Chapter 11

Application of Statistical Processing Techniques to the Impedance-based SHM for the Oil & Gas Industry

Chapter details

Chapter DOI:

https://doi.org/10.4322/978-65-86503-88-3.c11

Chapter suggested citation / reference style:

Silva, José W., et al. (2022). "Application of Statistical Processing Techniques to the Impedance-based SHM for the Oil & Gas Industry". *In* Jorge, Ariosto B., et al. (Eds.) *Uncertainty Modeling: Fundamental Concepts and Models*, Vol. III, UnB, Brasilia, DF, Brazil, pp. 337–383. Book series in Discrete Models, Inverse Methods, & Uncertainty Modeling in Structural Integrity.

P.S.: DOI may be included at the end of citation, for completeness.

Book details

Book: Uncertainty Modeling: Fundamental Concepts and Models

Edited by: Jorge, Ariosto B., Anflor, Carla T. M., Gomes, Guilherme F., & Carneiro, Sergio H. S.

Volume III of Book Series in:

Discrete Models, Inverse Methods, & Uncertainty Modeling in Structural Integrity

Published by: UnB City: Brasilia, DF, Brazil Year: 2022

DOI: https://doi.org/10.4322/978-65-86503-88-3

Application of Statistical Processing Techniques to the Impedance-based SHM for the Oil & Gas Industry

José Waldemar da Silva[1*], Quintiliano S. S. Nomelini[1],
Diogo de Souza Rabelo[2], José dos Reis V. Moura Jr[3],
Roberto M. Finzi Neto[4], Carlos A. Gallo[4] and Julio E. Ramos[5]

[1]Faculty of Mathematics, Federal University of Uberlândia, Brazil. E-mail: zewaldemar@ufu.br; quintiliano.nomelini@ufu.br
[2]Post-Graduate Program – Industrial Engineering, Fed. University of Goias, School of Sciences and Technology, Aparecida de Goiania, Brazil. e-mail: diogo.rabelo@ufg.br
[3]Post-Graduate Program – Modeling and Optimization, Fed. University of Catalao, Brazil. e-mail: zereis@ufcat.edu.br
[4]Post-Graduate Program – Mechanical Engineering, Federal University of Uberlandia, Brazil. e-mail: finzi@ufu.br; gallo@ufu.br
[5]Petrobras, Petroleo Brasileiro S.A., RD Center (CENPES), Brazil. e-mail: julio.ramos@petrobras.com.br

*Corresponding author

Abstract

This chapter presents methodologies for preprocessing electromechanical impedance-based structural health monitoring data sets with the objective of error reduction when inferring changes in the structure. Atypical signatures may not represent the actual state of the structure and should be excluded from the inference step. The impedance signature trend in the frequency domain is another aspect that must be considered because divergences between the samples regarding this feature can also lead to mistakes. Polynomial regression models for matching the effect of temperature on impedance signatures are also discussed in this chapter, and methodologies to infer the state of the structure from the CCD metric.

Keywords: Impedance-based structural health monitoring; Statistical preprocessing tools; Regression models

1 Data preprocessing in SHM

Structural health monitoring can be performed from the electromechanical impedance signature, and a set of signatures acquired under the same conditions are expected to be similar. In other words, repeatability of signatures is expected.

The monitored structure may be located in environments where it is impossible to control variables whose effect may interfere with the electromechanical impedance signature, reducing the degree of repeatability among measurements. The repeatability requirement may not be met due to atypical signatures. Abrupt actions resulting from a falling object, somebody walking, or a bird landing is an uncontrolled interference that can affect the electromechanical impedance signature.

The verification of outliers in the damage metrics, calculated within the same group of signature repetitions, can help detect atypical signatures, which should be removed from the database for further analysis. The procedures used for the necessary data cleansing will be called pre-processing (Section 1).

Further analyses, in this chapter, in particular, are defined as the processing (Section 2) of the baseline data for the determination of polynomial models for temperature compensation and the post-processing (Section 3) as the use of data to investigate the integrity of the structure.

