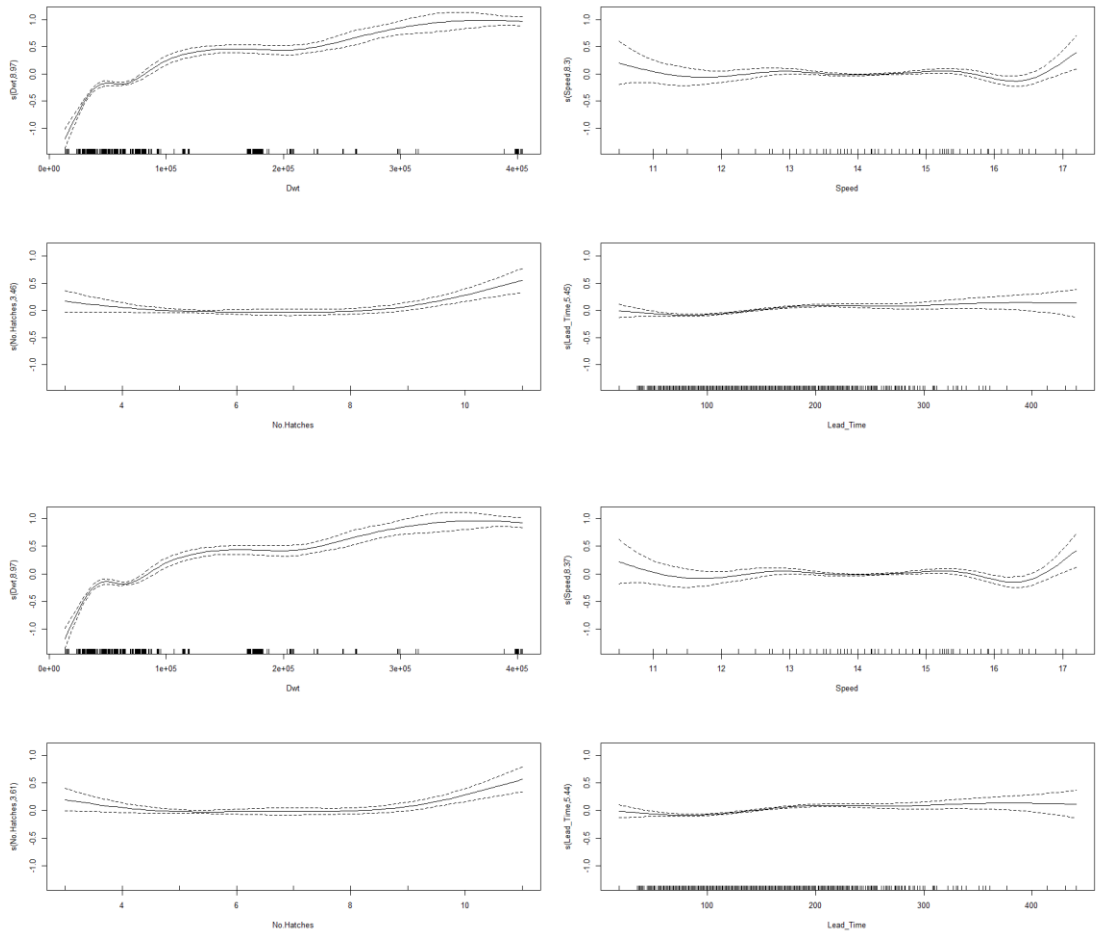

Figure A.1: Plots of Smooth Terms for Model (3.2) and (3.4)

The upper and lower four plots respectively present the plots of smooth terms for Model (3.2) and (3.4), and there is only little difference between them. Thus, which model is better cannot be judged from the plots of smooth terms.

This figure also suggests that Speed, No.Hatches and Lead Time might be regarded as parametric terms, because the plots for these terms are relatively steady and do not show obvious nonlinear relationships.

