Warranty Claim Quantity Forecasting via Multiple Linear Regressions

N Haris¹, A K Junoh² and W Z A Wan Muhamad³

¹,²,³Engineering Mathematics Institute, Faculty of Applied and Human Sciences, Universiti Malaysia Perlis, Pauh Putra Campus, 02600 Arau, Perlis, Malaysia

haris.nursafian@marelli.com

Abstract. A concrete justification unavailability of the roughly estimated warranty claim quantity is leading to also inability to understand defect trend behavior and its effect upcoming warranty claim quantity. This is where an equation model is derived by considering the previous actual warranty data, to represent warranty claim defect quantity impact. Taking into consideration, identified parameters which link with pricing and cost, it also includes the observation and monitoring of warranty trend from the existing actual warranty data, by plotting cumulative defect quantity over Vehicle Line-off Date, as well as plotting cumulative defect quantity over Vehicle Submission Date. Multiple Linear Regression is deployed to define the best Multiple Regression Equation. Predictors, response and predictors’ validations defined by using normality and probability test. The successfully developed equation model, takes into account the existing warranty data and trend. As a result, the equation model managed to provide forecasted warranty claim quantity, based on a complete 36-month warranty period cycle, which has a significant impact on a reliable and convincing figure – a key factor in warranty budgeting and accrual task.

1. Introduction

Quality has always been one of the hottest topics when it comes to judging the performance of any product, including automotive products. In simple words, quality is understood as the level how good of an item or service is, high level of worthiness or fineness. To be specific, IATF 16949 Standard defines quality as ‘fitness for use’, and the most important is, can be used when during decision making involving quality matters [1]. Quality as an overall satisfaction rank, was sustained by the self-induced principles of specific possessors who were highly inspired by their superiority in quality and service [2].

As far as product quality is concerned, it is needed to take into consideration that deviations from the expected performance may occur when the vehicle is in use [3]. The automotive industry spends roughly around $10 – $13 billion annually in America on warranty claims, even approaching $40 billion worldwide, eating up roughly 1% – 5.2% of original equipment manufacturers’ (OEM) product profit and around 0.5% – 1% of vendor’s product profit [4]. Warranty claims and complementary data comprise beneficial data regarding product quality and reliability [5]. This medium is where warranty budgeting and accrual task are always one of the concerned topics in most automotive-based company management discussions. This is also where due to the current technological development, making warranty claim quantity forecasting is the key of any manufacturing company towards services planning; warranty cost prediction and satisfaction of customer [6].
On the other hand, company warranty forecasts are usually without any base of, neither mathematical nor statistical structure. It is on a very rough estimation of what potential spend risks there are with no account for the magnitude of the actual warranty data. Warranty cost underestimation brings momentous profit loss consequently while warranty cost overestimation will reduce competitiveness of the company. As most vehicle manufacturers are billing warranty cost to vendors on back-to-back basis, vendors might undergo gigantic profit loss from their sales turnover should there be no factual warranty forecasting initiated considering both quantity and cost wise. This is where paying warranty claims is a cost to the vendors [7]. Therefore, extracting information from warranty data and manipulating it for warranty cost prediction, in parallel with warranty trend monitoring, are of particular interest to both vendors and vehicle manufacturer [3].

Based on the aforementioned complications, the problem statement can be described as no clear understanding and visualization of warranty MIS trend for production month of January 2019, for product BLM FL Headlamp. This is because of a warranty performance-analysis-tool, not available in Vehicle Lighting Company. This is leading to inability to understand defect trend behaviour and how it will affect warranty claim quantity in the future. Furthermore, there is no concrete justification of the roughly estimated warranty claim quantity, because it is not assessed in a proper mathematical method without any thorough study, making it difficult to convince the management team.

To have a clear overview of warranty situation, this study proposes Multiple Linear Regression Analysis (MLRA). Being a part of data mining technique, it is useful to be deployed, wherein [8], data quality mining functions as the cautious instrument of data mining techniques, serves for data quality measurement and improvement.

The key objective is developing MLRA equation model, that represents current actual warranty MIS trend, and then using it for risk assessment to obtain forecasted defect quantity at 36 MIS. This is an operative and modest technique for predicting warranty cost equation model, considering previous actual warranty data, to represent warranty quantity. It is a form of distinguishing, enumerating and explaining precise data quality insufficiencies in an extremely huge database. This is where the estimation of warranty cost from available warranty databases has to be utilized by discovering the data patterns inside it.

