Analysis of Regression Algorithm to Predict Administration, Production, and Delivery to Accuracy of Delivery of Products in Cosmetic Industry

W Gata1, H B Novitasari2, R Nurfalah3, R Hernawati4, M J Shidiq5
STMIK Nusa Mandiri, Jakarta, Indonesia
Email : gata81@gmail.com

Abstract. The purpose of this study is to apply a linear regression algorithm to predict the accuracy of product shipments using administration, production and delivery as attributes that will be compared with existing targets which are the results of agreements between companies and customers used as the dependent variable. In this study it was found that a linear method is feasible and effective for predicting administration, production and delivery of targets for accuracy of product shipments in the company. From the results of the prediction formula obtained, the prediction formula is implemented in the July 2018 data as test data. The results obtained from the prediction formula have an accuracy of 91.67%. From previous studies, which say that linear regression is feasible and effective for data use, they must make predictions. Likewise with this research where the results obtained in testing data are feasible and effective for the company.

1. Introduction
Production planning is the planning of production and manufacturing modules in a company or industry. It utilizes the resource allocation of activities of employees, materials and production capacity, in order to serve different customers [1]. The company engaged in the cosmetic industry in Indonesia and focuses on contract manufacturers or tolls requires good production planning, from the beginning of the sales order to the delivery process to the customer. But often planning does not match the reality in operation, so that it occurs not in accordance with the lead time. Because of this, the company needs data analysis or data mining in order to predict in the future. In this study, categorization consists of three things, namely: Administration, Production, and Delivery (APD) as attributes that will be compared with existing targets. It is expected that the research is to increase customer trust in the company, because the company can make predictions going forward based on existing data.

Administration time is meant is an Office Administration, where humans are the most important factor in achieving satisfactory and efficiency office services [2]. Production time is the beginning of the implementation of product manufacturing carried out by the production department, where the raw materials and packaging materials have been prepared in advance. Delivery Time is the time when the production process is complete and has become Finish Good until the Finish Good is sent to the customer. This activity is not just a product delivery but it is necessary to pay attention to the completeness of the documents and the condition of the product to be shipped in accordance with the handling requirements of the product.

The purpose of this study is to apply the Linear regression algorithm to predict the accuracy of product delivery using the APD model, where each time of APD are used as independent variable.
While the time of the target which is the result of an agreement between the company and the customer is used as a dependent variable. In this study it was found that a linear method is feasible and effective for predicting administration, production and delivery of targets for accuracy of product shipments in the company based on training data from January 2018 to June 2018 and tested with July 2018 data.

2. Method

Regression models are widely used for addressing scientific questions of interest regarding the associations among a set of variables. In particular, linear regression models describe how part of the natural individual-to-individual variation in a continuous response variable can be explained by one or more explanatory variables [3]. Linear regression describes how the mean values of a response variable (denoted by Y) vary as a linear function of a set of explanatory variables and X as independent variables.

The multiple linear regression model is built on the same foundation as simple linear regression, and the four fundamental assumptions made with simple linear regression must also be true for multiple linear regression [4]. Multiple linear regression models can be stated as follows:

\[
y_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + \cdots + \beta_p x_{pi} + \mu_i
\]  

(1)

where the meaning of the symbol above is:

- \(\beta_1\) = intercept of the model
- \(\beta_2, \beta_3, \ldots, \beta_p\) = Partial Regression coefficients of the i dependent variable
- \(x_{2i}\) = i-independent variables with parameters
- \(y_i\) = i dependent variable
- \(\mu_i\) = residual (error) for the first observation

There are several previous studies that used a linear regression classification model, including:

1. Linearity studies on detection and quantification of avermectins using linear. In this study, linear relationships between response and concentration were used to estimate the detection limit (DL) and quantification limit (QL) for five avermectins: emamectin, abemectin, doramectin, moxidectin, and ivemectin. Linearity assessment was performed by linear regression, which incorporated a regression model, outlier rejection, and evaluation of the assumption with a significant test [5].

2. Aero-Material consumption prediction based on linear regression models

In this paper, the mathematic model and calculation method of linear regression model are introduced. And the parameter estimation and model test method of linear regression model is discussed. The application of linear regression model in forecasting the consumption of aero-material spare parts is analyzed by examples. The results show that the linear regression model is feasible and effective for the prediction of aero-material spare parts consumption [6].