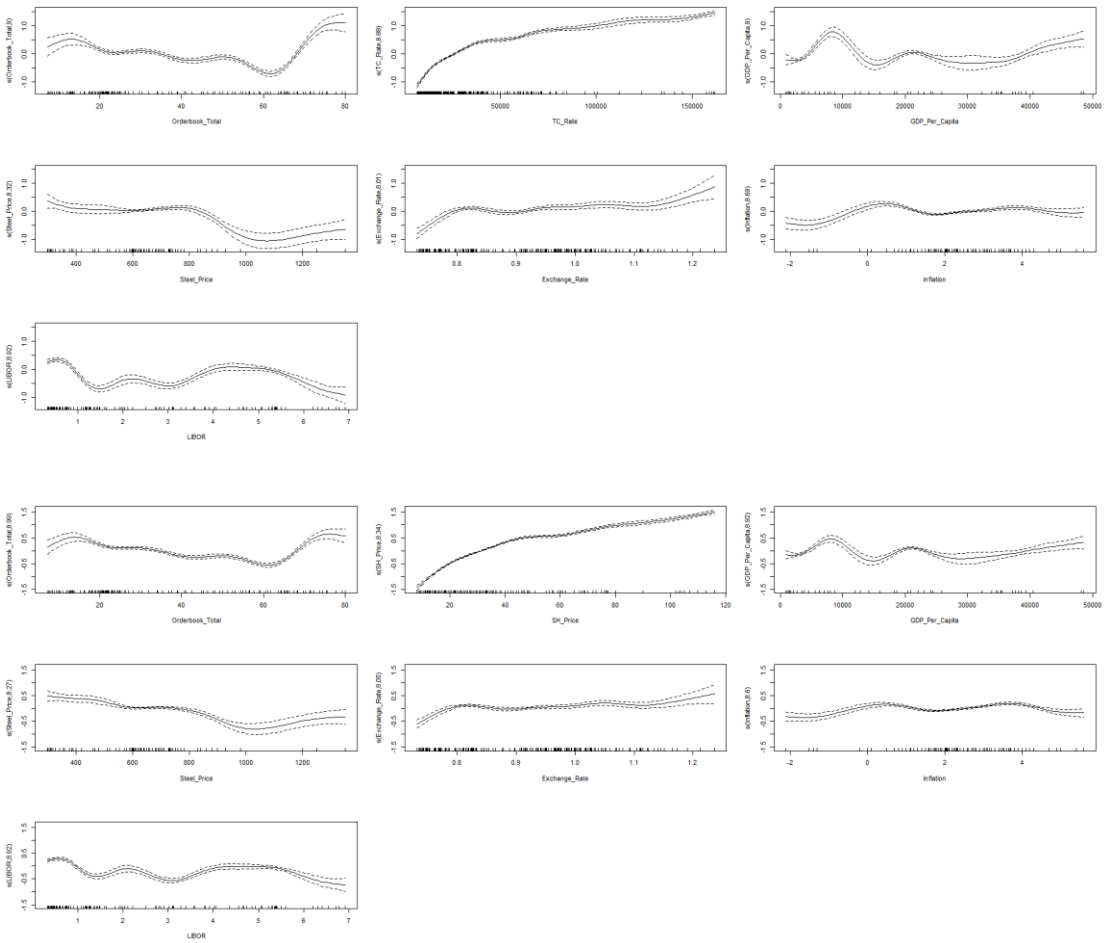

Figure A.2: Plots of Smooth Terms for Model (3.5) and (3.7)

The upper and lower plots respectively present the plots of smooth terms for Model (3.5) and (3.7), where the plots of Orderbook%Total are weird and not explainable compared to what they are in single tests.

