# Regression Models in Lubricant Oil Quality Assessment

Sandra M. Aleixo[1,2(✉)], Ricardo Enguiça[2,3], Francisco Sedas[1], Paulo Ferreira[4], and André Silva[4]

[1] Department of Mathematics, ISEL – Instituto Superior de Engenharia de Lisboa, IPL – Instituto Politécnico de Lisboa, Rua Conselheiro Emídio Navarro, 1, 1959-007 Lisboa, Portugal
sandra.aleixo@isel.pt

[2] CEAUL – Centro de Estatística e Aplicações da Universidade de lisboa, Faculdade de Ciências da Universidade de Lisboa, Bloco C6, Piso 4, 1749-016 Lisboa, Portugal
ricardo.roque@isel.pt

[3] CEMAT Ciências – Centro de Matemática Computacional e Estocástica - Instituto Superior Técnico, Av. Rovisco Pais, 1049-001 Lisboa, Portugal

[4] Galp, Quality & Technical Product Care, Avenida da Índia, 8, 1349-065 Lisboa, Portugal
{paulo.miguel.ferreira,andre.c.silva}@galp.com

**Abstract.** After a suitable data analysis, we adjust two types of regression models to support decision-making when monitoring lubricants in an industrial environment. The target variable is the remaining useful life of a lubricant, as it is an essential process in preventing damage to associated equipment. The main objective of this work is to predict at an early stage how long will an oil load still perform correctly.

The first model is a multiple linear regression, adjusting coefficients using cross-validation on several subgroups of the data set. Though the performance of this model is satisfactory, the corresponding adjustment is compromised by the fact that the observations are not independent, as they are grouped taking into account the oil loads to which they belonged. For this reason, a second type of regression was adjusted: a linear regression model with mixed effects. After adjusting the models, a *Python* algorithm was created to determine an optimal interval regarding the oil service hours after which the lubricant should be replaced.

**Keywords:** Predictive maintenance · Lubricants · Predictor selection · Correlation · Multiple linear regression model · Mixed effects linear regression · Stepwise selection methods · Optimal range

## 1 Introduction

Within the scope of a particular project, the Quality & Technical Product Care Department of the main Portuguese energy company (Galp) intends to carry out predictive maintenance of industrial equipment, in particular through a thorough data analysis and a corresponding assembly of a model to support decision-making when monitoring lubricants in service.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2024
O. Gervasi et al. (Eds.): ICCSA 2024 Workshops, LNCS 14815, pp. 393–411, 2024.
https://doi.org/10.1007/978-3-031-65154-0_24

Evaluating the remaining useful life of a lubricant is essential to prevent damage in the equipment. It is important to provide the customer with a forecast of the condition of the lubricant so that damage does not occur. This can be accomplished by means of data modeling of the lubricants in service.

In what follows we make a brief review of some literature on this subject and also on the techniques that were applied.

In [8], Pinheiro *et al.* perform several multiple linear regressions on the available predictors for assessing lubricant oil properties, where it was shown that there is no overall superior regression method and the choice is dependent on the predicted property. Also on the maintenance of lubricants framework, Wakiro *et al.* [9] develop a maintenance decision support system that automatically quantifies the influence of lubricant parameters on the unhealthy state of the machine using alternative classifiers. The classifiers are compared for robustness from which decision-makers select an appropriate classifier given a specific lubricant data set. In our framework, we want to use a reduced set of explanatory variables and keep the most relevant ones through all models. Note that we used a stepwise selection method to choose the best explanatory variables.

Another aspect that we enhance in our work is the determination of the confidence interval of service hours where we get the minimum wrong diagnosis, that is, getting a warning to change oil when it is still in good condition (false positive) or a tag considering that the oil is functional when it should be changed (false negative). An interesting applied study on this matter can be found in [7], where Malo *et al.* screen several chemical compounds to identify candidate "hits" accurately. This work also has its focus on finding hits' thresholds and the importance of minimizing false-positive and false-negative rates.

Recently, machine learning techniques are getting a much broader attention in all domains, and our framework is not an exception. For example, in the paper by Malaguti *et al.* [5], some of these tools are used to analyse lubricant oil data, enabling better information in the condition based maintenance approach to the process of diagnosing and predicting failures in diesel-powered vehicle fleets. The results presented in this analysis show that to explain 90% of the variation in the original data only a few number of variables are needed, just as is our case. After evaluating the quality of the variables, an experimental study to analyse the performance of various machine learning algorithms was conducted, taking into account the number of features as input data. In [3], we can find an application of machine learning tools to calibrate real-time condition monitoring of lubricating oil. The methods used are ensembles of regression trees. Ensembles are combinations of models that often are able to improve the results of individual models. Several ensemble methods were studied and the best results were obtained with Rotation Forests. In [6], Mota *et al.* explore Predictive Maintenance systems as an enabler to predict a machine's condition/lifespan, allowing for maintenance-effective manufacturing. This paper aims to address these issues by proposing a novel methodology to improve this performance, by proposing a machine learning training methodology, an automatic hyperparameter optimizer, and a retraining strategy for real-time application. To validate the proposed methodology a random forest and an artificial neural network model are

implemented as well as explored. A synthetic dataset, that replicates industrial machine data, was used to show the robustness of the proposed methodology. In our approach we opted not to consider advanced machine learning methods, namely neural networks in order to obtain an explicit model, which is harder to find using those tools.