After collecting a set of signatures representing the same state of the structure, the median signature can be taken as a reference signature. In this way, it is possible to calculate some damage metrics, such as the CCD (Correlation Coefficient Deviation), which in this situation will depict the degree of variability or similarity between the group's signatures. The median signature is chosen as a reference due to the generally asymmetrical character of the electromechanical impedance distribution at each frequency.

The correlation coefficient is defined as in Eq. 1

$$CC_{Z_1 Z_2} = \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{\left[Z_{1,i} - \bar{Z}_1\right] \left[Z_{2,i} - \bar{Z}_2\right]}{S_{Z_1} S_{Z_2}} \right\} \tag{1}$$

where $Z_{1,i}$ and $Z_{2,i}$ are the real parts of the electromechanical impedance of the two signals involved in calculating the metric, at frequency i, with $i = 1, 2, \cdots, n$ and n is the number of frequency points used to data acquisition. To calculate the correlation coefficient within the same group of signature repetitions, you can define Z_1 as any signature and Z_2 as the median signature. The averages of each sample set are represented respectively by $\bar{Z}_1$ and $\bar{Z}_2$ and, S_{Z_1} and S_{Z_2} are the standard deviations of them. The metric is then determined by Eq. 2.

$$CCD = 1 - CC_{Z_1, Z_2}. \tag{2}$$

Example 1. *Consider the 16 impedance signatures as presented in Table 7 (adapted data), measured from a real experiment developed at the Structural Mechanics Lab of the Mechanical Engineering School of the Federal University of Uberlandia. Calculating the CCD values associated with each measurement using the R software:*

Solution

The developed code is shown in the following image (Figure 1), whose output is the CCD values rounded to 4 decimal places.

```
> med=vector()
>  for(i in 1:nrow(data)){
+    med[i]=median(data[i,])} # median signature calculation
> ccd=vector()
> for(i in 1:ncol(data))
+    {
+     ccd[i]= 1-sum( (data[,i] - mean(data[,i]))*(med - mean(med))/
+                    (sd(data[,i]) * sd(med)) ) /nrow(data)
+       } # Eqs. 1 and 2
> round(ccd,digits=4)
 [1] 0.0202 0.0202 0.0201 0.0202 0.0201 0.0201 0.0201 0.0201
 [9] 0.0201 0.0202 0.0201 0.0223 0.0209 0.0209 0.0208 0.0207
>
```

Figure 1: Calculating CCD in R language.

It is assumed that outliers values of CCD will be obtained from atypical or discordant signatures about the others in the group. Therefore, the corresponding signatures must be retaken from the database for further analysis. There are some methods of identifying outliers, among them, the Tukey method [Tukey, 1977], better known as boxplot, and the method de Chauvenet [Reddy, 2011]. These two methods will be presented in subsections 1.1 and 1.2 to follow.

1.1 Tukey Method - Boxplot

The box plot or BoxPlot is defined by a rectangle in which, if presented vertically, the lower and upper sides coincide with the first (q_1) and the third (q_3) quartiles, respectively. The inner dash is the median value (q_2). The distance between q_1 and q_3 is the interquartile range, $IIQ = q_3 - q_1$. This interval contains 50% of the central observations, and its amplitude, as well as the full amplitude of the Boxplot, contains information about the dispersion of the data.

A value located between $q_1 - 1.5 \times IIQ$ and $q_1 - 3 \times IIQ$ or between $q_3 + 1.5 \times IIQ$ and $q_3 + 3 \times IIQ$ is called outlier. A point beyond $q_3 + 3 \times IIQ$ or short of $q_1 - 3 \times IIQ$ is called an extreme value. In this text we will not make this distinction and extreme values will also be called outlier (Figure 4).