2. Literature Review
This section explains MRLA as the main equation model covering concepts of warranty period, month-in-service and variables validation. Research flow chart illustrated as below figure 1 for clearer overview of the whole process.
2.1. Warranty Period
Automotive warranty is a treaty from the vehicle manufacturer to the consumers, which inaugurates accountability between among each other in any occurrence of product malfunction and dissatisfactory. Alternatively, warranty also affects a buyer informationally, in a way that occasionally buyers may conclude that a vehicle should be more dependable and robust when higher warranty coverage is included upon purchase [9]. This also functions as a market instrument, triggering reliability, and quality indicator. Hence, the consumer can be assured with the products execution as designed with no early failures [10].

Free labour charge and free defect part replacement are covered in automotive warranties, as long as the claimed vehicle is within both age and mileage range. Taking $l_t$ and $l_m$ as the age range (time-based) and mileage range (distance traveled based) for a warranty, the range of $l_t = 36$ months and $l_m = 100,000$ kilometres are deployed, where it indirectly explains the anticipated product performance and willingness to perform repair or replacement should there be any malfunction or dissatisfaction. Figure 2 below illustrates warranty period concept.

2.2. Months-in-service
Months in service (MIS), explained as the number of days (months) between vehicle sold date until failure reporting date, and mileage on the odometer at the time of failure reporting, are the two significant information associated with a certain failure mode [9]. MIS is defined as the service timeframe duration until it is repaired and reported [6]. To calculate MIS, four main variables as stated below are treated to
be the independent variables to determine the predictors. Terms may differ for every vehicle manufacturers. However, it boils up to a similar key point.

- **Vehicle Line-off Date (LO)**, or often called Vehicle Assembly Date, is the day when the vehicle assembly process finishes and is lined-off from the production assembly line, and ready for delivery to the vehicle dealers. This date is important to trace the concerned production day when the problem occurs.
- **Vehicle Initial Registration Date (IR)**, sometimes called Sold Date, is the particular date when the vehicle is out from the vehicle dealers and starts to service its’ respective owners on the road, and a reference of warranty period start date.
- **Vehicle Failure Date (FD)**, or sometimes-called Repair Date is when a repair job or part replacement is made to the particular complained failure.
- **Vehicle Submission Date (SD)**, also known as claim entry date, is reporting day to the Warranty Portal, usually done by the respective dealers who perform the repair job or part replacement.

There are four methods to determine MIS. The first is by considering Vehicle line-off date until repair date. Alternatively is to consider vehicle sold date minus repair date. Thirdly is by finding the duration of vehicle sold to date to repair date, Finally Yet Importantly, is by calculating vehicle sold date to submission date.

2.3. **Multiple Linear Regression**

As far as warranty claim prediction is concerned, there are already several manuscripts, venturing about it. By exploring previous warranty data as the main resources, Limon, Yadav and Nepal, [11] recommended a warranty projection based on the usage behaviour. The collected field data are then scrutinized to estimate lifetime parameters, which is used to anticipate the upcoming warranty claims.

Recently, there is a manuscript concerning utilizing MLRA for warranty cost forecasting, as in [12]. The manuscript proposed a method to acclimatize previous maintenance warranty data, differentiating Artificial Neural Network (ANN) and MLRA. Besides, reference [13] also proposed MLRA utilization on warranty claim quantity forecasting for electrical home appliances for products, which are still within the warranty period. The authors established a machine-learning model, with the direction to reduce forecast inaccuracy, resulting in improved warranty claim procedures, which are more transparent in the name of ideal supplies formation.

There is even a manuscript involving Volvo personnel, exploring MLRA in one of their failure rate calculation approaches. With the assumption of quality issues as an unanticipated surge in specific component failure rates, the authors developed the quick detection approach, by comparing two different machine-learning methods. An MLRA auto-regression model of the automobile failure rates, taking into account previous data serves as the main method. Meanwhile, the other one is singular vehicle failure forecast accumulation based on singular usage of each vehicle respectively [14].

Finding a linear connection among a dependent variable and numerous independent variables is the key determinant of multiple linear regression [15]. This also comprises hypothesis testing, estimating, and predicting linear connections, with the objective of demonstrating techniques that enables us to construct models, resulting in accurate response prediction for any upcoming observations [16]. Many applications of regression analysis involve situations in which there is more than one regressor variable [17], also called an independent variable.