For a generic reference on Regression Diagnostics, we mention the book by Belsley *et al.* [2], where we can find good practices to assess quality and reliability of regression estimates. This book emphasizes diagnostics and includes suggestions for remedial action. Another reference for the impact of different undesirable scenarios is the paper by Afreixo *et al.* [1], where issues like collinearity, Simpson's paradox, variable interaction or Freedman's paradox are analysed in order to find criteria for adequate feature selection and coefficients' estimation, namely through traditional methodologies, such as automatic selection (e.g., stepwise using Akaike criterion and Bayesian criterion) and penalized regression (e.g., least absolute shrinkage and selection operator (LASSO), elastic net, relaxed and adaptive LASSO, minimax concave penalty and smoothly clipped absolute deviation penalty, penalized regression with second-generation p-values).

In this work, we perform two different regression models applied to a data set, using the accumulated service hours of the lubricating oil as the response variable. First, we considered a multiple linear regression model, where to estimate coefficients we applied cross-validation. Given that the observations are not independent (since they can naturally bee grouped taking into account the oil load to which they belonged), in a second approach we opted to adjust the data to a linear regression model with mixed effects.

Comparing the results of the two models, we conclude that the linear regression model with mixed effects seems to be more appropriate, due to the relevance given to the grouping structure present in the data.

On the subject of regression models mixed linear effects, we indicate the reference book by Galecki and Burzykowski [4].

This paper is organised in the following way: in Sect. 2 we explore the considered data set, in Sect. 3 we detail the associate models, in Sect. 4 we built the optimal intervals that minimize misclassified observations and we end by stating the main conclusions in Sect. 5.

## 2 Data Set

Galp's Quality & Technical Product Care department deals with issues related to quality, technical assistance and services associated to sold products, whether fuels, chemicals or lubricants.

This work focuses on monitoring lubricants, with the primary objective being to make a decision regarding the time interval in which oil renewal should occur, through the analysis of laboratory data. Thus, data samples were collected from the equipment and sent to the laboratory, where they were classified according to the recorded values in each of the physicochemical tests.

Each test corresponds to the results for several independent variables (of a total of $m$), and the results for each of these variables can be seen as a vector represented by $\mathbf{X_j}$, corresponding to one of the columns in the data matrix $\mathbf{X}$. The last column will correspond to the response variable, $y$, which is the number of service hours the oil has been in use. Each of the analysed samples corresponds to a line in the matrix, thus being considered as a single observation, of a total of $n$.

The included explanatory variables result from laboratory analysis, carried out on lubricant samples, and can be grouped into three categories: variables related to the physicochemical state, variables resulting from the wear of a metal and variables linked to another type of contamination.

The variables in the first group are those that constitute the physicochemical composition of the lubricant. Throughout the useful life of the oil, these values are compared with the initial condition of the oil. The considered variables are:

- Oxy, the generic term oxidation is a chemical reaction of the lubricant with oxygen. Oxidation corrosion results in serious effects on both the oil and the equipment's engine. In this context, oxidation is a measure of the byproducts of this reaction in the oil, whose unit of measurement is Abs/cm.
- pH, a dimensionless numerical scale used to specify the acidity or basicity of a substance. Over time, the lubricant degrades, making the pH more acidic.
- Sul, sulfation is a chemical reaction of the lubricant with sulfur. This reaction results in acidic compounds. This variable is measured in Abs/cm.
- TAN, the total acid number evaluates the amount of acidic compounds present in the lubricating oil. This variable is measured in $\mathrm{mg}_{KOH}/\mathrm{g}$.
- TBN, the total basic number evaluates the amount of alkaline reserve present in the lubricating oil. This variable is measured in $\mathrm{mg}_{KOH}/\mathrm{g}$.
- Ca, the calcium content indicates the concentration of Calcium present in the lubricating oil. This variable is measured in ppm.
- P, the phosphorus content indicates the concentration of phosphorus present in the lubricating oil. This variable is measured in ppm.
- S, the sulphur content indicates the concentration of Sulfur present in the lubricating oil. This variable is measured in ppm.
- Zn, the zinc content indicates the concentration of Zinc present in the lubricating oil. This variable is measured in ppm.
- $\mathrm{Vis}_{40}$ and $\mathrm{Vis}_{100}$, the viscosity at 40 °C and at 100 °C, reflect the dispersion of the lubricant in a given area over a given period of time. They are both measured in cSt.

The second group of variables reflect oil contamination due to wear:

- PQ Index, an indirect measure of the concentration of ferromagnetic metals in the lubricating oil, capturing particles from 14 microns. This variable is dimensionless;
- Al, aluminium content that measures the concentration of this metal in the lubricating oil, and a high value may indicate that there is wear on pistons, bearings, pumps and rotors. This variable is measured in ppm;

- Cu, copper content that measures the concentration of this element in the lubricating oil. This variable is also measured in ppm;
- Fe, iron content that measures the concentration of ferrous material in the lubricating oil. This variable is also measured in ppm.

The last group of explanatory variables concern external contamination:

- Na, sodium content that frequently originates from the refrigerant. A high value of this variable may indicate leaks in the cooling system. This variable is measured in ppm.
- Si, silicon content that can come from dust, the wear of the rubber that seals the oil compartments and the gas used in the gas engine. This variable is also measured in ppm.

The response or dependent variable is the one that we indent to be explained by all previously described. Within the scope of the project developed, this variable will be the number of hours that the lubricating oil is at service, $h_{oil}$.

The lubricating oil in service is only removed and, consequently completely renewed, when after a sample analysis, it is diagnosed as being in danger. This means that observations corresponding to the same oil load follow an increasing function in the variable $h_{oil}$, which is only interrupted when the next observation corresponds to another load. When this happens, it is possible to verify that the next observation will have a lower oil service hours value. This lubricant total lifetime can be influenced by several factors, namely the quality of the lubricant, the environment outside the engine, or the operation of the equipment itself. Therefore, the periodic analysis of lubricant samples in service is due to the variability of these factors, allowing us to determine when the lubricant should be replaced.