Example 2. *Consider the CCD values presented in the example solution 1 and check the existence of outliers values using Tukey's method.*

Solution

Getting $q_1 = \dfrac{x_{(4)} + x_{(5)}}{2} = 0.0201$ *and* $q_3 = \dfrac{x_{(12)} + x_{(13)}}{2} = 0.02075$*, values greater than* $0.02075 + 1.5 \times (0.02075 - 0.0201) = 0.0217$ *or less than* $0.0201 - 1.5 \times (0.02075 - 0.0201) = 0.0191$ *are classified as outliers. The index in parentheses,* $x_{(i)}$*, indicates the value of* x *in the i-th position after all values are sorted in ascending order. In this example, only the value* 0.0223 *is classified as outliers.*

The descriptive analysis presented in the Example 2 can be performed in the Software R [R Core Team, 2021]. However, a different methodology for determining the quartiles. There are several methods for estimating quartiles [Langford, 2006], and they may differ in the way of determining their position and their calculation itself.

The same data explored in the Example 2 were used to illustrate the obtaining of outliers from the standard methodology implemented in the R software (Figure 2).

```
> boxplot(ccd)$out
[1] 0.0223
>
```

Figure 2: Calculating outlier in R language.

The outlier point, as well as the behavior of the data distribution, are shown in Figure 3.

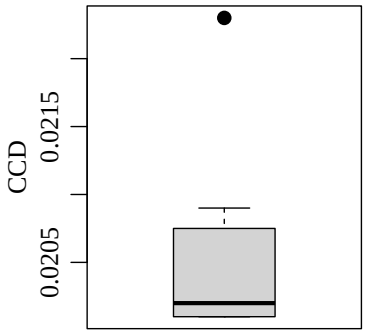

Figure 3: Boxplot of the outlier as well as the CCD distribution.

Based on this criterion, the twelfth signature (CCD=0.0223) may have been acquired under the effect of some atypical event and therefore does not represent the actual status of the structure and thus must be removed from the database for future analysis.

Another example of detecting *outliers* at the two extremes of the data distribution, from a set of data arising from the observation of any variable Y, is shown in Figure 4.

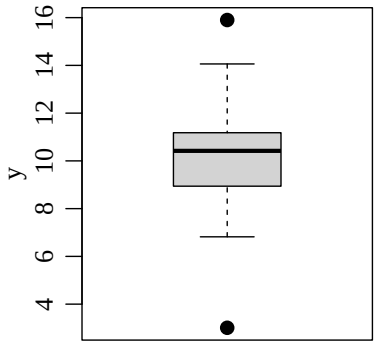

Figure 4: Boxplot of the outliers at the extreme right (top) and left (bottom) of the data distribution.

The values of outliers shown in Figure 4, as well as the positions of these values in the data list, can be obtained with the R commands shown in the following image (Figure 5)

```
> y=c(10.23,  7.39, 10.84,  8.82,  6.82, 15.90, 10.86, 8.94,
+     10.42, 10.04, 11.10,7.47,  11.87, 14.06, 13.43, 10.92,
+     11.18, 12.48, 10.11, 10.42, 3.01,  9.29)
> boxplot(y)$out
[1] 15.90  3.01
> match(boxplot(y)$out,y)
[1]  6 21
```

Figure 5: Identification of outlier values and their respective positions in the data vector.

The position of an outlier damage metric in the list of values makes it easier to identify the atypical impedance signature associated with that value.

1.2 Chauvenet's method

This method is based on the number of standard deviations z (Eq. 3) that a value, (y) is far from the mean $(\bar{y})$.

$$z = \frac{y - \bar{y}}{s} \Longrightarrow (y - \bar{y}) = zs \tag{3}$$

If z is greater than the maximum allowable deviation (d_{max}) or less than the minimum permissible deviation (d_{min}), then the corresponding value y is considered an outlier.

The thresholds d_{max} and d_{min} are determined from the normal distribution. Therefore, the hypothesis of normality of the variable (Y), from which measurements or observations are made, cannot be rejected. This is a necessary assumption for applying the Chauvenet criterion [Taylor, 2012].

According to this criterion, a value is considered non-typical if the corresponding number of deviations from the mean z belongs to a range of values, symmetrically delimited by d_{min} and d_{max} around the mean, with probability $1-1/2n$ where n is the number of elements in the sample.

