



Ordinary Least Squares Regression for Predict Container Delay

Ning Ratih Fauziah¹, Indra Ranggadara², Anita Ratnasari³

^{1,2,3}Faculty of Computer Science, Mercu Buana University, Jakarta, Indonesia

41816110062@student.mercubuana.ac.id¹, indra.ranggadara@mercubuana.ac.id², anita.ratnasari@mercubuana.ac.id³

ABSTRACT

In Indonesia, many manufacturing companies use logistics services, commonly referred to as transporters, to distribute goods. In practice in the field experienced many obstacles transporter. One of them is during the door to the port process. Process complex business is one of the reasons the door's length to the port process, and one same factor effect is during the loading process in the warehouse Sender. So that the door to the port process can be faster, factors which effects must be minimized. In this research, the author uses two variables: delay in the loading warehouse as variable X or independent variable, and process delay door to port as they are variable Y or dependent. This study aims to determine how much the influence of the X variable on the Y variable. Equation OLS regression from this study obtained $Y = 2.2564 + 1.3546X + 0$. The Independent variable that affects the dependent variable based on OLS regression gets an R^2 value of 0.9584. This shows that the warehouse's delay capability (X) influencing the door to port delay (Y) is 96%. Researchers also used the MAPE method to calculate system accuracy value. The MAPE value obtained is 1.676%, which means that the system accuracy value is quite good, namely amounted to 98.324%.

Key words: information systems, logistics, sea transportation, monitoring, OLS.

1. INTRODUCTION

According to data at Bappenas 2018, Indonesia's population in 2019 is projected to reach 266.91 million souls. To meet the community's needs from Sabang to Merauke, necessary to distribute the goods to make it possible up to consumers. Many manufacturing companies are using logistics services or commonly known as freight forwarders, distributing these goods.

A freight forwarder is a business entity provide services for all activities that are shared implementation of delivery, transportation, and receipt goods using multimodal transport either through land, sea, and air[1]

A forwarder is a service-based industry, and its quality of service is an essential factor for the success of business.

Service quality is a crucial issue of the organizational to pursue customer satisfaction as the goal of business. Many scholars already discuss liner shipping and third-party logistics service providers' service quality issues in the literature. However, the past researches apply quality function deployment (QFD) to explore international freight forwarders' service quality requirements[1].

In practice, there are many obstacles experienced by logistic companies in the distribution of goods in the field. One of them is during the door to the port process. Like waiting in line in the warehouse, congestion on the road, when picking up containers empty, queued when the full stack and others.

From the results of observations made by researchers on one of the logistic companies in the Jakarta and Surabaya branches, here are three main factors that cause it to delay the door to port process.

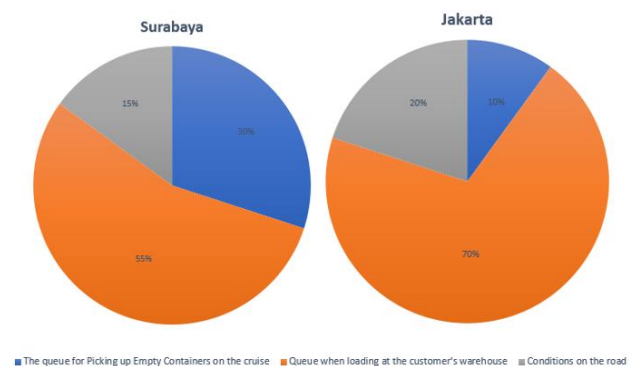


Figure 1: Factors that cause delay door to port

Figure 1 explains both in Surabaya and in Jakarta is the main factor causing delays in the door to port process is at the customer's warehouse.

Regression analysis is a study of the relationship of one dependent variable with one or more independent variables. If the independent variable is more than one, it is called multiple linear regression because the influence of several independent variables will be imposed on the dependent variable. In general, to estimate the regression coefficient in a regression model, the Ordinary Least Squares (OLS) method is used, namely by minimizing the number of residual squares[2].

In this study, the OLS method applied to the monitoring information system. To analyze how much influence load delays to the door to the port process. The trial was carried out using the length of the load delay done by the transporter, which becomes the independent variable(x) and the process door's length to port as a bound variable (y) to get value R^2 .

2. STUDY LITERATURE AND PREVIOUSRESEARCH

2.1 Freight Forwarder

A forwarder is a service-based industry, and its quality of service is an essential factor for the success of business. Service quality is a crucial issue of the organizational to pursue customer satisfaction as the goal of business. In the literature, some of scribe already discuss the service quality issues of liner shipping and third-party as logistics service providers; however, few research types apply quality function deployment (QFD) to explore international freight forwarders' service quality requirements[1].