We also considered a non-numerical information resulting from laboratory analysis, which is the diagnosis category. This categorical variable is represented by $C_i$ for the $i^{th}$ observation and admits three possible states: "Normal", "Watch" or "Risk". This variable will later be associated to the number of hours of service of a given load, being this association the one that reduces most wrong diagnosis.

The data set on which we based our model comes from a biogas motor generator. A biogas motor generator consists of an internal combustion engine that generates heat and electrical energy and, as in any other type of engine, the correct functioning of the lubrication system plays a critical role, for example in reducing friction between parts and protecting against corrosion.

The data set involves 172 observations, analysed between April 2018 and December 2021. As stated before, since new oil charges are always topped up and when considered obsolete are fully removed, the majority of sequential observations belong to the same lubricant charge.

## 3 Modeling Approaches

In a first and more simplistic approach, a multiple linear regression model was adjusted using the entire amount of samples. This was made with the awareness

that it is not the most appropriate model, due to the fact that observations belonging to the same oil load are not independent. However, in order to mitigate effects of dependence in observations, coefficients for this regression were estimated using a cross-validation technique.

In a second approach, a linear regression model with mixed effects was adjusted, accommodating observations dependent on each oil load, and also removing observations that belonged to loads with two or fewer observations (in order to avoid the performance of a regression with probable outliers, that did not represent the standard population of oil loads).

All the results stated ahead were obtained through algorithms developed in *Python* programming language and using appropriate available packages.

### 3.1 Multiple Linear Regression Model

In a first phase, a multiple linear regression model was adjusted. The first model must include independent explanatory variables that contribute to the best explanation of the response variable.

An analysis of correlations between all variables (explanatory and response) was initially carried out (using Pearson correlation coefficient) (see the matrix in Fig. 1). Below the main diagonal we have the sample correlation values between each pair of variables and the corresponding p-values associated to the significance test of the population correlation coefficient (p-values are presented with two decimal places).

Significantly high sample correlation coefficient values (greater than 0.7) were identified. These correspond to p-values lower than the 5% significance level, leading to the rejection of the respective null hypothesis. For each pair of variables under these conditions, that is, for each two linearly correlated explanatory variables, the one with a lower correlation coefficient value with the response variable is removed.

After the execution of this process, only five explanatory variables remained, namely Sulfation, Aluminium Content, Silicon Content, Viscosity at 100 °C and Viscosity at 40 °C.

The first model was adjusted following two different procedures. In the first, variable selection was made through hypothesis tests for each model's coefficients, while in the second, a stepwise variable selection algorithm was used.

For the first case, the regression model was adjusted with the five explanatory variables determined previously, obtaining a globally significant model, since the p-value of the F test is practically null. Subsequently, those variables whose p-value relative to the significance test of the respective coefficient was higher than the significance level of 5% were excluded, implying the exclusion of two more variables: Al content and Viscosity at 40 °C.

Alternatively, using the stepwise method with the Both Forward-Backward algorithm, the variables that better explained the response variable were selected among the five remaining ones, optimizing the number of variables. The algorithm returned the same three previously found by the other selection method.

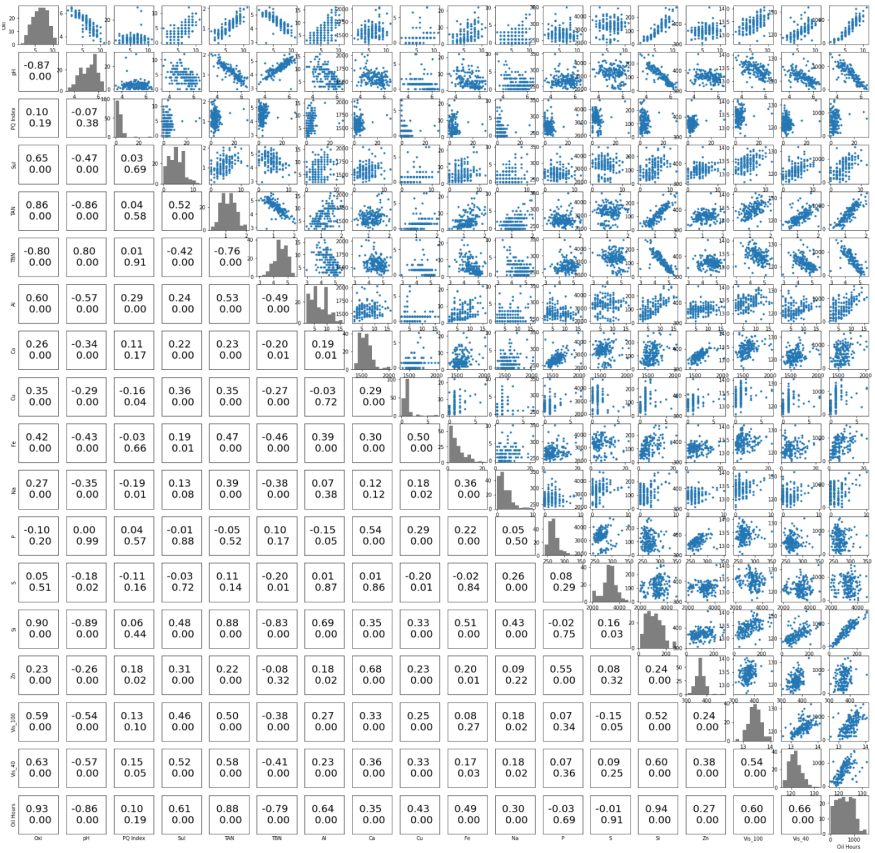

**Fig. 1.** Correlation matrix.

Therefore, the following explanatory variables were set as predictors for the multiple linear regression model: Sulfation, Silicon Content and Viscosity at 100 °C.