Since z, presented in Eq. 3 is a standardized score, the critical values of the Chauvenet test (d_{max} and d_{min}) are determined from the resolution of the integral as presented in Eq. 4.

$$P(-d_{max} < Z < d_{max}) = \int_{-d_{max}}^{d_{max}} \frac{1}{\sqrt{2\pi}} exp\left\{\frac{z^2}{2}\right\} dz = 1 - \frac{1}{2n} \tag{4}$$

Figure 6 shows the area corresponding to the probability presented in Eq. 4, for any n.

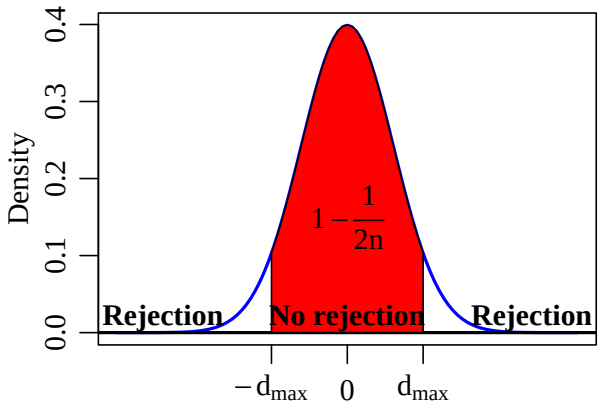

Figure 6: Illustration of non-rejection and rejection data regions using the Chauvenet criterion.

To eliminate the need to determine two thresholds, d_{max} and d_{min}, just use the absolute value of the difference in Eq. 3, that is $z = \frac{|x - \bar{x}|}{s}$ and x will be outliers if $z > d_{max}$. This strategy is justified by the symmetry property of the normal distribution.

Failure to comply with the assumption of normality may result in not-so-promising results. Therefore, in this case, other techniques for detecting outliers should be used, such as Tukey's (Section 1.1).

In practical terms, the critical value d_{max} can be obtained from some software that returns quantiles of the normal distribution for the probability $1 - \frac{1}{2n}$ (Eq. 4).

Example 3. *Consider the dataset presented in Figure 5 and check for the presence of outlier values using the Chauvenet method. If they exist, identify their positions in the dataset.*

Solution

```
> chauvenet=function(x){
+   n=length(x)
+   prob=1-1/(2*n)
+   q1=qnorm((1-prob)/2)
+   q2=abs(q1)
+   z=(x-mean(x))/sd(x)
+   out=y[z<q1 | z>q2]
+   c(out)
+ }
> data=c(10.23,  7.39, 10.84,  8.82, 6.82, 15.90, 10.86,  8.94,
+        10.42, 10.04, 11.10, 7.47, 11.87, 14.06, 13.43, 10.92,
+        11.18, 12.48, 10.11, 10.42, 3.01, 9.29)
> out=chauvenet(data);out
[1] 3.01
> match(out,y) # outliers position
[1] 21
```

Figure 7: R source code to identify outliers and respective positions.

Using the chauvenet function presented in Figure 7 can be obtained the value 3.01. It is classified as an outlier by the Chauvenet method. Match function indicates that this value occupies the 21^{a} position in the dataset.

It is recommended that the Chauvenet criterion be applied only once (Reddy [2011];Taylor [2012]).

1.3 Trend Analysis

The trending effect on signatures can lead to erroneous inferences about the state of the structure or temperature compensation polynomials if this effect is not the same in all signatures. Therefore, verifying its presence and removal is necessary for further processing or inference.

It is suggested that if trend correction is necessary for some impedance signatures, the same procedure is performed in the others to match them.

The significance of the exponential model (Eq. 5), or its correspondent in linearized form (Eq. 6), for some signature, indicates that the trend characteristic is present and must be adjusted.

The significance of the regression model can be investigated through the analysis of variance. Also, it can be done by verification of the null hypothesis for the parameter associated with the independent regressor variable, in the case of simple linear regression.

The determination coefficient R^2 is another measure that helps in the decision regarding the suitability of the model.

The methodology and basic concepts necessary for regression analysis are presented in the 1.4 section.