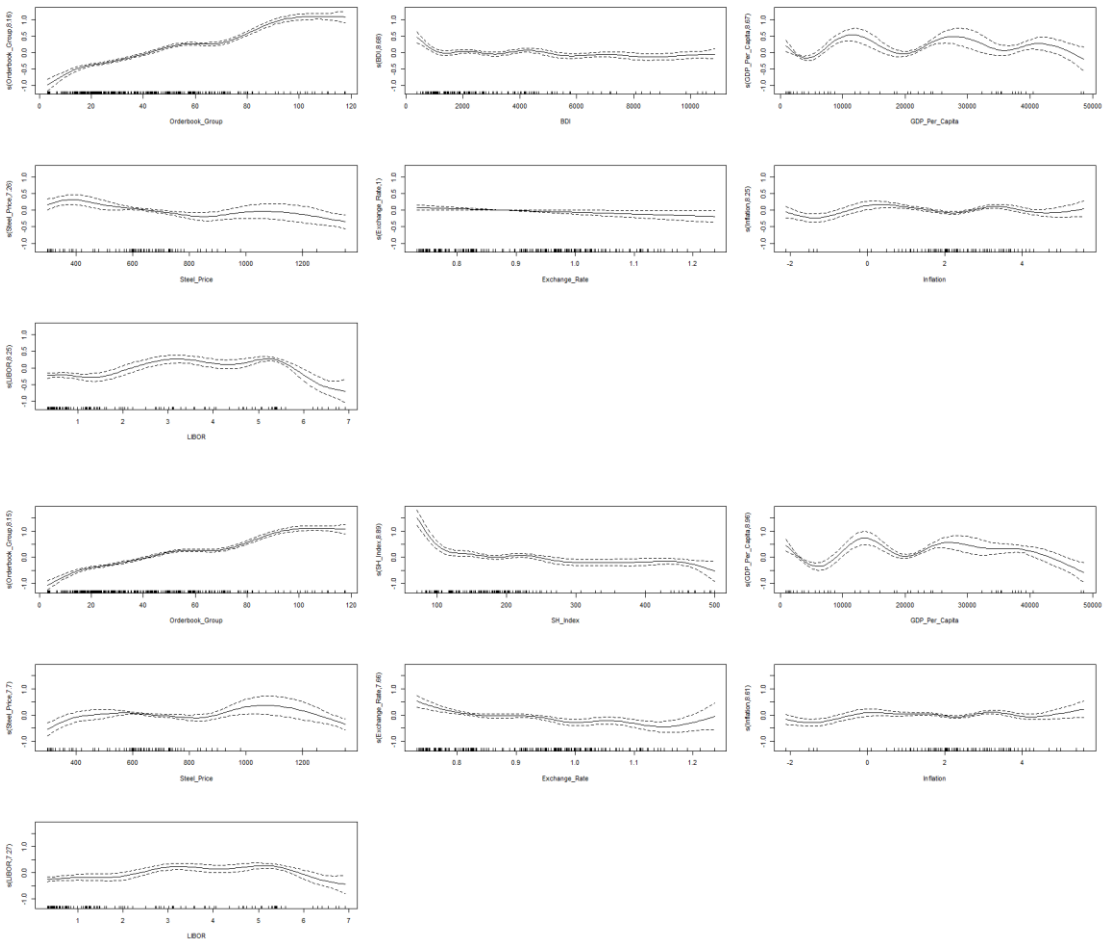

Figure A.3: Plots of Smooth Terms for Model (3.10) and (3.12)

The upper and lower plots respectively present the plots of smooth terms for Model (3.10) and (3.12), where the plots of BDI and SH Index are weird and not explainable compared to what they are in single tests.

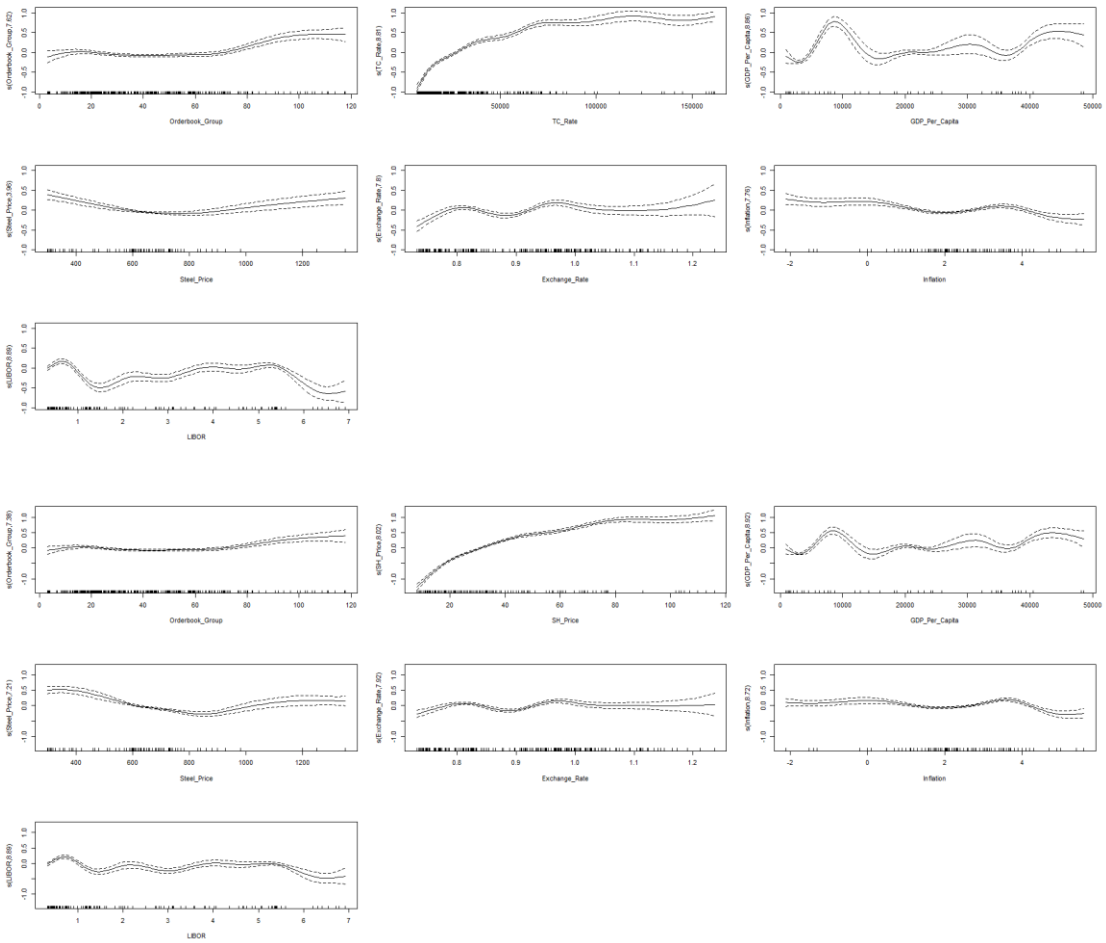

Figure A.3: Plots of Smooth Terms for Model (3.9) and (3.11)

The upper and lower plots respectively present the plots of smooth terms for Model (3.9) and (3.11), where the plots of Steel Price are different from what they are in single tests. Although the difference is little and at an acceptable level, Steel Price is still discarded for more reliable estimations.