There are four assumptions that should be fulfilled in order to initiate multiple linear regressions [18]. Firstly, data must be in a normal distribution form whereby the predictors and response are presumed to have a linear association. Every research struggles for the normal distribution data, as this is the finest research technique and data pattern [19]. Secondly, a normal and common multiple regression modeling has the ability to project figures of response from the predictors in a proper manner if the relationship between these two, are naturally linear. Thirdly, the regression line variance is similar for all predictor variable (X) values. Finally, unreliable measurement and data are avoided, so that relation between predictors and response is neither under-estimated nor over-estimated.
After the above assumptions are fulfilled, now multiple regression model can be explained as below equation.

\[ Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 \]  

which comprises a deterministic component involving the five regression coefficients (\(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \) and \(\beta_5\)).

There are three main functions for multiple linear regression analysis, besides causal analysis. Initially, it may function as an effective strength identification of the independent variables, or often called predictors, have on a dependent variable, or response often called.

Furthermore, MLRA is found beneficial for the effects or impacts of change predictions. It has been utilized extensively to accommodate inquiries such as forecasting gold prices in half a year time [20]. Many researchers had been implementing this method, to justify questions like, what is the strength of the relationship between dose and effect, vehicle sales forecast for the next few years [21] and marketing spend, and many other figures related matters from nature-related river flow regimes forecasting [22], and many others.

3. Methodology

3.1. Data Acquisition

Warranty data set considered is downloaded from a vehicle manufacturer warranty portal, consisting of 1086 data, containing January 2018 until November 2019 submission dates, and ranging from January 2017 - November 2019 Line-off Date [23]. After understanding and detecting a high rejection trend, the desired Line-off Date (January 2019) is selected.

This research focused on only January 2019 Line-off Month, consisting of 248 data (warranty claims). From 248 data, this section only explains examples of a few data to demonstrate how data is processed until it is ready for Multiple Linear Regression Analysis.

Below figure 3 is how the 1086 data, containing January 2017 until November 2019 Line-off dates look like. Warranty data is organized in a way that cumulative defect quantity is plotted over Vehicle Line-off Date. In this way, the performance of every line-off month can be visualized.

![Figure 3. Defect quantity vs vehicle line-off month graph of BLM Headlamp.](image-url)

Being a part of data mining objective, this is a form of huge volumes of data investigation to realize ‘binding, original, theoretically useful, and ultimately plausible patterns’ [24]. Therefore, this must be paired with plotting defect quantity over vehicle repair month (Figure 4). The main purpose is to visualize the defect quantity that is being reported and received on a time basis.
From Figure 3 & 4, it is discovered an unusual and alarming trend of warranty claims from vehicle line-off month January 2019, whereby the cumulative defect quantity is almost triple among all line-off months.

3.2. Response (Cumulative Defect Quantity) Calculation

There are 4 methods of calculating MIS. First, the timeframe between Vehicle Line-off Date to Vehicle Repair Date. \( X_1 = \text{Duration of Vehicle Line-off Date to Repair Date (in months)}, \) derived as per below equation.

\[
X_1 = FD - LO
\]

Secondly, the timeframe between Vehicle Line-off Date to Claim Submission Date, where \( X_2 = \text{Duration of Vehicle Line-off Date to Claim Submission Date (in months)}, \) derived as per below equation.

\[
X_2 = SD - LO
\]

Thirdly, timeframe, between Vehicle Initial Registration Date to Claim Submission Date, where \( X_3 = \text{Duration of Vehicle Initial Registration Date to Claim Submission Date (in months)}, \) derived as per below equation.

\[
X_3 = FD - IR
\]

Fourthly, timeframe between Vehicle Initial Registration Date to Vehicle Repair Date, where \( X_4 = \text{Duration of Vehicle Initial Registration Date to Vehicle Date (in months)}, \) derived as per below.

\[
X_4 = SD - IR
\]

Below is the example of MIS \( X_1, X_2, X_3 \) and \( X_4 \) calculations.