1.4 Regression Model for Trend Adjustment

Assuming that the trend effect is exponential, the model presented in Eq. 5 can be used. In this model, impedance is considered a function of frequency:

$$y_i = \beta_0 e^{\beta_1 x_i} \times \epsilon_i; \quad i = 1, 2, \cdots, n, \tag{5}$$

where y_i and x_i are respectively the impedance and frequency values obtained in the *i-th* observation to which the error ϵ_i is also associated; n is the frequency sample at which impedances were collected.

Assuming that the model error described in Eq. 5 is multiplicative, then this model can be approached in its linearized form (Eq. 6). Linearization is advantageous since, in this case, the system of ordinary equations associated with the model has an analytical solution. It, therefore, does not depend on the convergence of algorithms to determine the estimates of the coefficients as in the nonlinear case.

$$ln(y_i) = ln(\beta_0) + \beta_1 x_i + ln(\epsilon_i) \tag{6}$$

If the proposed model is indeed effective then the trend-free signature will be obtained by $y_i - exp(\widehat{ln(y_i)})$.

To present some theoretical aspects of the simple linear regression model, consider two variables y and x where y is a variable that linearly depends on x. The statistical model of a simple linear regression is shown in Eq. 7.

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \tag{7}$$

where y_i: represents the i-th observed value; x_i: the independent variable, i = 1, 2,..., n; ϵ_i : is the unobservable error associated with i-th observation; β_0 and β_1: are the intercept and the slope, parameters of the model.

In a linear regression model, the following assumptions must be met:

-
- The x values are fixed, that is, known before the observation of y and therefore x is not a random variable;
- The mean error is zero, ie $E[\epsilon_i] = 0, \forall i = 1, 2, \ldots, n$;
- For a given value of x, the error variance is always constant, that is, $V[\epsilon_i] = E[\epsilon_i^2] - (E[\epsilon_i])^2 = E[\epsilon_i^2] = \sigma^2, \forall i = 1, 2, \ldots, n$, The error is then said to be homoscedastic;
- The error of one observation is uncorrelated with the error of another observation (the errors are independent), ie $E[\epsilon_i \epsilon_j] = 0$, for $i \neq j$;
- The error has a Normal distribution with zero mean and constant variance (σ^2), that is, $\epsilon_i \sim N(0, \sigma^2)$. Concluding, the errors are independent and identically distributed, that is, $\epsilon \sim iidN(0, \sigma^2)$.

This last item assembles the necessary assumptions so that the inference on the model parameters can be made from the parametric point of view, the ordinary least-squares method. On the other hand, if these assumptions are not met, the inference must be made by alternative means such as the bootstrap method Efron and Tibshirani [1993].

A dependent variable can be modeled as a function of a single regressor variable (Eq. 7), by several regressors (Eq. 8)

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i, \tag{8}$$

where x_1, x_2, $\cdots$, x_k represent the k explanatory variables considered. Yet another possibility is the modeling of y by a single regressor variable but in polynomial form (Eq. 9) and in this case, k represents the degree of the polynomial

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_k x_i^k + \epsilon_i. \tag{9}$$

The model presented in Eq. 7 is a particular case of both the model presented in Eq. 8 and Eq. 9. It should be noted that other variations are possible as multiple linear regression models, in addition to the linear terms, quadratic, cubic terms, etc., are considered for some or some regressors in addition to interactions between them. The real need for these terms and even for the variables in the modeling can be verified from the techniques of selection of variables.

The models presented in Eqs. 7, 9 and 8 can be written in matrix form as in Eq. 11 since for each i, there is a line as shown in Eq. 10.

$$\underbrace{\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}}_{\boldsymbol{Y}} = \underbrace{\begin{pmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1k} \\ 1 & x_{21} & x_{22} & \ldots & x_{2k} \\ \vdots & \vdots & \vdots & \ldots & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{nk} \end{pmatrix}}_{\boldsymbol{X}} \underbrace{\begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}}_{\boldsymbol{\beta}} + \underbrace{\begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}}_{\boldsymbol{\epsilon}} \tag{10}$$

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{11}$$

When $k = 1$ in Eq. 10 it means a simple linear regression model (Eq. 7) and when the columns of the matrix $\boldsymbol{X}$ correspond to the quadratic terms x^2, x^3, etc, from this same variable x, represents the polynomial model.