Besides, this figure also indicates that GDP Per Capita, Exchange Rate, Inflation and LIBOR might be regarded as parametric terms in the model.

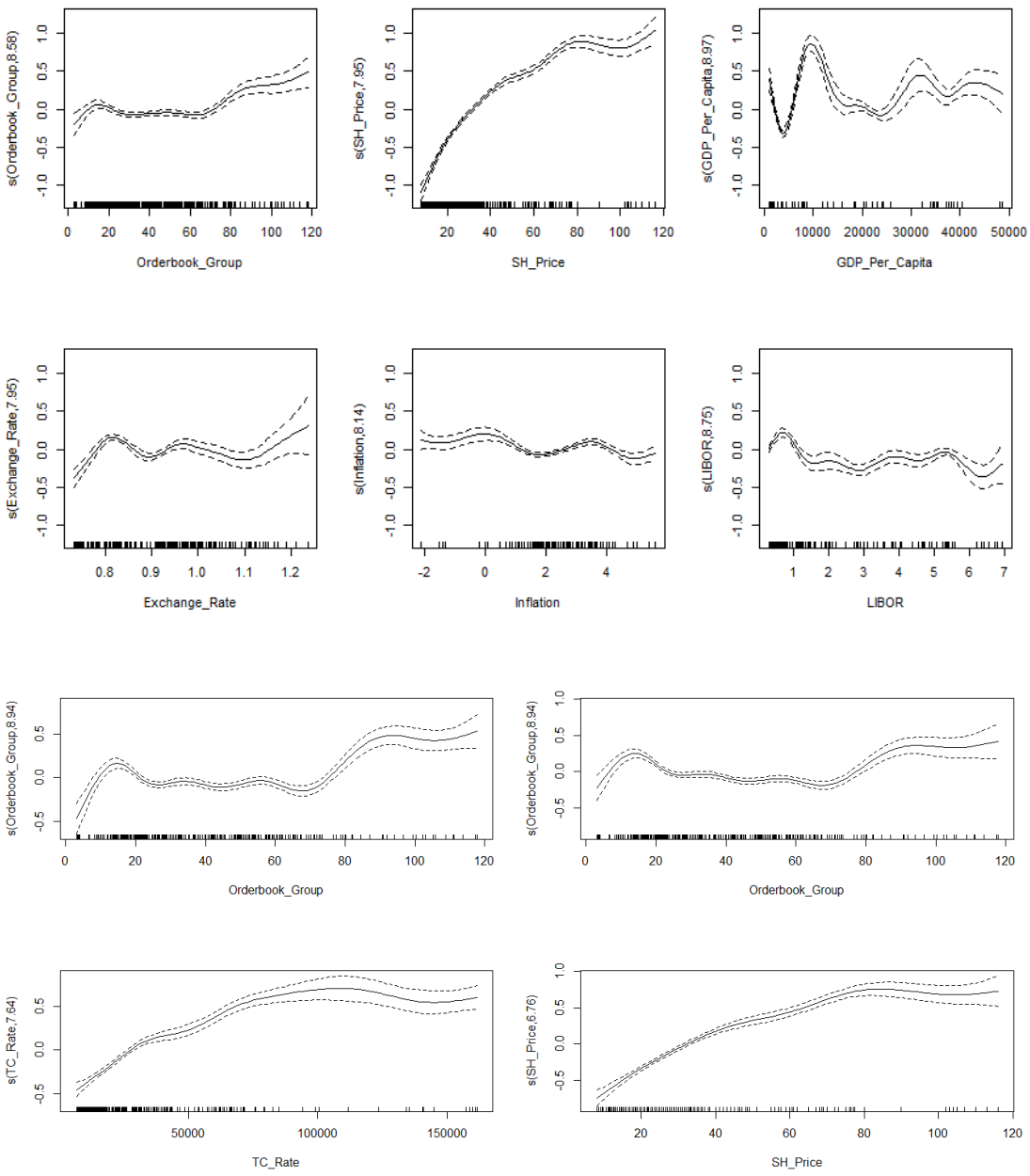

Figure A.4: Plots of Smooth Terms for Model (3.14), (3.15) and (3.16)

The upper, lower left and lower right plots respectively present the plots of smooth terms for Model (3.14), (3.15) and (3.16). From the upper plot, the curves of GDP Per Capita, Exchange Rate, Inflation and LIBOR random fluctuates between a certain range, suggesting that regarding them as parametric terms is a better choice.

The lower two plots show that the model with TC Rate has boarded confidence intervals than those of the model with SH Price. Considering the better statistic results, Model (3.16) is more appropriate than (3.15).