To estimate the coefficients we used cross-validation. The initial sample was divided into five groups and five multiple regressions were carried out, considering data from four of these groups, allowing a random effect in the calculation of these coefficients. Each coefficient was calculated as the arithmetic mean of the coefficients obtained in the five regressions. The estimated model equation is the following:

$$\hat{h}_{oil} = -2150.996 + 4.975Sul + 24.574Si + 157.606Vis_{100}$$

For each coefficient, the obtained value represents the change in the number of oil service hours per unit of the corresponding independent variable, keeping the values of the other independent variables constant. For example, if Sulfation increases 1 Abs/cm, the number of hours that the lubricating oil remains

in service increases about 5 h (considering that the values Silicon Content and Viscosity at 100 °C remain constant).

To validate the estimated model, an analysis of residuals and of the existence of influential observations was carried out.

For the adjusted model to be valid, the residuals must meet three assumptions: independence, homoscedasticity and normality. Through Fig. 2, a brief graphical analysis of the residuals can be made.

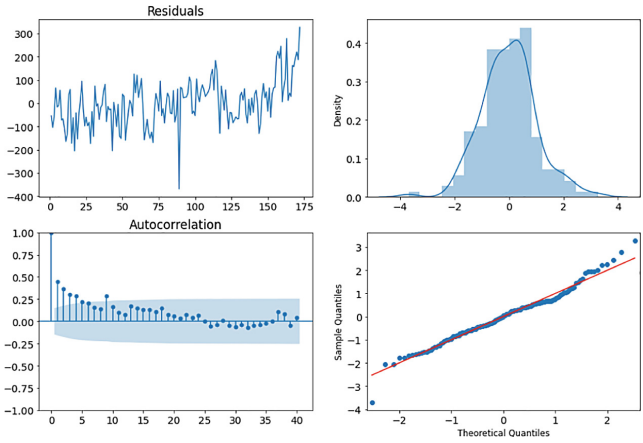


**Fig. 2.** Graphs for Residuals Analysis.

Concerning residuals independence, given that all observations come from the same equipment and that many of them belong to the same oil load, it is known a priori that this assumption will not be verified. If the observations are dependent, then the residuals must be dependent as well. When analysing the correlogram, in the bottom left corner of Fig. 2, it can be seen that not all points are within the blue band, which is an indicator of correlation between residuals. This graphical analysis can be validated by performing the Durbin-Watson test, whose null hypothesis states that the residuals are independent. The test value obtained was 1.04, so the fact that it is below 1.5 means that the null hypothesis is rejected, and therefore the existence of positive autocorrelation between the residuals can be assumed.

On the residuals homoscedasticity assumption, we need to see if the residuals variance is constant.

Through the scatter diagram, it can only be concluded that this variance is constant if the points are dispersed randomly, without presenting any type of pattern. Additionally, to analytically confirm what is observed in the graph, the Breusch-Pagan homoscedasticity hypothesis test can also be carried out. In Fig. 3 we have the Residuals vs Fitted graph, and it can be seen that an increase in residual variability is observed for higher adjusted values. Performing the Breusch-Pagan hypothesis test, a p-value of 0.65 was obtained, which is higher than the significance level of 5%, so we can conclude that the variance of residuals appears to be constant.

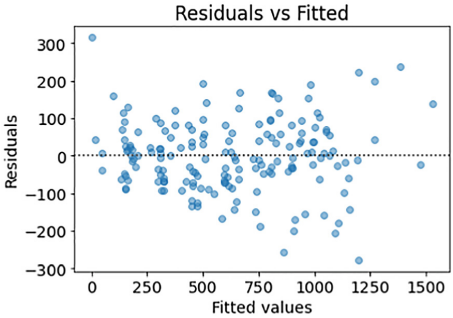


**Fig. 3.** Dispersion diagram between residuals and adjusted values.

To check the normality of residuals, we initially carry out a graphical analysis of the QQ-Plot, in the bottom right corner of Fig. 2. It seems reasonable to conclude that the residuals follow a normal distribution, as most of the blue points are on the red diagonal line. This analysis can be complemented by analysing the shape of the histogram and the respective frequency polygon, that resembles the normal density curve (in the top right corner of Fig. 2). To complement this analysis, Shapiro-Wilk hypothesis test was carried out to evaluate the hypothesis that residuals come from a population with normal distribution. For this test, a p-value of 0.036 was obtained, which is below the significance level of 5%. Therefore, the null hypothesis is rejected, concluding that the residuals do not appear to come from a normal distribution. (Note that if the significance level considered was 1%, normality would already be verified).

With this, we can only conclude that the homoscedasticity assumption is verified. Regarding independence, given the presence of samples from the same oil cargo, it was known from the beginning that it would not be verified. The non-validation of the assumption of normality would not invalidate the model in its whole if it was the only assumption not verified, which was not the case.

In the Residuals vs Fitted graph in Fig. 3, it is possible to see two values with absolute value close to 300, and, therefore, potentially consisting in influential observations.