| Claim No | Line Off Date (LO) | Initial Registration Date (IR) | Failure date (FD) | Submission date (SD) | Odometer | MIS X1 (FD - LO) | MIS X2 (SD - LO) | MIS X3 (FD - IR) | MIS X4 (SD - IR) |
|----------|--------------------|--------------------------------|-------------------|----------------------|-----------|-----------------|-----------------|-----------------|-----------------|
| 594144   | 08/01/2019         | 14/01/2019                     | 09/02/2019        | 03/02/2019           | 13/02/2019| 553             | 1               | 1               | 1               |
| 655220   | 28/01/2019         | 03/02/2019                     | 25/03/2019        | 03/04/2019           | 1895      | 2               | 2               | 2               | 2               |
| 691957   | 18/01/2019         | 17/03/2019                     | 17/06/2019        | 09/07/2019           | 5301      | 5               | 6               | 3               | 4               |
| 662079   | 02/01/2019         | 08/02/2019                     | 01/09/2019        | 03/09/2019           | 9129      | 8               | 8               | 7               | 7               |
| 670603   | 17/01/2019         | 23/01/2019                     | 16/09/2019        | 08/10/2019           | 13233     | 8               | 9               | 8               | 9               |
Then, data is arranged according to MIS, from MIS1 until MIS9. Taking into account all 248 data of January 2019 Line-off Date, tabulation of data versus MIS is as below.

| MIS (X_n) | (ΣX_1) FD - LO total | (ΣX_2) SD - LO total | (ΣX_3) FD - IR total | (ΣX_4) SD - IR total |
|-----------|----------------------|----------------------|----------------------|----------------------|
| 1         | 2                    | 2                    | 2                    | 2                    |
| 2         | 13                   | 8                    | 17                   | 12                   |
| 3         | 30                   | 26                   | 34                   | 30                   |
| 4         | 25                   | 20                   | 52                   | 35                   |
| 5         | 77                   | 63                   | 47                   | 68                   |
| 6         | 25                   | 37                   | 29                   | 23                   |
| 7         | 24                   | 32                   | 26                   | 26                   |
| 8         | 16                   | 19                   | 25                   | 16                   |
| 9         | 36                   | 33                   | 16                   | 36                   |
| Total     | 248                  | 240                  | 248                  | 248                  |

After MIS in terms of X_1, X_2, X_3, and X_4 are defined, then it is possible to calculate the involved cumulative defect quantity from the existing data.

- The response is defined as Y_1, cumulative defect quantity in terms of X_1.
- The response is defined as Y_2, cumulative defect quantity in terms of X_2.
- The response is defined as Y_3, cumulative defect quantity in terms of X_3.
- The response is defined as Y_4, cumulative defect quantity in terms of X_4.

Below table 3 is the tabulation of Response (Cumulative Defect Quantity), calculated by accumulating each X_1, X_2, X_3, and X_4 respectively. Then it is visualized by plotting it as per figure 5.

| MIS (X_n) | (Y_1) FD - LO total | (Y_2) SD - LO total | (Y_3) FD - IR total | (Y_4) SD - IR total |
|-----------|----------------------|----------------------|----------------------|----------------------|
| 1         | 2                    | 2                    | 2                    | 2                    |
| 2         | 15                   | 10                   | 19                   | 14                   |
| 3         | 45                   | 36                   | 53                   | 44                   |
| 4         | 70                   | 56                   | 105                  | 79                   |
| 5         | 147                  | 119                  | 152                  | 147                  |
| 6         | 172                  | 156                  | 181                  | 170                  |
| 7         | 196                  | 188                  | 207                  | 196                  |
| 8         | 212                  | 207                  | 232                  | 212                  |
| 9         | 248                  | 240                  | 248                  | 248                  |

**Table 2. Defect Quantities according to MIS1 to MIS9 for January 2019 Line-off Date.**

**Table 3. Cumulative Defect Quantities according to MIS1 to MIS9 for January 2019 Line-off Date.**

**Figure 5.** Response (Cumulative Defect Quantity, Y_n) vs Predictors (Months-in-service, X_n).

---

The 1st International Conference on Engineering and Technology (ICoEngTech) 2021
Journal of Physics: Conference Series 1962 (2021) 012043 doi:10.1088/1742-6596/1962/1/012043
As illustrated in figure 5, the value of response $Y_1$, $Y_2$, $Y_3$, and $Y_4$ increase in a timely manner, as a result of how predictors $X_1$, $X_2$, $X_3$, and $X_4$ behave, as months-in-service increases.