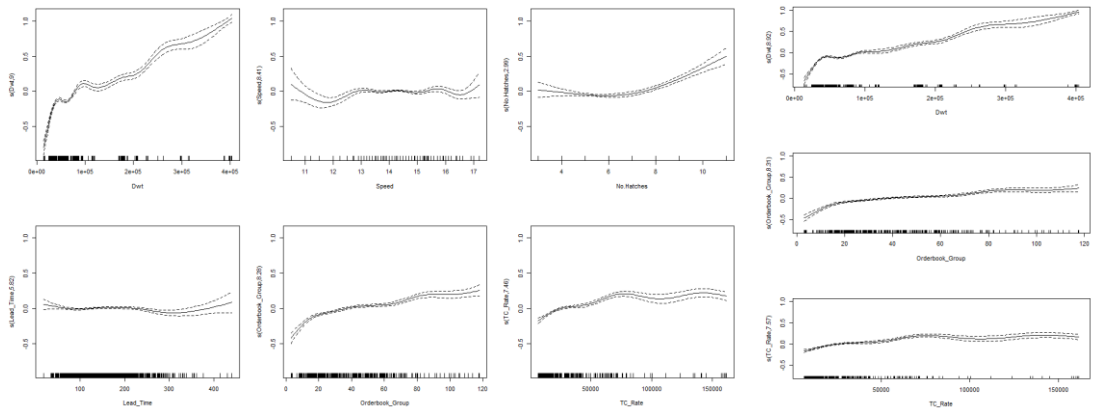

Figure A.5: Plots of Smooth Terms for Model (3.18) and (3.20)

The left and right plots respectively present the plots of smooth terms for Model (3.18) and (3.20). The statistic results of the four models are very close, under which condition Model (3.18) and (3.20) are distinguished since they use fewer independent variables to well estimate the dependent variable. For the same reason, Model (3.20) is better more appropriate than (3.18) for it uses fewer smooth terms, and the plots also suggest that Speed, No.Hatches and Lead Time should be regarded as parametric terms.

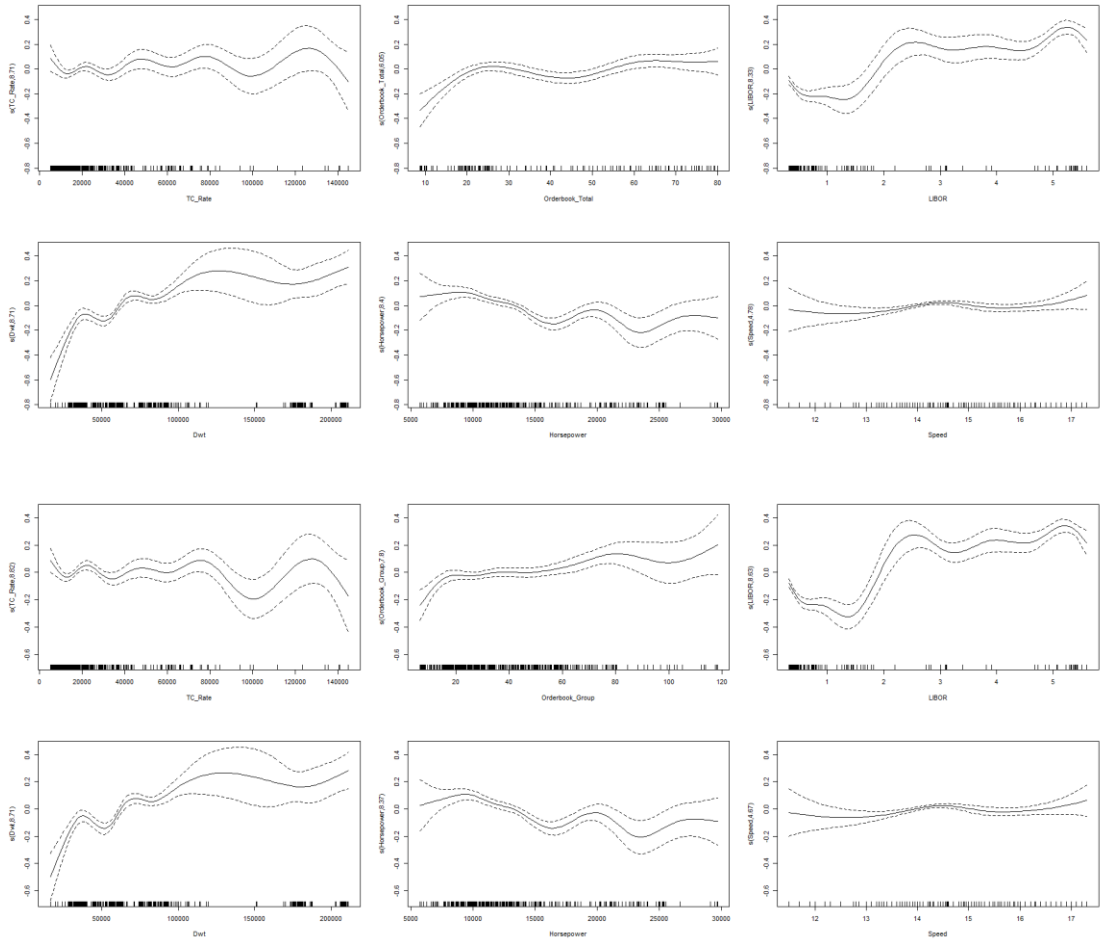

Figure A.6: Plots of Smooth Terms for Model (4.3) and (4.4)