1.4.1 Least-squares estimators

The least-squares estimator for $\boldsymbol{\beta}$ is determined by minimizing the sum of squared errors, presented in Eq. 12 which can be interpreted as a function of $\boldsymbol{\beta}$, which will be named $Q(\boldsymbol{\beta})$.

$$\begin{aligned} Q(\boldsymbol{\beta}) &= \boldsymbol{\epsilon'}\boldsymbol{\epsilon} = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \\ &= \mathbf{Y}'\mathbf{Y} - \boldsymbol{Y'X}\boldsymbol{\beta} - \boldsymbol{\beta'X'Y} + \boldsymbol{\beta'X'X}\boldsymbol{\beta}. \end{aligned} \tag{12}$$

Note that $\boldsymbol{Y'X}\boldsymbol{\beta}$ and $\boldsymbol{\beta'X'Y}$ are scalar and therefore equal. Then Eq 12 can be presented as in Then Eq 13,

$$Q(\boldsymbol{\beta}) = \mathbf{Y}'\mathbf{Y} - 2\boldsymbol{\beta'X'Y} + \boldsymbol{\beta'X'X}\boldsymbol{\beta}. \tag{13}$$

Differentiating $Q(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$ is obtained:

$$\frac{\partial Q(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -2\boldsymbol{X'Y} + 2\boldsymbol{X'X}\boldsymbol{\beta} \tag{14}$$

and equating to Eq. 14 set to zero, specific values are admitted for the vector $\boldsymbol{\beta}$, allocated in a vector $\boldsymbol{b}$, which satisfies Eq. 15

$$-2\boldsymbol{X'Y} + 2\boldsymbol{X'X}\boldsymbol{b} = 0, \tag{15}$$

and therefore the least-squares estimator of $\boldsymbol{\beta}$ is as shown in Eq. 16

$$\boldsymbol{b} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'Y}. \tag{16}$$

Determining the regression equation or the estimation of regression coefficients does not reflect any information about the significance of the model, that is, how important the estimated model is to explain the response variable. The regression analysis of variance allows this study and is presented in the 1.4.2 section below.

1.4.2 Analysis of variance for regression

The analysis of variance for regression consists of verifying how much variation of the dependent variable is explained by the model adopted. The total variance quantified through the total sum of the square, $SSTotal$, can be written as the sum of the variance explained by the regression model, $SSReg$, with the unexplained or residual variance, $SSRes$.

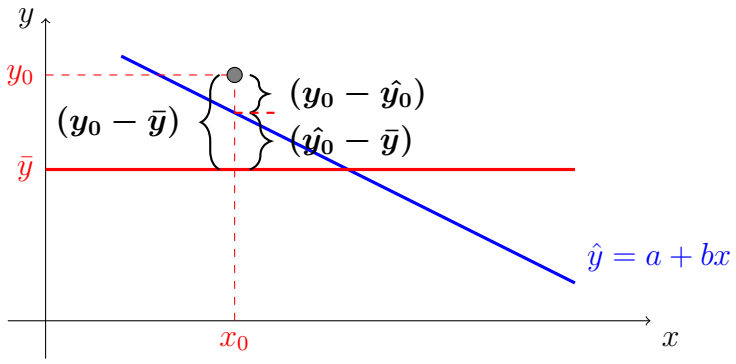

Figure 8: Decomposition of deviations from the mean in a simple linear regression.