3.3. Variables Validation
To define the best multiple regression equation modelling representing January 2019 line-off month, the predictors and responses are identified and validated. Statistical significance is conducted to fulfill the assumption of regression, where data sets are obeying normality and probability distribution, linearity behaviour, and significantly corresponds to each factor. The normality test is used to indicate whether to accept or reject the null hypothesis that data originated from a normally distributed population. It functions to decide whether an observed dataset is possible to be represented by a normal distribution [25]. If the residuals basically follow a straight line, it indicates the data are normally distributed. [26]

Then Anderson-Darling statistic evaluates the wellness of the data, obeying the fitted line in a probability plot [27]. Here, statistical significance is obtained by the observing p-value to be smaller than the defined significance level.

Probability plot functions to evaluate whether distribution fits the data, percentiles approximation, and different sample distributions judgment. It also allows distribution fit decision by observing points’ distribution along the line, where predictors linearity can be visualized. The key points here are performing probability plots and scatter plots, as mentioned in [28].

With an unambiguous function of demonstrating how much one variable affects another and finding connections between two variables, also called correlation, scatterplot resembles line graphs where during data points plotting, horizontal and vertical axes are utilized. A scatterplot shows the relationships between two variables [29].

Another useful statistical test validating and gauging connectivity between 2 or more continuous variables is the Pearson’s correlation coefficient ($r$), derived as below equation.

$$r = \frac{(N\sum xy - \sum x \sum y)}{\sqrt{(N\sum x^2 - (\sum x)^2)(N\sum y^2 - (\sum y)^2)}} \quad (6)$$

It is beneficial for multicollinearity detection, where coefficient value ‘$r$’ will vary from -1 to 1, indicating either variables are positively or negatively correlating, or even does not correlates at all when $r = 0$ [30].

4. Results and Discussions

4.1. Predictors and Response Validation
Warranty For response $Y_1$ evaluation, Anderson-Darling test found that $P_{value} = <0.005$ and normal. The probability plot shows predictors $X_1$ correlates linearly and achieve 95% Confidence Interval. Therefore, predictor $X_1$ is proven to be normally distributed, and the null hypothesis is accepted. It is also observed that the increase of months in service over cumulative defect quantity in proportional basis, where correlates in a linear manner. Figures below illustrate the result.

Figure 6. Anderson Darling Normality Test of Predictor $X_1$. 

$$\text{Normal}$$
For response \( Y_2 \) evaluation, Anderson-Darling test also found that \( P_{\text{value}} < 0.005 \) and normal. The probability plot shows predictors \( X_2 \) correlates linearly and achieve 95% confidence interval. Therefore, predictor \( X_2 \) is proven to be normally distributed and the null hypothesis is accepted. It is also observed that the increase of months in service over cumulative defect quantity in proportional basis, where correlates in a linear manner. Figures below illustrates the result.

Figure 9. Anderson Darling Normality Test of Predictor \( X_2 \).
For response Y_3 evaluation, Anderson-Darling test also found that P-value = <0.005 and normal. The probability plot shows predictors X_3 correlates linearly and achieve 95% confidence interval. Therefore, predictor X_3 is proven to be normally distributed, and the null hypothesis is accepted. It is also observed that the increase of months in service over cumulative defect quantity in proportional basis, where correlates in a linear manner. Figures below illustrate the result.
Finally for response $Y_4$ evaluation, Anderson-Darling test also found that $P_{value} = <0.005$ and normal. The probability plot shows predictors $X_4$ correlates linearly and achieve 95% Confidence Interval. Therefore, predictor $X_4$ is proven to be normally distributed, and the null hypothesis is accepted. It is also observed that the increase of months in service over cumulative defect quantity in proportional basis, where correlates in a linear manner. Figures below illustrate the result.

**Figure 13.** Probability Plot of Predictor $X_3$.

**Figure 14.** Scatterplot of cumulative defect quantity, $Y_3$ vs $X_1$, $X_2$, $X_3$, and $X_4$.

**Figure 15.** Anderson Darling Normality Test of Predictor $X_4$. 
From figure above, it is needed to validate multicollinearity among $X_1, X_2, X_3,$ and $X_4$. Utilizing Pearson’s correlation coefficient, the coefficient values are tabulated as below.

| Predictors | $X_1$ | $X_2$ | $X_3$ | $X_4$ |
|------------|-------|-------|-------|-------|
| $X_1$      | 0.989 | 0.989 | 0.998 | 0.998 |
| $X_2$      | 0.989 | 0.970 | 0.985 |       |
| $X_3$      | 0.989 | 0.970 | 0.994 | 0.994 |
| $X_4$      | 0.998 | 0.985 | 0.994 | 0.994 |

It is learned from Table 4 that there is a strong linear connection among variables (predictors). Coefficients that value more than 0.9 is considered strong correlation, hence showing that multicollinearity exists in this model [31]. Therefore, among $X_1, X_2, X_3,$ and $X_4$, only 1 variable is needed to be selected to represent the equation model.