The upper and lower plots respectively present the plots of smooth terms for Model (4.3) and (4.4), where the curves for Horsepower change a lot compared to the single tests, indicating multicollinearity issues.

Appendix B: Models for Newbuilding Prices with Price/Dwt

Variable	Dwt	Grain Capacity	Horsepower	Speed	No.Hatches	Lead Time	Steel Price	TC Rate	BDI
R-sq.(adj)	0.551	0.538	0.309	0.147	0.335	0.0329	0.124	0.0435	0.195
P-value	< 2E-16	< 2E-16	< 2E-16	< 2E-16	< 2E-16	< 2E-16	< 2E-16	< 2E-16	< 2E-16
Significance	***	***	***	***	***	***	***	***	***
Variable	**SH Price**	**SH Index**	**Exchange Rate**	**Inflation**	**LIBOR**	**GDP Per Capita**	**GDP Per Capita Growth**	**Orderbook %Total**	**Orderbook %Group**
R-sq.(adj)	0.0322	0.215	0.156	0.11	0.165	0.166	0.124	0.159	0.0505
P-value	< 2E-16	< 2E-16	< 2E-16	< 2E-16	< 2E-16	< 2E-16	< 2E-16	< 2E-16	< 2E-16
Significance	***	***	***	***	***	***	***	***	***