Suyono (2005) provides an understanding of freight forwarder, namely a business entity that aims to provide services or management of all activities involved to implement delivery, transportation, and receipt of goods using multimode transport either by land, sea, and air. Then, the Freight forwarder also carries out processing procedures and formalities documentation required by government regulations of export countries, transit countries, and import countries under the scope of business. Next, the freight forwarder also settles the costs incurred due to transportation activities, cargo handling at the port or warehouse, handling documentation, and covers insurance liabilities that the owner generally requires.

2.2 Logistics Information System

The logistic information system is a communication network broader existing within the organization, or company preparations must be made to make or send and receive information conveyed by the organization or company. A logistics system is made up of a series of facilities connected to transportation services. Logistics Information Systems include procurement of raw materials, recording of goods orders, storage of raw materials and finished goods and distribution, or delivery item[3]. Including secure data archiving by moving data to cloud storage[4]. Innovative technology like IoT, cloud computing, big data, and data mining has generated much interest due to playing a vital role in creating a new business and converting the world to an infinite number of information and communication systems[5].

2.3 Ordinary Least Squares (OLS)

Multiple linear regression analysis is one statistical method used to model and investigate the relationship between one dependent variable and two or more independent variables. One of the methods used Ordinary method is used to estimate regression parameters Least Squares (OLS). The OLS method is susceptible to the presence of deviations from the assumptions on the data. Some of the assumptions are between another is the residuals must be normally distributed, the

variance homogeneous, and no autocorrelation occurs. If the data is not fulfilled, one assumption is that the OLS estimator obtained is not efficient due to an outlier existence.

The OLS regression model can be used to find the best relationship between the independent variable X and the variable dependent Y , measures the strength of the relationship, and performs prediction of variable Y through observed variable X . The method of estimating parameters in OLS regression uses the Least Square Method (MKT). Linear regression analysis model doubles are depicted with

$$Y_i = \beta_0 + \beta_1 X_i + e \quad (1)$$

Where:

Y : dependent variable (response variable)

X : independent variable (predictor variable)

β_0 : constant (intercept)

β_1 : parameter in variable to k , where $i: 1, 2, \dots, p$ [6]

2.4 Previous Research

Previous research is a reference in research making this time. In dealing with society ASEAN, economies need to apply a technological system informatics[3] and to face Industry 4.0, and companies logistics must also apply digitization so that business processes to be more efficient[7]. A very logistic information system is needed in the logistic process to support performance because it can help company leaders supervise logistic goods[8]. Information systems logistics are also needed to simplify services and monitor entering and leaving goods[9]. Information system logistics are also needed to assist the management process data and distribution so that goals are achieved and can well be monitored and managed[10].

Many Ordinary Least Square Regression Models are used in various fields, one of which is in the field management[11]. Apart from that, the OLS method is also used in research in agriculture[12]. The OLS method is also used in Health[13][14]. The OLS method can also be used in social research[2].

Based on literature studies from several scientific journals and problem backgrounds, the OLS Regression Method is considered good enough to apply to a monitoring system. For that research, this time intends to contribute by making the system logistic monitoring information using the OLS method to automatically detect how much effect the delay has on Warehouse loading against the door to port processing speed.

3. RESEARCH METHOD

3.1 Data Collection

This time, the research took place in Jakarta's logistics company using data delay door to port in 2019. Data retrieval refers to the module container loading to get data on how long the process took Cargo warehouse and Container Exchange modules to forget the data how long it took the door to port process like depicted in figure 2. for later made application of the OLS method into the module. In data processing, the author uses the Matlab version R2013a (8.1.0.604).



Figure 2: Container Exchange

3.2 Ordinary Least Squares (OLS) Method

The OLS regression model can be used to find the best relationship between the independent variable X and the variable dependent Y, measures the strength of the relationship, and performs prediction of variable Y through observed variable X[15]. The method of estimating parameters in OLS regression uses the Least Square Method (MKT). Linear regression analysis model doubles are depicted with

$$Y_i = \beta_0 + \beta_1 X_i + e \tag{1}$$

Where:

- Y : dependent variable (response variable)
- X : independent variable (predictor variable)
- e : eror
- β_0 : constant (intercept)
- β_1 : parameter in variable to k, where i: 1,2,..., p[6]

The coefficient of determination (R^2) measures the proportioned diversity or variation in the total response variable explained by predictor variables. Incredible value R^2 can be retrieved from[16]

$$R^2 = 1 - \frac{RSS}{TSS} \tag{2}$$

Rated R^2 ranges between $0 \leq R^2 \leq 1$. The higher the value of R^2 the higher the predictor variable's power to explain the response variable in regression analysis. So that better predictions of response variables can be obtained[16].