Finalizing the result of multiple linear regressions, the regression equation, which can be used to calculate the response (Cumulative Defect Quantity) until 36-month-in-use, can be written as:

$$Y_1 = -26.3 + 32.6X_3$$

(7)
4.2. Justification of Selected Response
From four available responses (Y₁, Y₂, Y₃, and Y₄), usually, it is decided to select Y₁ to represent the model. Y₁ is the result of months-in-service, calculated by considering the difference between Line-off Month and Repair Month (LO – FD).

Line off date is well explained as the date when a vehicle is out from its assembly line [32]. Vehicle manufacturers and also vendors are both speaking the same language by making this date as a reference. This is made possible when it is used as a form of traceability in terms of 4M (Man, Machine, Material, Method) should there are any quality issues. Based on this justification, using Line-off Date is crucial in this case rather than the Initial Registration Date.

Failure date can be explained as the date of part failure occurrence. Considering this date as the most exact and accurate reference, this contributes to providing the best failure part life-span (months-in-service).

Besides, the actual cumulative defect quantity (Y₁) is also normally distributed. Figure 18 below explains the normality test conducted.

![Figure 18. Anderson Darling Normality Test of Response (Cumulative Defect Quantity, Y₁)](image)

From Table 5 below, correlation coefficients for X₃ records the highest value. Therefore, X₃ is selected to represent the variable and acts as the predictor.

Table 5. Pearson Correlation Coefficients (r) between variables and response, Y₁

| Predictors | X₁ | X₂ | X₃ | X₄ |
|------------|----|----|----|----|
| Y₁         | 0.987 | 0.974 | 0.991 | 0.989 |

4.3. Forecasted Cumulative Defect Quantity Result
Using regression Equation (6), response Y₁(Forecast) is calculated and displayed as per Table 5.

Table 6. Y₁ (Forecast) versus Predictors (Months-in-service)

| Predictors (Months-in-service) | Response (Cumulative Defect Quantity) | Response (Cumulative Defect Quantity) |
|---------------------------------|--------------------------------------|--------------------------------------|
| X₁                              | Y₁ Actual                             | Y₁ Forecast                          |
| 1                               | 2                                    |                                      |
| 2                               | 15                                   |                                      |
| 3                               | 45                                   |                                      |
| 4                               | 70                                   |                                      |
| 5                               | 147                                  |                                      |
| 6                               | 172                                  |                                      |
| 7                               | 196                                  |                                      |
| 8                               | 212                                  |                                      |
| 9                               | 248                                  |                                      |
| 10                              |                                      | 352                                  |
| 11                              |                                      | 385                                  |
As regression equation is derived considering the trends of actual cumulative defect quantity, written as the actual response \( Y_1 \), from 1 month-in-use \( (X_1 = 1) \) until 9 months-in-use \( (X_1 = 9) \), Equation 4 represents the defect trend. Response \( Y_1 \) is calculated where it is forecasted that cumulative defect quantity \( Y_1 \) at \( X_1 = 36 \) during 36-months-in-use is up to 1157pcs.

Forecasted Cumulative Defect Quantity Results are plotted and figures are obtained, where \( Y_1 \) at \( X_1 = 36 \) during 36-months-in-use is up to 1157pcs. To find the anticipated defect quantity, which will occur in the future, is by deducting \( Y_1 \) at \( X_1 = 36 \) with the actual \( Y_1 \) at \( X_1 = 9 \). It is explained below.

\[
Y_{\text{forecast}} = Y_1 \text{ at } X_1 = 36 - Y_1 \text{ at } X_1 = 9
\]
\[
Y_{\text{forecast}} = 1200 - 248 = 952 \text{ defects quantity.}
\]

Figure 19. Regression Result of Cumulative Defect Quantity.