Figure 8 illustrates the decomposition of the variation in $\boldsymbol{y}$ considering a simple linear regression. Note, from this Figure, that $(y_0 - \bar{y}) = (y_0 - \hat{y_0}) + (\hat{y_0} - \bar{y})$ and, generalizing, for any pair (x_i, y_i) obtains,

$$(y_i - \bar{y}) = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y}) \tag{17}$$

It is possible to show that

$$SSTotal = SSReg + SSError$$

or, that

$$\underbrace{\sum_{i=1}^{n}(y_i - \bar{y})^2}_{SSTotal} = \underbrace{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}_{SSReg} + \underbrace{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}_{SSError}. \tag{18}$$

To show the validity of the equality in Eq. 18, apply the sum to the square of the deviations, term to the right of the equality in Eq. 17, after subtracting and conveniently adding $\hat{Y}_i$, that is,

$$\begin{aligned} SSTotal &= \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2 = \sum_{i=1}^{n}((y_i - \hat{y}_i) + (\hat{y}_i - \bar{y}))^2 \\ &= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + 2\underbrace{\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \bar{y})}_{=0} + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \end{aligned} \tag{19}$$

and verify that

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = 0. \tag{20}$$

The validity of Eq. 20 can be shown by noting that

$$\sum_{i=1}^{n}(y_i-\hat{y}_i)(\hat{y}_i-\bar{y})=\underbrace{\sum_{i=1}^{n}(y_i-\hat{y}_i)\hat{y}_i}_{I}+\bar{y}\underbrace{\sum_{i=1}^{n}(y_i-\hat{y}_i)}_{II}=0. \tag{21}$$

Writing I in Eq. 21 in a matrix form obtains

$$\begin{aligned} I &= \sum_{i=1}^{n}(y_i-\hat{y}_i)\hat{y}_i=(\boldsymbol{Y}-\boldsymbol{Xb})'\boldsymbol{Xb}=\boldsymbol{Y}'\boldsymbol{Xb}-\boldsymbol{b}'\boldsymbol{X}'\boldsymbol{Xb}=\boldsymbol{Y}'\boldsymbol{Xb}-\left[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}\right]'\boldsymbol{X}'\boldsymbol{Xb} \\ &= \boldsymbol{Y}'\boldsymbol{Xb}-\boldsymbol{Y}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Xb}=\boldsymbol{Y}'\boldsymbol{Xb}-\boldsymbol{Y}'\boldsymbol{Xb}=0 \end{aligned}$$

and further, knowing that the least-squares method consists of minimizing, that is, determining the values β_s that minimize the sum of squares in Eq. 22,

$$\sum_{i=1}^{n}(y_i-\hat{y}_i)=\sum_{i=1}^{n}(Y_i-[\beta_0+\beta_1x_{1i}+\beta_2x_{2i}+\cdots+\beta_kx_{ki}])^2 \tag{22}$$

applying this method will lead to a system with $k+1$ equations and $k+1$ unknowns in which the first equation is obtained by differentiating the given function in Eq. 22 and setting the result to zero. So the derivative concerning β_0

$$\begin{aligned} &\frac{\partial\sum_{i=1}^{n}(y_i-[\beta_0+\beta_1x_{1i}+\beta_2x_{2i}+\cdots+\beta_kx_{ki}])^2}{\partial\beta_0} \\ =\ &-2\sum_{i=1}^{n}(y_i-[\beta_0+\beta_1x_{1i}+\beta_2x_{2i}+\cdots+\beta_kx_{ki}])^2 \end{aligned} \tag{23}$$

and equating to zero, specific values are allowed for β_0, β_1, β_2, $\cdots$, β_k, which b_0, b_1, b_2, $\cdots$, b_k satisfy Eq. 24, below.

$$\sum_{i=1}^{n}(y_i-[b_0+b_1x_{1i}+b_2x_{2i}+\cdots+b_kx_{ki}])=0. \tag{24}$$

then,

$$II=\sum_{i=1}^{n}(y_i-\hat{y}_i)=0 \tag{25}$$

because,

$$II=\sum_{i=1}^{n}(y_i-\hat{y}_i)=\sum_{i=1}^{n}(y_i-[b_0+b_1x_{1i}+b_2x_{2i}+\cdots+b_kx_{ki}])=0. \tag{26}$$

From the results presented in Eqs. 22 and 26 it follows that the sum in Eq. 20 is in fact zero and therefore the total change in Y, quantified by $SSTotal$ is decomposed into the model-explained change, $SSReg$, and the residual change, $SSError$.