3.3 Percentage Error (MAPE)

MAPE is the most accurate prediction model, each model is considered to be applied to the past data, and a model with the minimum total error is chosen. Second, it is necessary to assess the accuracy periodically, detect abnormalities and deficiencies in the model may arise later. To evaluate the accuracy and forecasting the performance of different models. This research adopts the Percentage Error index (MAPE). MAPE is an absolute value of the data's percentage error to the mean or formulated[17].

$$MAPE = \frac{\sum (\frac{Prediksi - Actual}{Actual}) \times 100}{n} \tag{3}$$

4. RESULT AND DISCUSSION

The first process of OLS is to do calculations on the data set. Searching for the OLS equation begins by calculating the parameter value in the variable or β_1 , then looking for a constant value or β_0 then looking for an error value term e . After that, to find out how significant the independent variable's influence on the dependent variable must seek the value of R^2 . Rated R^2 ranges from 0 to 1 with the conclusions. The greater the value, the better the prediction variable can be obtained.

4.1 OLS Equation Calculation

The data set used the door to port delay data in 2019 in Jakarta with 365 data. Started by finding the value of β_1 in the following steps

$$\beta_1 = \frac{(\sum xy - \frac{(\sum x)(\sum y)}{n})}{(\sum x^2 - \frac{(\sum x)^2}{n})} = \frac{(798813 - \frac{(12045 \times 17143)}{365})}{(569561 - \frac{(145082025)}{365})} = 1,3546 \tag{4}$$

After obtaining the variable coefficient value or β_1 , Stepnext is to find the value for β_0 , with steps asfollowing

$$\beta_0 = \frac{\sum y}{n} - \left(\beta_1 \times \frac{\sum x}{n} \right) = \frac{17143}{365} - \left(1,3546 \times \frac{12045}{365} \right) = 2,2564 \tag{5}$$

Then look for the term e error value as follows

$$\sum y - \sum \hat{y} = 17143 - 17143 = 0 \tag{6}$$

From the results of these calculations, the OLS equation is obtained as follows $y = 2,2564 + 1,3546x + 0$.

4.2 Calculation of the coefficient of determination (R^2)

After obtaining the OLS equation, it is also necessary to look for the coefficient of determination to determine how much large ability variable x data influences variable y. The known OLS equation proceeds with prediction calculations and looks for the Total Sum of Square value (TSS) and the Residual Sum of Square (RSS). From the data set studied, the TSS value was 329467.6055, and the RSS value of 13718,6757. So that it can be calculated the value of R^2 as follows:

$$R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{13718,6757}{329467,6055} = 0,9584 \tag{2}$$

So we get the value of R^2 of 0.9584 or the 96% who shows that the ability of the X variable to the Y variable is quite large.

4.3 Calculation of Percentage Error (MAPE)

After getting the predicted value, calculate the system accuracy value with the formula 100% reduced MAPE value. The MAPE value is obtained as follows:

$$MAPE = \frac{\sum \frac{(\text{Prediksi}-\text{Actual})}{\text{Actual}} \times 100}{n} = \frac{611,8346}{365} = 1,676\% \quad (3)$$

So that the system accuracy value is $100\% - 1,676\% = 98,324\%$.

4.4 Visualization

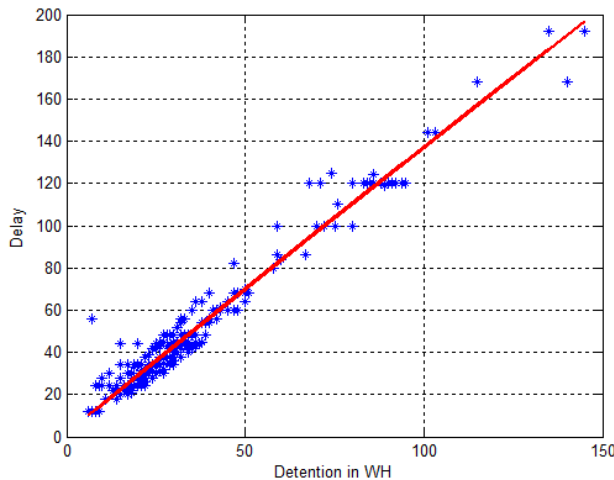


Figure 3: Scatter Plot Visualization

Figure 3 is a connecting scatter plot of the door's value to port delay on the y-axis. The value ranges from 0 to 200 in hours, with a delay value in the warehouse on the x-axis, which has a value range of 0 up to 150. From the scatter plot above, it can be seen that the distribution of data or coordinate points is very close to the linear line; this shows that the error is relatively small. From the overall linear line, you can see the pattern from the bottom left to the top right, which can be said if the delay in the warehouse is getting longer, the door to port delay will also be longer.

5. CONCLUSION

Based on the analysis results, the OLS regression model is obtained in this case study, and it is $y = 2,2564 + 1,3546x + 0$. The β_0 value is 2,2564, which means that if the loading warehouse's delay is zero, the door to the port process can only occur in 2.2564 hours, and the β_1 value is 1,3546, which means that if the delay in the loading warehouse increases, 1 hour, the door to port process will take 1.3546 hours. Score R^2 in this study is 96% explained that the delay ability in the loading warehouse is very influential on the door to port processing speed.