Therefore, the regression equation successfully forecast that there are 952 defects warranty defects anticipated in the future, from January 2019 Line-off Month, by using the regression equation.
5. Conclusion
With the implementation of the MLRA technique, now visualization of warranty budgeting and accrual task, as a concerned topic in most automotive-based company management discussions, is now possible and with a strong base of the mathematical and statistical structure. It is now no longer on a very rough estimation of what potential spend risks where it is taking into account the actual warranty data magnitude. Underestimating warranty cost nor overestimating warranty cost is not an issue anymore.

A clear overview of warranty situation is now visible, that data mining technique has been deployed estimate warranty quantity, results which will lead to warranty cost from available warranty database. It has been utilized and data patterns of defect quantity versus Line-off Date and defect quantity versus repair date have been presented in an appropriate way. Multiple Linear Regression is about finding a linear connection between a dependent variable and several independent variables. The effect of more than one independent variable on response has been examined by defining the predictors and response correctly at the input and desired output, validating the actual data to have fulfilled the assumptions and hypotheses of multiple linear regression. Assumptions and hypotheses are fulfilled and accepted. These are ensured where the predictors and responses are normally distributed and are in linear correlation.

An effective and relatively simple method for estimating the warranty cost equation model successfully has been derived, whereby it took into account, the previous actual warranty data, to represent warranty quantity which will lead to cost impact, taking into consideration, identified parameters that link with pricing and cost.

References

[1] D. Hoyle, Automotive Quality Systems Handbook. First Edition. Oxford: Butterworth-Heinemann Ltd, 2000.
[2] M. U. Thomas, Reliability and warranties: Methods for product development and quality improvement. Florida: CRC Press, 2006.
[3] Verband der Automobilindustrie, Joint Quality Management in the Supply Chain Marketing and Service, Field failure analysis, 1st ed., no. 1. Frankfurt am Main: Henrich Druck + Medien, 2009.
[4] C. Zhou, R. B. Chinnam, and A. Korostelev, “Hazard rate models for early detection of reliability problems using information from warranty databases and upstream supply chain,” Int. J. Prod. Econ., vol. 139, no. 1, pp. 180–195, 2012, doi: 10.1016/j.ijpe.2012.04.007.
[5] W. Shaomin, “Warranty Data Analysis: A Review,” Qual. Reliab. Eng. Int., vol. 28, no. 0748–8017, pp. 795–805, 2012.
[6] B. Ahammed and M. Mesbahul Alam, “Forecasting failure number using warranty claims in multiplicative composite scale,” in International Conference on Statistical Data Mining for Bioinformatics Health Agriculture and Environment, 2012, pp. 540–653.
[7] D. N. P. Murthy and I. Djamaludin, “New product warranty: A literature review,” Int. J. Prod. Econ., vol. 79, no. 3, pp. 231–260, 2002, doi: 10.1016/S0925-5273(02)00153-6.
[8] J. Hipp, U. Günther, and U. Grimmer, “Data Quality Mining -- Making a Virtue of Necessity,” Proc. 6th Acm Sigmod Work. Res. Issues Data Min. Knowl. Discov. (Dmkd 2001, vol. 20, pp. 52–57, 2001, [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.21.2453.
[9] B. K. Rai and N. Singh, RELIABILITY ANALYSIS and PREDICTION with WARRANTY DATA. Issues, Strategies, and Methods. Taylor & Francis Group LLC, 2009.
[10] M. Kurvinen, I. Töyrylä, and D. N. P. Murthy, Warranty Fraud Management: Reducing Fraud and Other Excess Costs in Warranty and Service Operations, 1st ed. John Wiley & Sons, 2016.
[11] S. Limon and O. Yadav, “Forecasting Warranty Claim by Using Past Warranty and Field Usage Data,” in IIE Annual Conference. Proceedings, 2015, p. 1627.
[12] M. F. Darmawan, N. I. Jamahir, R. R. Saedudin, and S. Kasim, “Comparison between ANN and multiple linear regression models for prediction of warranty cost,” Int. J. Integr. Eng., vol. 10, no. 6, pp. 193–196, 2018, doi: 10.30880/ijie.2018.10.06.027.
[13] A. Ferencek and M. Kljajić Borštnar, “Data quality assessment in product failure prediction models,” J. Decis. Syst., vol. 00, no. 00, pp. 1–8, 2020, doi: 10.1080/12460125.2020.1776927.

[14] R. Khoshkangini et al., “Early prediction of quality issues in automotive modern industry,” Inf., vol. 11, no. 71, p. 15, 2020, doi: 10.3390/info11070354.