Observing the graph of Residuals vs Leverage in Fig. 4, it is possible to see that all points are within the dashed lines. Combining the information on the two

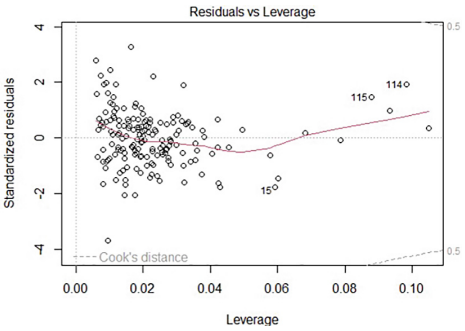


**Fig. 4.** Scatter diagram between residuals and fitted values.

graphs, the existence of two influential observations can be assumed, even if the Cook distance is not very high for them. The details of these observations are stated in Table 1.

**Table 1.** Influential observations noted.

| Index | Sulfation | Silicon Content | Viscosity to 100 °C | Oil Hours |
|---|---|---|---|---|
| 89 | 3 | 70 | 13.31 | 0 |
| 172 | 4 | 154 | 13.68 | 1196 |

These observations were consequently removed and the multiple linear regression model was readjusted to the data, without these observations, following the same procedure, and the following formula was obtained:

$$\hat{h}_{oil} = -2004.012 + 4.930Sul + 25.183Si + 146.869Vis_{100}$$

For the first adjusted linear multiple regression model, the $R_a^2$ value is 0.913, meaning a good adjustment, explaining about 91.3% of the response variable. The corresponding AIC coefficient is 2078. For the adjusted model without influential observations, $R_a^2$ improved to 92.3% and the AIC to 2029, enabling us to conclude that the quality of this new model is slightly better.

In Fig. 5, we find a representation of the estimated values and real observations. The resemblance is evident, confirming that the model is well adjusted.

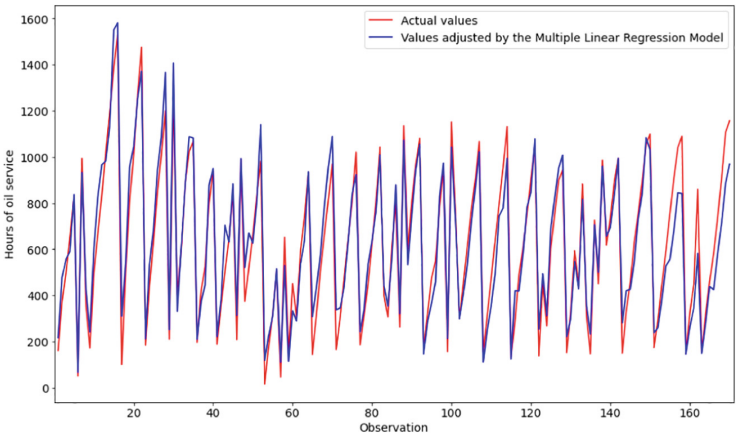


**Fig. 5.** Comparison between the values estimated by the model and the real values.

In terms of residuals analysis, the assumption of independence remains unverified. However, the assumption of normality of residuals was already been verified (the p-value for the Shapiro-Wilk test is 0.133).

This improved Multiple Linear Regression Model is well adjusted, however, it is not the most suitable for this data set due to the non independence of observations. Therefore, we proceed to find another model, that better fits the properties of our data: the Multiple Linear Regression Model with Mixed Effects.

### 3.2 Multiple Linear Regression Model with Mixed Effects

By the reasons already stated above, it is clear that observations are dependent within the same oil load, so, in order to built an alternative model, we opted to consider groups of observations taking into account the lubricating oil load to which they belong. To improve the statistical relevance of the data set, it was decided to remove groups with a total of observations smaller than three, resulting in a slight decrease in the total number of observations. Furthermore, influential observations previously identified and removed were taken into account, none of which influenced the removal of groups. The final data set consists of 144 observations, split into 26 groups (we kept the original id's of each group in the original set), corresponding to different oil loads.

Mixed-effects linear regression models can include random intercepts, random slopes, or both, so, in the following we will settle which is more appropriate.

We started with a procedure analogous to that previously described. Through the correlation matrix and the respective tests on the correlation coefficients, the uncorrelated variables were chosen to enter the model, which were the same: Sulfation, Aluminium Content, Silicon Content, Viscosity at 100 °C and Viscosity at 40 °C. After that, using variable selection methods, three more variables were removed: Aluminium Content, Viscosity at 100 °C and Viscosity at 40 °C. So, the model will only include two variables: Sulfation and Silicon Content.

In order to understand whether it makes sense to introduce random intersections, we analysed the boxplots relating to the response variable, taking into account the group to which observations belong, which are represented in Fig. 6.

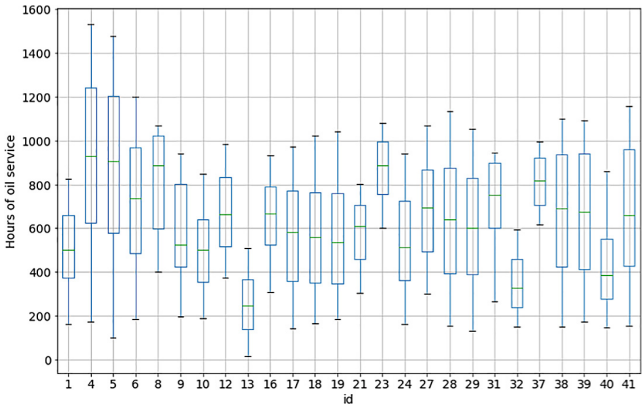


**Fig. 6.** Boxplots of the response variable grouped by id.

It can be seen that there is a clear difference in the number of oil service hours between groups: some are around 1000 h, while others around 300 h. Due to these differences, it makes sense to insert random intersections for each group, meaning that the mixed model will estimate a different random intersection for each group. For each group we will consider an equation of this type:

$$h_{oil} = \beta_0 + \beta_1 Sul + \beta_2 Si + (u_0)_{id=i} + \epsilon$$

where $i = 1, \cdots, 26$ and $(u_0)_{id=i}$ represents the random intersection for $i^{th}$ group. To investigate the possibility of introducing random slopes, we analyse the graphs containing simple linear regressions between a predictor and the output variable, for each of the groups. In Fig. 7, the left graph refers to simple linear regressions between the variable Silicon Content and the output variable for each of the groups; on the right we have the same for the Sulfation variable.
For Silicon Content, all regressions present similar slopes, so there will be no need to apply a random element to the coefficient of this variable. For Sulfation, it can be seen that some regressions' slope are quite different from the majority, but in general most regressions are similar. Therefore, it was considered that the increase in model complexity inherent in the estimation of many more parameters if random slopes were introduced is not justifiable and, therefore, the considered model for each group $i$ is:

$$\hat{h}_{oil} = -106.116 + 12.946 Sul + 5.999 Si + (\hat{u_0})_{id=i}$$

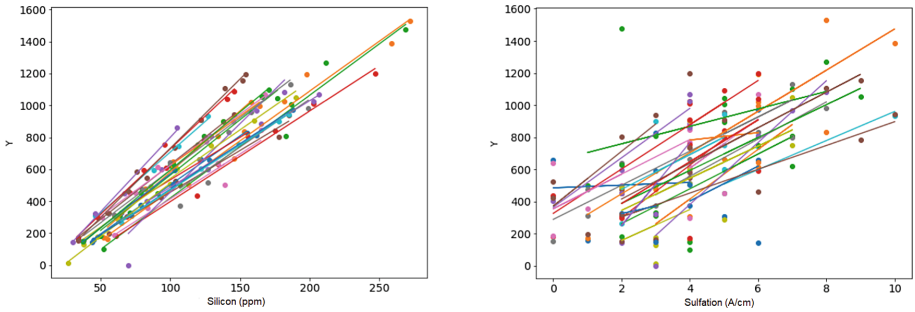


**Fig. 7.** Simple regressions between a predictor and the output variable for each group.

The range of random intersection values for the linear mixed model vary from −139.01 in group 6 to 152.19 in group 39. For example, for the group with *id* 1 of the first oil load, whose random intercept value is -18.24, the equation would be:

$$\hat{h}_{oil} = -124.356 + 12.946Sul + 5.999Si$$

To validate the estimated model, an analysis of the residuals and the existence of influential observations was carried out.

As already mentioned, the model will have to verify three assumptions. Contrary to what was previously observed in the adjustment of the multiple linear regression model, it is expected that the assumption of independence will be verified. When analysing the correlogram, in the lower left corner of Fig. 8, it can be seen that only the first and third points are not within the blue band, with the third point being very close to the limit. This means that there appears to be no correlation between the residuals. This interpretation can be validated by using the Durbin-Watson test. An observed test statistic value of 1.91 was obtained, which is in the interval [1.5, 2.5], so the null hypothesis is not rejected and it can be assumed that there is no autocorrelation between the residuals.
To graphically verify the assumption of homoscedasticity of the residuals, we proceed to analyze the Residuals vs Fitted graph, presented in Fig. 9.
As can be seen, the points are dispersed randomly without, apparently, presenting any type of pattern. Thus, the variance of the residuals appears to be constant. To confirm this suspicion, the Breusch-Pagan hypothesis test was carried out, obtaining a p-value of 0.22, which is higher than the significance level of 5%, so that the variance of the residuals appears to be, in fact, constant.

To assess the normality of the residuals, we analyze the QQ-Plot, in the bottom right corner of Fig. 8, and it seems possible to conclude that the residuals come from a population with normal distribution, because most of the blue points are on the red line. To analytically confirm this suspicion, the Shapiro-Wilk hypothesis test was carried out, obtaining a p-value of 0.57, which is higher than the significance level of 5%. In this way, the normality of residuals is confirmed.

It is therefore concluded that all assumptions were verified when adjusting the multiple linear regression model with mixed effects, contrary to what happened when adjusting the multiple linear regression model.

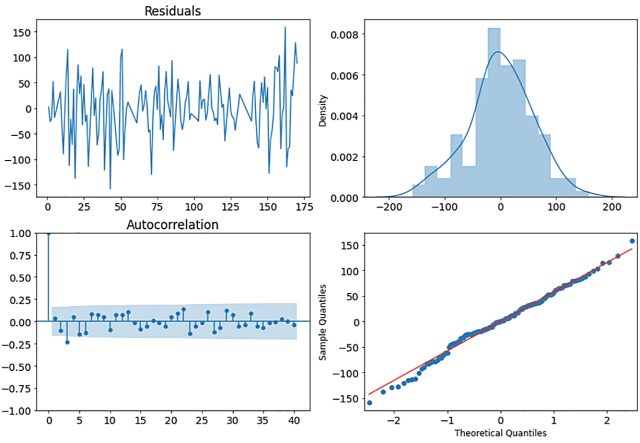


**Fig. 8.** Graphs for residuals analysis.

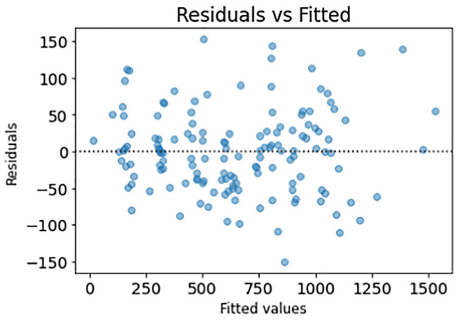


**Fig. 9.** Scatter diagram between residuals and fitted values.

Analysing the Residuals vs Fitted graph in Fig. 9 there do not appear to be any values that differ significantly from the others. Therefore, there do not appear to be any influential observations.