3. Predicting ethnicity with first names in online social media networks. This study provide a method to upgrade individual detail in terms of ethnicity in data gathered from social media via the use of register data. This research aim to predict the most likely value of ethnicity, given one’s first name, and to show how one can test hypotheses with the predicted values for ethnicity as an independent variable while simultaneously accounting for the uncertainty in these predictions [7].

4. Linear regression models for prediction of annual heating and cooling demand in representative Australian residential dwellings. This paper presents the development methodology of linear regression models that were developed for the prediction of annual thermal loads in representative residential buildings across three major climates in New South Wales, Australia, and the assessment of the impact of building envelope upgrades. The results presented show that the linear models with simple independent variables can predict the requirements for space heating and cooling of the residential buildings in the specific climates within acceptable errors [8].
5. An impact of linear regression models for improving the software quality with estimated cost. In this article, an application of defects removal effectiveness to improve the software quality and fault prone analysis, methods are finding the solution of parameters in linear regression models with cost estimating method. It is enabling to improve software development processes by target reducing the estimated cost for software products and improve the techniques for more effectively and efficiently [9].

6. In a multistage manufacturing process, extensive amounts of observational data are obtained by the measurement of product quality features, process variables, and material properties. These data have temporal and spatial relationships and may have a non-linear data structure. It is a challenging task to model the variation and its propagation using these data and then use the model for feedforward control purposes. This article proposes a methodology for feedforward control that is based on a piecewise linear model. An engineering-driven reconfiguration method for piecewise linear regression trees is proposed. The model complexity is further reduced by merging the leaf nodes with the constraint of the control accuracy requirement. A case study on a multistage wafer manufacturing process is conducted to illustrate the procedure and effectiveness of the proposed method [10].

3. Results And Discussion
The KDD model can be seen in Figure 6.

![Diagram of KDD model]

**Figure 1.** Method of Knowledge Discovery Database.

In this study using the Knowledge Discovery Database (KDD) method, the KDD stages in Figure 1, above are as follows:

1. Data Selection
   At this stage, data is selected from a set of operational data. In this study, PT Cedefindo's production history data was selected in the period of the delivery date in 2018, starting from one January 2018 to 30 June 2018 with a total history of 712 data.

2. Pre Processing Data
   Data cleaning stage is carried out including removing data duplication, checking data inconsistencies and data errors. Production history data used in this study, initially as many as 712 data became 640 data.
3. Data Transformation
After going through the pre-processing stage as many as 640 data consists of 7 attributes, namely Order Date, Order Number, Delivery Target, Product Code, Production Date, Product Transfer Note (PTN) Date, and Send Date are transformed into 4 attributes, namely Target, Administration, Production and Delivery.

4. Data Mining
This study uses a data mining algorithm to process data, namely the Linear Regression algorithm. Data that has gone through the transformation phase is then processed in Rapid Miner using the Linear Regression algorithm.

5. Interpretation or Evaluation
The pattern that is generated from the data mining process is presented in the form of information that is easy to understand and checks whether the pattern or information found is contrary to the facts or hypotheses that existed before. It is expected that this research can help improve the effectiveness and efficiency of work that affects the target delivery of products.

In this study using a software tool to produce analysis data from linear regression algorithms [7]. The software tools are Rapidminer version 9.0, in order to be able to use the software tools, we need to prepare production history data that has gone through the stages of pre-processing and data transformation data into Excel format. The following is an example of production history data after going through the selection phase in table 1.