In this study, only one variable was used independently, namely delay in the loading warehouse. Therefore the next research can add several other variables that also affect the door to port processing speed and can using other regression models to compare the result.

REFERENCES

[1] S. T. Huang, E. Bulut, and O. Duru, "Service quality evaluation of international freight forwarders: an

empirical research in East Asia," *J. Shipp. Trade*, vol. 4, no. 1, 2019, doi: 10.1186/s41072-019-0053-6.

[2] I. Joseph and N. Elda, "Homicide and Social Media: Global Empirical Evidence," 2019.

[3] P. Kampan, "Strategic Development of ASEAN Logistics Infrastructure," *Open Transp. J.*, vol. 11, no. 1, pp. 67–80, 2017, doi: 10.2174/1874447801711010067.

[4] I. Ranggadara, N. Legowo, and Sfenrianto, "Migrating applications to clouds with TOGAF framework at MERAKI digital Indonesia," *J. Theor. Appl. Inf. Technol.*, vol. 97, no. 9, pp. 2515–2525, 2019.

[5] M. Al-Fawa'reh and M. Al-Fayoumiy, "Detecting stealth-based attacks in large campus networks," *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 9, no. 4, pp. 4262–4277, 2020, doi: 10.30534/ijatcse/2020/15942020.

[6] U. Madubuko Cyril, N. Ifeyinwa Elizabeth, and M. Chris Chukwuemeka, "Impact of Fair Value Accounting on Biological Assets in Agricultural Sector in Nigeria," *Int. J. Financ. Bank. Res.*, vol. 5, no. 4, p. 64, 2019, doi: 10.11648/j.ijfbr.20190504.11.

[7] S. Winkelhaus and E. H. Grosse, "Logistics 4.0: a systematic review towards a new logistics system," *Int. J. Prod. Res.*, vol. 58, no. 1, pp. 18–43, 2020, doi: 10.1080/00207543.2019.1612964.

[8] W. Ul-Hameed, M. S. Shabbir, M. Imran, A. Raza, and R. Salman, "Remedies of low performance among pakistani E-logistic companies: The role of firm's IT capability and information communication technology (ICT)," *Uncertain Supply Chain Manag.*, vol. 7, no. 2, pp. 369–380, 2019, doi: 10.5267/j.uscm.2018.6.002.

[9] M. Gordon *et al.*, "MODELS AND APPLIED INFORMATION TECHNOLOGY FOR SUPPLY," vol. 1, no. 7, pp. 59–68, 2019.

[10] M. Jamil and M. R. Lessy, "Web Application to Monitor Logistics Distribution of Disaster Relief Using the CodeIgniter Framework," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 325, no. 1, 2018, doi: 10.1088/1757-899X/325/1/012015.

[11] B. George, S. Van de Walle, and G. Hammerschmid, "Institutions or Contingencies? A Cross-Country Analysis of Management Tool Use by Public Sector Executives," *Public Adm. Rev.*, vol. 79, no. 3, pp. 330–342, 2019, doi: 10.1111/puar.13018.

[12] T. Narunart and V. Panjakajornsak, "An empirical analysis of factors affecting customer loyalty to sea freight forwarders in Thailand," *Asia-Pacific Soc. Sci. Rev.*, vol. 19, no. 3, pp. 128–143, 2019.

[13] T. Smith, E. Johns-wolfe, G. S. Shields, J. Malat, F. Jacquez, and G. M. Slavich, "Associations between lifetime stress exposure and prenatal health behaviors," vol. 36, no. 3, pp. 384–395, 2020, doi: 10.1002/smi.2933.Associations.

[14] S. A. Shinde and P. Raja Rajeswari, "A novel hybrid framework for cuff-less blood pressure estimation based on vital bio signals processing using machine learning," *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 9, no. 2, pp. 1556–1561, 2020, doi: 10.30534/ijatcse/2020/99922020.

[15] S. F. Husin, M. Mamat, M. A. H. Ibrahim, M. Rivaie,

- and M. A. Mohamed, "Modification of search direction in steepest descent method and its application in regression analysis," *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 9, no. 1.1 Special Issue, pp. 90–95, 2020, doi: 10.30534/ijatcse/2020/1791.12020.
- [16] A. T. Bon, "Model of Physical Condition of Leg Muscle Explosive Power , Nutrition Status and Confidence Towards Skill Smash Sepaktakraw," pp. 2062–2069, 2019.
- [17] A. de Myttenaere, B. Golden, B. Le Grand, and F. Rossi, "Mean Absolute Percentage Error for regression models," *Neurocomputing*, vol. 192, pp. 38–48, 2016, doi: 10.1016/j.neucom.2015.12.114.