[15] X. Gang Su and X. Yan, Linear regression analysis: Theory and computing, 1st ed., vol. 140, no. 4. Singapore: World Scientific Publishing Co. Pte. Ltd, 2009.

[16] F. E. Harrell Jr, Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis, 2nd ed. Springer International Publishing Switzerland, 2015.

[17] D. C. Montgomery and G. C. Runger, “Applied Statistics and Probability for Engineers,” Eur. J. Eng. Educ., vol. 19, no. 3, p. 383, 2003, doi: 10.1080/03043799408928333.

[18] J. W. Osborne and E. Waters, “Four Assumptions of Multiple Regression That Researchers Should Always Test. - Practical Assessment, Research & Evaluation,” Pract. Assessment, Res. Eval., vol. 8, no. 2, pp. 1–5, 2002, [Online]. Available: http://pareonline.net/getvn.asp?n=2&v=8.

[19] A. Fitrianto and L. Y. Chin, “Assessing normality for data with different sample sizes using SAS, minitab and R,” ARPN J. Eng. Appl. Sci., vol. 11, no. 18, pp. 10845–10850, 2016.

[20] Z. Ismail, A. Yahya, and A. Shabri, “Forecasting Gold Prices Using Multiple Linear Regression Method Department of Mathematics , Faculty of Science Department of Basic Education Department of Education,” Am. J. Appl. Sci., vol. 6, no. 8, pp. 1509–1514, 2009.

[21] D. Fantazzini and Z. Toktamysova, “Forecasting German Car Sales Using Google Data and Multivariate Models,” Int. J. Prod. Econ., vol. 1, no. 170, pp. 97–135, 2015, [Online]. Available: https://mpra.ub.uni-muenchen.de/67110/1/MPRA_paper_67110.pdf.

[22] P. W. Seelbach, L. C. Hinz, M. J. Wiley, and A. R. Cooper, “Use of multiple linear regression to estimate flow regimes for all rivers across Illinois, Michigan, and Wisconsin,” Fish. Res. Rep, vol. 2095, pp. 1–35, 2011.

[23] Perusahaan Otomobil Nasional Berhad, “Compass.proton.com.” http://compass.proton.com/irj/portal (accessed Jan. 06, 2020).

[24] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, “From Data Mining to Knowledge Discovery in Databases,” in AI magazine, 1996, vol. 17, no. 3, pp. 37–54.

[25] U. H. Nosakhare and A. F. Bright, “Evaluation of Techniques for Univariate Normality Test Using Monte Carlo Simulation,” Am. J. Theor. Appl. Stat. Spec. Issue Stat. Distrib. Model. Appl. Math., vol. 6, no. 5–1, pp. 51–61, 2017, doi: 10.11648/j.ajtas.s.2017060501.18.

[26] E. G. M. Goh, Learn R for applied statistics, with data visualizations, regressions, and statistics. Singapore: Apress Media LLC, 2019.

[27] C.-F. Lee, J. Lee, J.-R. Chang, and T. Tai, Essentials of Excel, Excel VBA, SAS, and Minitab for statistical and financial analysis, 1st ed. Cham: Springer International Publishing Switzerland, 2016.

[28] S. A. Darestani, Y. Ismail, and A. Napsiah, “Recognition of Nonrandom Patterns on Supply Performance by Employing Statistical Monitoring,” J. Am. Sci., vol. 6, no. 4, pp. 114–122, 2010.

[29] H. C. Thode Jr, Testing for normality, 1st ed. Basel: Marcel Dekker Inc, 2002.

[30] H. Xu and Y. Deng, “Dependent Evidence Combination Based on Shearman Coefficient and Pearson Coefficient,” IEEE Access, vol. 6, pp. 11634–11640, 2017, doi: 10.1109/ACCESS.2017.2783320.

[31] P. Schober, C. Boer, and L. A. Schwarte, “Correlation coefficients: Appropriate use and interpretation,” Anesth. Analg., vol. 126, no. 5, pp. 1763–1768. 2018, doi: 10.1213/ANE.0000000000002864.

[32] M. E. Bergen, P. Van Beek, and T. Carchrae, “Constraint-based vehicle assembly line sequencing,” InConference Can. Soc. Comput. Stud. Intell., vol. 2056, pp. 88–99, 2001, doi: 10.1007/3-540-45153-6_9.