Table B.1: Statistic Results of Single Tests

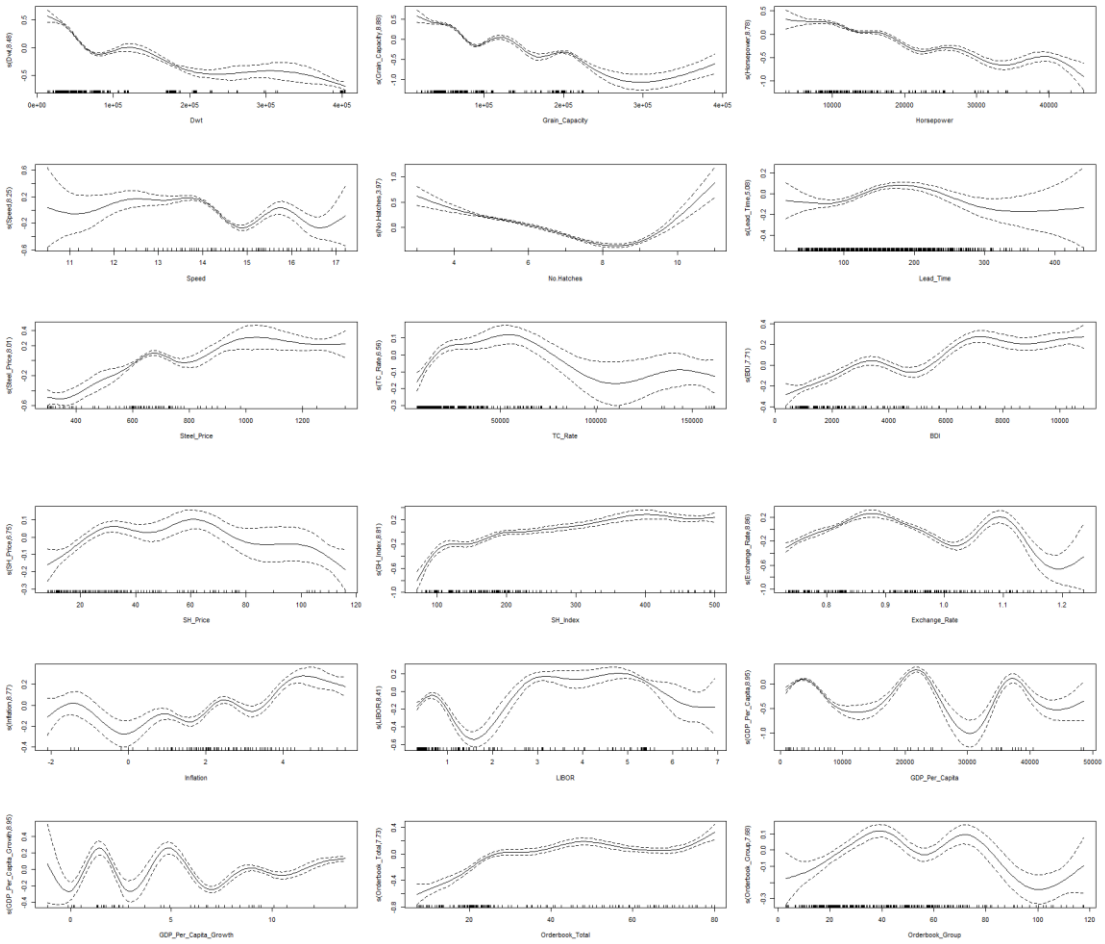

Figure B.1: Smooth Terms of Single Tests

Table B.1 and Figure B.1 respectively present the statistic results and smooth terms of single tests for Price/Dwt, where it can be noticed that Dwt is still the most influential single variable.

Dwt	**Grain Capacity**	**Horsepower**	**No.Hatches**		
	Y	Y	N		
TC Rate	**BDI**	**SH Price**	**SH Index**	**Orderbook%Total**	**Orderbook%Group**
	Y	Y	Y	N	N
BDI	**SH Price**	**SH Index**	**Inflation**	**LIBOR**	
	Y	Y	N	N	
SH Price	**Orderbook%Total**	**Orderbook%Group**	**Steel Price**		
	Y/N	N	Y/N		
SH Index	**Orderbook%Total**	**Orderbook%Group**	**Steel Price**		
	N	N	N		
Steel Price	**Orderbook%Total**	**Orderbook%Group**			
	Y/N	N			

Table B.2: Summary of Multicollinearity Tests

The results of multicollinearity tests are similar: Dwt, Grain Capacity and Horsepower are multicollinear and only Dwt is kept; TC Rate, BDI, SH Price and SH Index are multicollinear and we need to make a choice among them during the model specification.

The model specification also starts with the contract-specific variables. Similarly, there are four models for this part, and the equations are same as those of models with Price. Table B.3 presents their statistic results.

Model	(B.1)			(B.2)			(B.3)			(B.4)		
Parametric Terms	Estimate	P-value	Significance	Estimate	P-value	Significance	Estimate	P-value	Significance	Estimate	P-value	Significance
Intercept	-7.5518	<2E-16	***	-7.5604	<2E-16	***	-7.5973	<2E-16	***	-7.5348	<2E-16	***
Gear	-	-	-	-	-	-	-0.1381	4.83E-03	**	-0.0553	0.051	.
Ice Class_Yes	-	-	-	-	-	-	-0.1039	3.09E-03	**	-0.0900	6.79E-05	***
Strengthened_Yes	-	-	-	-	-	-	3.23E-03	0.850		0.0190	0.114	
Yard Size_Mega	-	-	-	-	-	-	0.0458	0.039	*	-	-	-
Yard Size_Medium	-	-	-	-	-	-	0.0770	5.17E-04	***	-	-	-
Yard Size_Small	-	-	-	-	-	-	9.91E-04	0.993		-	-	-
Yard Size_Very Small	-	-	-	-	-	-	0.0570	0.212		-	-	-
Smooth Terms	EDF	P-value	Significance	EDF	P-value	Significance	EDF	P-value	Significance	EDF	P-value	Significance
s(Dwt)	8.481	<2E-16	***	8.889	<2E-16	***	8.430	<2E-16	***	8.845	<2E-16	***
s(Speed)	-	-	-	8.311	3.60E-03	**	7.034	0.003	**	1.002	0.872	
s(No.Hatches)	-	-	-	3.628	1.05E-06	***	3.799	1.55E-09	***	3.343	4.33E-06	***
s(Lead Time)	-	-	-	5.423	<2E-16	***	6.550	1.66E-11	***	5.419	<2E-16	***
R-sq.(adj)	0.551			0.62			0.634			0.62		
GCV	0.05920			0.05297			0.05709			0.05296		
N	1625			1622			928			1622		

Signifi. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table B.3: Statistic Results of Models with Contract-specific Variables.

Comparing Model (B.3) and (B.4), the loss of Yard Size does not give a statistically significant deterioration of the fitting result and the significance level of Yard Size in (B.3) also suggests Yard Size does not have to be considered. Comparing Model (B.2) and (B.4), the inclusion of Gear, Ice Class and Strengthened does not show any statistical improvements, under which condition these variables should not be included in the model. Besides, the better statistic results of Model (B.2) than that of (B.1)

suggest that (B.2) is the most appropriate model in this part. In addition, Speed, No.Hatches and Lead Time should be also considered as parametric terms.

For the models with macro variables, there are also eight different models to compare, whose statistic results are presented in Table B.4.