Through the quotient between the variance of the random intercept and the sum between this and the error variance, it is possible to evaluate the quality of fit of the model. The value of this indicator must be in the range [0,1], with a value closer to 1 meaning that the estimation of the random effect for the group is relevant. In this specific case, the following result was obtained:

$$\frac{7158.44}{7158.44 + 3990.80} = 0.642\,,$$

and given that it is larger than 0.5, the grouping of data taking into account the oil load can be considered significant.

For this mixed effects model with random intercepts, the AIC value is 1674, which can be assumed to be a better option compared to the multiple linear

regression model, where the AIC is 2029. Therefore, this model is better suited to this data than the multiple linear regression model without mixed effects.

Another way to evaluate the quality of this type of model with mixed effects is through two determination coefficients. A first one, the Adjusted Marginal Coefficient of Determination, which concerns the proportion of variance that is explained by the fixed effects. This coefficient is the quotient between the variance of the fixed effects and the total variance of the model (sum of the variances of the fixed effects, the mixed effects and the error), which in this case is given by:

$$\frac{113779.16}{124928.41} = 0.911\,.$$

The second is the Conditional Adjusted Coefficient of Determination, which concerns the proportion of variance that is explained by both fixed effects and mixed effects, that is, the quotient between the sum of the variances of the fixed and mixed effects and the total variance of the model, being for the data under study given by:

$$\frac{113779.16 + 7158.44}{124928.41} = 0.968\,.$$

Given the high values of both coefficients (above 0.9), it can be considered that the model is well adjusted.

For the multiple linear regression with random intersections (MLRRI), the results were much better when compared to the results of multiple linear regression (MLR). The main validation and quality measures of the models are presented in Table 2.

**Table 2.** Linear Regression Model Comparison Table.

| Validation and Quality Indicators | MLR | MLRRI |
|---|---|---|
| Number of Predictors | 3 | 2 |
| $R_a^2$ | 92.3 % | 91.1 % (Marginal) |
| AIC | 2029 | 1674 |
| Residuals Independence | No | Yes |
| Residuals Normality, Residuals Homoscedasticity | Yes | Yes |

## 4 Optimal Intervals for Lubricant Replacement Based on Regression Results

In this section we find an optimal confidence interval, through which it is possible to infer from the total service hours when to renew the lubricating oil, avoiding the performance of extra testing.

Note that the correct diagnoses are only guaranteed by laboratory analysis, and that our goal is to be able to provide an estimate for how long should a load remain on a machine using preliminary laboratory analysis values.

Although the multiple linear regression model is not the most suitable for adjusting this type of data, we opted to consider it as the basis to find the optimal interval for renewing the lubricating. Later we will also consider a version using the mixed effects model.

The first step consisted in representing the total number of anomalous diagnoses if the classification was defined by a single threshold value concerning the total number of hours at service of a given oil load.

Anomalous diagnosis can be split in diagnoses considered Normal above an Upper Control Limit (UCL) and diagnoses labelled as Watch or Risk but below the Lower Control Limit (LCL). If $UCL = LCL$, Fig. 10 shows us the corresponding numbers of misclassifications.

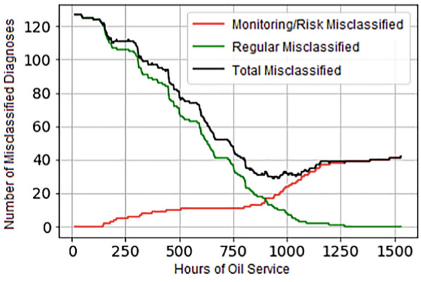


**Fig. 10.** Graphic Representation of the Number of Anomalous Diagnoses

The minimum of total misclassifications is located at the 932 h mark (for a total of 29 misclassified diagnoses). This value is the natural choice for our threshold, but since it is important to give the customer some room for maneuver, we consider a 95% confidence interval centered at it. To calculate the limits of the confidence interval it is necessary to have a vector with the values of the explanatory variables.

An algorithm was created in *Python*, which checks all possible combinations of integer values of the explanatory variables and returns the one whose value of the response variable is closer to a target value. For example, considering a target value of 932 h, we get a Sulfation value of 11, a Silicon Content of 152 and a Viscosity at 100 °C equal to 13.

Having the values of the explanatory variables, it is now possible to determine the limits of the 95% confidence interval for the 932 h forecast: $[726, 1136]$. Using this interval, the number of anomalous diagnoses is 13, with 2 Normal diagnoses above the upper limit of the range, 9 Watch diagnoses below the lower limit of the range, and 2 Risk diagnoses below the lower limit of the range.

In the next step, we allow for a shift of this interval, maintaining its amplitude and search for the minimum number of anomalous diagnoses. On the considered

case, if we add 59 hours to our interval, we avoid one extra misclassification, getting the optimal interval $[785, 1195]$.

The last step consists in reusing the algorithm to find the predictors integer values for the new center of the interval.

To obtain the optimal range of oil service hours using linear regression with mixed effects, the optimization algorithm underwent some changes.

The grouping of observations must be taken into consideration when setting the interval for a future observation.

As an example, in order to explain the changes made to the optimization algorithm, we consider a sample with 299 hours of service, with a Sulfation level of 5 and Silicon content equal to 41.

This is a sample that was taken from a very "young" oil, which makes it a suitable example as you can compare it to the available data for subsequent samples of the same load.