| No | Order Date | Order Number | Delivery Target | Product Code | Production Date | PTN Date | Send Date |
|----|------------|--------------|-----------------|--------------|----------------|----------|-----------|
| 1  | 2017-05-30 | OPE-17-05-00119 | 2018-03-30 | 6773JMLRGR10 | 2018-05-30 | 2018-06-29 | 2018-06-30 |
| 2  | 2017-08-16 | OPE-17-08-00144 | 2018-06-29 | 63135BHMPM10 | 2018-06-04 | 2018-06-11 | 2018-06-21 |
| 3  | 2017-07-21 | OPE-17-07-00132 | 2018-04-12 | 5803ANNVJS10 | 2018-04-12 | 2018-04-19 | 2018-04-20 |
| 4  | 2017-08-16 | OPE-17-08-00147 | 2018-05-04 | 63135PMFCM10 | 2018-05-03 | 2018-05-11 | 2018-05-11 |
| 5  | 2017-08-12 | OPE-17-08-00129 | 2018-05-29 | 5803ANNVIP20 | 2018-02-23 | 2018-03-01 | 2018-03-06 |
| 6  | 2017-08-22 | OPE-17-08-00130 | 2018-02-28 | 5803ANLMFC10 | 2018-02-14 | 2018-02-22 | 2018-02-26 |
| 7  | 2017-08-16 | OPE-17-08-00148 | 2018-04-30 | 63135MRHRWN20 | 2018-04-05 | 2018-04-12 | 2018-04-13 |
| 8  | 2017-06-22 | OPE-17-06-00131 | 2018-02-28 | 5803ANLMIBN10 | 2018-02-08 | 2018-02-19 | 2018-02-20 |
| 9  | 2017-06-20 | OPE-17-06-00077 | 2018-02-28 | 5313DAIS8410 | 2018-02-02 | 2018-03-15 | 2018-03-15 |
| 10 | 2017-06-22 | OPE-17-06-00123 | 2018-02-09 | 5803ANLM5SD10 | 2018-01-30 | 2018-02-12 | 2018-02-19 |

As explained in Table 1, with the KDD method after performing the data selection, it is necessary to do the pre-processing and data transformation data stages. From the production history data, the original data as much as 712 data becomes 640 data and produces the following attributes:

| Target          | Total days from Order Date to Delivery Target |
|-----------------|-----------------------------------------------|
| Administration  | Total days from Order Date to Production Date |
| Production      | Total days from Production Date to PTN Date   |
| Delivery        | Total days from PTN Date to Send Date         |

The following are data that have been determined attributes and components X, Y that will be processed on Rapidminer as shown on Table 2.
Based on data processing using Rapidminer version 9.0, it can be seen the output of data processing results found in Table 3.

| Attribute     | Coefficient (b) | Standard Error | Tolerance | t-Stat | p-Value | Code |
|---------------|-----------------|----------------|-----------|--------|---------|------|
| Administration (X1) | 0.925 | 0.019 | 0.856 | 0.992 | 48.139 | 0.000 **** |
| Production (X2)   | 0.648 | 0.060 | 0.191 | 0.993 | 10.738 | 0.000 **** |
| Delivery (X3)    | 0.191 | 0.067 | 0.050 | 1.000 | 2.835  | 0.005 ***  |
| Intercept        | 6.105 | 2.068 |       |       | 2.953  | 0.003 *** |

From the above output it can be formulated a mathematical regression equation with administration, production and delivery variables as follows:

Prediction (yi) = Intercept + b1X1 + b2X2 + b3X3  
Prediction (yi) = 6.105 + (0.925 x Administration) + (0.648 x Production) + (0.191 x Delivery)  

The intercept or constant value of 6.105 states that if there are no administrative, production and shipping variables, then the target of delivery is 6,105 days [10]. The coefficient for the independent administrative variable (X1) is positive at 0.925 indicating that if the other independent variables are fixed and the administration increases 1%, the delivery target will increase by 0.925. The coefficient for the independent variable of production (X2) has a positive value of 0.648 indicating that if the other independent variables are fixed values and production has increased by 1%, the delivery target will increase by 0.648. The coefficient value for the independent variable of delivery (X3) positive value of 0.191 shows that if the other independent variables are fixed values and shipments increase by 1%, the delivery target will increase by 0.191. From the above equation, the data can be generated as show in Table 4.
Table 4. Results of linear regression calculations.

| Target (Y) | Administration (X1) | Production (X2) | Delivery (X3) | Prediction (yi) |
|------------|---------------------|----------------|--------------|----------------|
| 304.00     | 365.00              | 30.00          | 1.00         | 363.361        |
| 317.00     | 292.00              | 7.00           | 10.00        | 282.651        |
| 265.00     | 265.00              | 7.00           | 1.00         | 255.957        |
| 261.00     | 260.00              | 8.00           | 0.00         | 251.789        |
| 260.00     | 246.00              | 6.00           | 5.00         | 238.498        |
| 251.00     | 237.00              | 8.00           | 4.00         | 231.278        |
| 257.00     | 232.00              | 7.00           | 1.00         | 225.432        |
| 251.00     | 231.00              | 11.00          | 1.00         | 227.099        |
| 253.00     | 227.00              | 41.00          | 0.00         | 242.648        |
| 232.00     | 222.00              | 13.00          | 7.00         | 221.216        |
| 253.00     | 219.00              | 18.00          | 0.00         | 220.344        |
| 168.00     | 217.00              | 7.00           | 4.00         | 212.130        |