Model	(B.5)			(B.6)			(B.7)			(B.8)		
Parametric Terms	Estimate	P-value	Significance	Estimate	P-value	Significance	Estimate	P-value	Significance	Estimate	P-value	Significance
Intercept	-7.5653	<2E-16	***	-7.5419	<2E-16	**	-7.5708	<2E-16	***	-7.5419	<2E-16	***
Smooth Terms	EDF	P-value	Significance	EDF	P-value	Significance	EDF	P-value	Significance	EDF	P-value	Significance
s(Orderbook%Total)	8.782	<2E-16	***	4.509	5.07E-03	***	8.606	<2E-16	***	8.332	3.33E-04	***
s(Orderbook%Group)	-	-	-	-	-	-	-	-	-	-	-	-
s(TC Rate)	8.924	<2E-16	***	-	-	-	-	-	-	-	-	-
s(BDI)	-	-	-	7.322	2.58E-06	***	-	-	-	-	-	-
s(SH Price)	-	-	-	-	-	-	8.905	<2E-16	***	-	-	-
s(SH Index)	-	-	-	-	-	-	-	-	-	8.929	6.34E-06	***
s(GDP Per Capita)	8.998	<2E-16	***	8.900	<2E-16	***	8.998	<2E-16	***	8.831	5.19E-15	***
s(Steel Price)	8.364	8.03E-14	***	8.721	5.31E-11	***	8.045	<2E-16	***	8.044	1.24E-03	***
s(Exchange Rate)	8.212	<2E-16	***	7.885	2.02E-07	***	8.334	<2E-16	***	8.523	5.67E-09	***
s(Inflation)	8.038	1.95E-04	***	8.546	4.20E-05	***	7.341	3.08E-07	***	8.695	4.30E-05	***
s(LIBOR)	8.796	<2E-16	***	8.341	2.03E-12	***	8.648	<2E-16	***	7.993	6.30E-07	***
R-sq.(adj)		0.715			0.331			0.772			0.325	
GCV		0.04126			0.09119			0.02948			0.09177	
N		1625			1625			1625			1625	
Model	(B.9)			(B.10)			(B.11)			(B.12)		
Parametric Terms	Estimate	P-value	Significance	Estimate	P-value	Significance	Estimate	P-value	Significance	Estimate	P-value	Significance
Intercept	-7.5606	<2E-16	***	-7.5560	<2E-16	***	-7.5663	<2E-16	***	-7.5570	<2E-16	***
Smooth Terms	EDF	P-value	Significance	EDF	P-value	Significance	EDF	P-value	Significance	EDF	P-value	Significance
s(Orderbook%Total)	-	-	-	-	-	-	-	-	-	-	-	-
s(Orderbook%Group)	8.180	3.77E-05	***	8.288	<2E-16	***	8.236	1.13E-06	***	8.107	<2E-16	***
s(TC Rate)	8.955	<2E-16	***	-	-	-	-	-	-	-	-	-
s(BDI)	-	-	-	8.522	<2E-16	***	-	-	-	-	-	-
s(SH Price)	-	-	-	-	-	-	8.917	<2E-16	***	-	-	-
s(SH Index)	-	-	-	-	-	-	-	-	-	8.954	<2E-16	***
s(GDP Per Capita)	8.990	<2E-16	***	8.760	<2E-16	***	8.995	<2E-16	***	8.764	<2E-16	***
s(Steel Price)	8.132	<2E-16	***	7.860	<2E-16	***	7.539	<2E-16	***	8.272	2.35E-16	***
s(Exchange Rate)	8.756	<2E-16	***	7.793	4.35E-10	***	8.843	<2E-16	***	6.664	3.20E-11	***
s(Inflation)	7.846	7.20E-09	***	8.593	3.78E-13	***	8.161	2.63E-16	***	8.743	2.62E-12	***
s(LIBOR)	8.134	6.95E-16	***	6.894	<2E-16	***	8.128	<2E-16	***	8.675	<2E-16	***
R-sq.(adj)		0.634			0.581			0.694			0.588	
GCV		0.05130			0.06112			0.03914			0.05906	
N		1625			1625			1625			1625	

Signifi. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table B.4: Statistic Results of Models with Macro Variables

After checking the plots of smooth terms, there are multicollinearity issues for all the models with Orderbook%Group, TC Rate, SH Price, indicating these models cannot be selected. Steel Price again indicates a potential multicollinearity issue, thus being discarded. Besides, the model with BDI performs better than the model with SH Index in both statistic results and plots of smooth terms.

After integration, it is indicated that Speed, Lead Time and Inflation are not significant while No.Hatches, GDP Per Capita, Exchange Rate and LIBOR should be regarded as parametric terms. The final estimation model with Price/Dwt is as follows.

$$g\left(E\left(NB_i\right)\right) = \beta_0 + s\left(Dwt_i\right) + s\left(Orderbook\%Total_i\right) + s\left(BDI_i\right) + \beta_1 \cdot No.Hatches + \beta_2 \cdot GDP\ Per\ Capita_i + \beta_3 \cdot Exchange\ Rate_i + \beta_4 \cdot LIBOR_i \quad \text{(B.1)}$$