Unlike the multiple linear regression case, it is not necessary to find values for the predictors. However, the first step is to understand which of the 26 groups should the interval will be calculated for. To do this, with the associated model and the values of the new observation, the random intersection value is determined. For observation mentioned above, a value of 94.42 was obtained. Then, the absolute difference between this value and that of the random intersections of all groups was calculated and the one whose difference was minimal considered. This guarantees that if the new observation is inserted into the group, its random intersection would not change significantly. In this case, the association was with group 16, which contains only 4 observations.

This first step is the main difference in the algorithm, as the following steps are similar to the previous case.

The associated threshold of hours that minimizes anomalous diagnoses is 932 h, where no anomalous diagnoses were recorded. Then, as had already been done, the combination of integer values to be given to the predictors was calculated, so that the model responded with a value close to 932 h. For Sulfation, the value was 9 and for Silicon Content, the value was 126.

With these values, the 95% confidence interval is $[807, 1020]$, and there was no need to implement the optimization algorithm for this interval, since if it had zero amplitude there would not be any anomalous diagnoses, therefore, located at the same value with a greater amplitude, the number of anomalous diagnoses will also be zero.

**Comparison with the Real Sample.** A good methodology for validating the range will be to compare it with the real sample. As mentioned, this observation belongs to a data set with 24 observations grouped into 4 groups. This way we can compare the interval with the later observations of the oil load corresponding to this observation.

When analyzing the Table 3, it appears that the last two samples already indicated a possible degradation of the lubricating oil. However, it was only replaced

**Table 3.** Oil Load to which the observation belongs.

| Oil Service Hours | Status |
|---|---|
| 299, 452, 595, 750 | Normal |
| 906 | Risk |
| 1094 | Watch |

at 1094 h. However, the calculated interval defines a maximum value for renewal of 1020 h, with the central value located at 913 h.

## 5 Conclusions and Future Work

Conclusions regarding the relevance of each of the considered models were stated in previous sections, but we enhance that the dependence of observations between samples from the same load implied that a mixed effects multiple linear regression model with random intersections was the most appropriate for our framework.

Obviously, a larger dataset, not necessarily for oil cargoes always associated to the same machine could allow us to draw better conclusion for our mixed effects regression model, eventually with a clearer influence of random slopes, instead of only on random intersections. Somehow we feel that different values for the intersection can be associated to an oil that degrades at a similar pace, but starts working with different level of quality. On another hand, random slopes can be related to a degradation speed, maybe associated to the lubricant's purpose on different machines, or even to different types of oil. We find that there is still room to improve results by applying these type of models.

The determination of the appropriate thresholds for the categorical classification (Normal, Watch and Risk) can also follow the mixed effects version, by assuming different intervals depending on the first analysis on the corresponding load. We provided a grounding work on this subject, but we feel that improvements can be made in a future work.

**Acknowledgements.** Research partially funded by FCT - Fundação para a Ciência e a Tecnologia, Portugal, under the projects UIDB/00006/2020, DOI: 10.54499/UIDB/00006/2020 (CEAUL), UIDB/04621/2020, DOI: 10.54499/UIDB/04621/2020 (CEMAT) and ISEL.

## References

1. Afreixo, V., Cabral, J., Macedo, P.: Comparison of feature selection methods in regression modeling: a simulation study. In: Gervasi, O., et al. (eds.) ICCSA 2023. LNCS, vol. 14112, pp. 150–159. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-37129-5_13
2. Belsley, D.A., Kuh, E., Welsch, R.E.: Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. Wiley, Hoboken (2005)

3. Bustillo, A., Villar, A., Gorritxategi, E., Ferreiro, S., Rodríguez, J.J.: Using ensembles of regression trees to monitor lubricating oil quality. In: Mehrotra, K.G., Mohan, C.K., Oh, J.C., Varshney, P.K., Ali, M. (eds.) IEA/AIE 2011. LNCS (LNAI), vol. 6703, pp. 199–206. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-21822-4_21
4. Galecki, A., Burzykowski, T.: Linear mixed-effects model. In: Galecki, A., Burzykowski, T. (eds.) Linear Mixed-Effects Models Using R. Springer Texts in Statistics, pp. 245–273. Springer, New York (2013). https://doi.org/10.1007/978-1-4614-3900-4_13
5. Malaguti, R., Lourenço, N., Silva, C.: A supervised machine learning model for determining lubricant oil operating conditions. Expert Syst. **40**(5), e13116 (2023). https://doi.org/10.1111/exsy.13116
6. Mota, B., Faria, P., Ramos, C.: Machine learning applied to industrial machines for an efficient maintenance strategy: a predictive maintenance approach. In: Jorgensen, B.N., da Silva, L.C.P., Ma, Z. (eds.) Energy Informatics. LNCS, vol. 14467. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-48649-4_17
7. Malo, N., Hanley, J., Cerquozzi, S., et al.: Statistical practice in high-throughput screening data analysis. Nat. Biotechnol. **24**, 167–175 (2006). https://doi.org/10.1038/nbt1186
8. Pinheiro, C., Rendall, R., Quina, M., Reis, M., Gando-Ferreira, L.: Assessment and prediction of lubricant oil properties using infrared spectroscopy and advanced predictive analytics. Energy Fuels **31**(1), 179–187 (2017). https://doi.org/10.1021/acs.energyfuels.6b01958
9. Wakiru, J., Pintelon, L., Muchiri, P., Chemweno, P.: A data mining approach for lubricant-based fault diagnosis. J. Qual. Maint. Eng. **27**(2), 264–291 (2021). https://doi.org/10.1108/JQME-03-2018-0027
10. Yu, J.: A hybrid feature selection scheme and self-organizing map model for machine health assessment. Appl. Soft Comput. **11**(5), 4041–4054 (2011). https://doi.org/10.1016/j.asoc.2011.03.026. ISSN 1568-4946