From the results of the linear regression calculation above, it can be seen that in the Target (Y) column for 265 days, Administration (X1) for 265 days, Production (X2) for 7 days and Delivery (X3) for 1 day then the results of Prediction (yi) are 255,957 days, this indicates that there is a difference between Target (Y) and Prediction (yi) of 9.04 days. If Target (Y) for 253 days, Administration (X1) for 227 days, Production (X2) 41 days and Delivery (X3) for 0 days, the result of Prediction (yi) is 242,648 days, indicating that there is a difference between Target (Y) and Prediction (yi) of 10.35 days. If Target (Y) for 253 days, Administration (X1) for 219 days, Production (X2) 18 days and Delivery (X3) for 0 days, the result of Prediction (yi) is 220,344 days, indicating that there is a difference between Target (Y) and Prediction (yi) of 32.66 days. From the overall data of this study, it can be calculated that the total Target (Y) is equal to 55901 and the total Prediction (yi) is 55890.84, indicating that there is a difference between Target (Y) and Prediction (yi) which is 10.16 days.

From the results of the prediction formula obtained, the prediction formula is implemented in July 2018 data as testing data. The results obtained from the prediction formula have accuracy of 91.67%. From previous research, which said that linear regression is feasible and effective for the use of data that they have to make predictions. Likewise with this research where the results obtained in testing data are feasible and effective for the company.

4. Conclusion
From this study it was found that the linear method is feasible and effective to predict the administration, production and delivery of targets for the accuracy of product delivery in the company based on training data from January 2018 to June 2018 and tested with July 2018 data. So companies are advised to use the results of this study to predict future targets to improve service to customers in accordance with the research objectives.

References
[1] Fargher, H. E., & Smith, R. A. (1996). U.S. Patent No. 5,586,021. Washington, DC: U.S. Patent and Trademark Office.
[2] Upadhyay, R., Ladhe, Y. P., Rai, R. K., Bhatkar, C. B., & Upadhyay, R. (2015). OFFICE MANAGEMENT SYSTEM OF AN EDUCATIONAL INSTITUTE-A CASE STUDY ON SHRI DADAJI INSTITUTE OF TECHNOLOGY & SCIENCE, KHANDWA (MP). International Journal of Mechanical Engineering and Robotics Research, 4(2), 72.
[3] Fitzmaurice, G. M. (2016). Regression. Diagnostic Histopathology, 22(7), 271-278.
[4] Marill, K. A. (2004). Advanced statistics: linear regression, part II: multiple linear regression. *Academic emergency medicine, 11*(1), 94-102.

[5] Ismail, R., Lee, H. Y., Mahyudin, N. A., & Bakar, F. A. (2014). Linearity study on detection and quantification limits for the determination of avermectins using linear regression. *Journal of food and drug analysis, 22*(4), 407-412.

[6] Yang, Y., Sun, L., & Guo, C. (2018). Aero-material consumption prediction based on linear regression model. *Procedia computer science, 131*, 825-831.

[7] Hofstra, B., & de Schipper, N. C. (2018). Predicting ethnicity with first names in online social media networks. *Big Data & Society, 5*(1), 2053951718761141.

[8] Aghdaei, N., Kokogiannakis, G., Daly, D., & McCarthy, T. (2017). Linear regression models for prediction of annual heating and cooling demand in representative Australian residential dwellings. *Energy Procedia, 121*, 79-86.

[9] Marandi, A. K., & Khan, D. A. (2015). An Impact of Linear Regression Models for Improving the Software Quality with Estimated Cost. *Procedia Computer Science, 54*, 335-342.

[10] Jin, R., & Shi, J. (2012). Reconfigured piecewise linear regression tree for multistage manufacturing process control. *IIE Transactions, 44*(4), 249-261.