Under the assumption that the errors are independent and identically distributed according to the normal distribution with zero mean and constant variance, σ^2, the F test of the analysis of variance for the linear regression model (Table 1) allows testing the hypothesis that the variance explained by the model, with p coefficients, is significant. This hypothesis is equivalent to the idea that the slope parameter of the line, or slope of the line, β_1 is zero, in the case of simple linear regression, or that the coefficients of the explanatory variables are all null as shown in Eq. 27. The alternative hypothesis is that at least one of these coefficients differs from zero.

Table 1: Analysis of variance for the complete model

Source	DF	SS	MS	F-Value
Regression	$p-1$	$SSReg$	$MSReg$	$MSReg/MSError$
Error	$n-p$	$SSError$	$MSError$	
Total	$n-1$	$SSTotal$		

The hypotheses verified from the analysis of variance can be described as follows,

$$\begin{cases} H_0 : \beta_0 = \beta_1 = \cdots = 0 \\ H_a : \beta_j \neq 0 \quad \text{para pelo menos um } j, \text{ com } j = 0, 1, 2, \cdots, k \end{cases} \tag{27}$$

Thus, H_0 is rejected, at the α level of significance, if $F_0 \geq F_{\alpha,p-1,n-p}$ where $F_{\alpha,p-1,n-p}$ is a quantile of the *F* distribution of *Snedecor* with $p-1$ and $n-p$ degrees of freedom such that $P(F > F_{\alpha,p-1,n-p}) = \alpha$. The values $MSReg$ and $MSError$ are obtained by doing $SSReg/(p-1)$ and $SSError/(n-p)$, respectively.

1.4.3 Goodness of fit

As mentioned earlier, parametric inference on a regression model requires the fulfillment of some assumptions such as normality, independence, and homoscedasticity of errors. The lack of knowledge of the population model makes it necessary to consider the evidence for rejection or non-rejection of these assumptions. This evidence is obtained from the sample data and summarized in the estimated model. The decision on whether or not to reject the assumptions is made based on appropriate hypothesis tests. When any of these assumptions are not met, alternative methodologies must be used, such as data transformations in cases of non-normality and homogeneity or dependence modeling in cases of non-independence, or even non-parametric methodologies such as the bootstrap technique.

The assumptions necessary for the validity of the parametric inference from models estimated via the ordinary least-squares method are:

- **Normality of errors:** errors must be normally distributed. Normality is required so that hypothesis tests based on the normal, t, chi-square, and F distribution are unbiased. There are some normality test methods such as Kolmogorov-Smirnov, Lilliefors, Shapiro-Wilk Royston [1995];
- **Independence of errors or residuals:** the errors must be independent, that is, the probability of occurrence of any error must not depend on other errors. There are some tests to

verify the independence hypothesis, among which the Durbin-Watson Durbin [1969] test, which is widely used;

- **Homogeneity of errors:** The lack of homogeneity can interfere with the hypothesis tests (F test and t-test) and the quality of the estimates of the model coefficients. In the presence of heterogeneity, least squares estimators may not have a minimum variance compared to other estimators, and therefore they are no longer efficient. To verify the hypothesis of homoscedasticity, the Breusch-Pagan [Breusch and Pagan, 1979] or Levene [Fox and Weisberg, 2018] tests can be used.

The Shapiro-Wilk, Durbin-Watson and Breusch-Pagan tests are implemented in R software through the functions *shapiro.test*, *dwtest* and *bptest*. These last two are in the *lmtest* [Zeileis and Hothorn, 2002] package.

The Shapiro-Wilk test is limited to samples with sizes between 3 and 5000, but there are several other tests in the literature with the same purpose [Khatun et al., 2021]. Among them is the Anderson-Darling test [Thode, 2002b] that can be applied from the function *ad.test*, implemented in the package *nortest* [Gross and Ligges, 2015], of the software R.

Graphical methods involving the residuals can also be used to verify the adequacy of the model. Some graphs and their usefulness are presented below.

- Residuals versus adjusted/predicted values:
  - It may indicate poor specification of the model;
  - Evaluate whether the errors have constant variance;
  - Identify *outliers*.

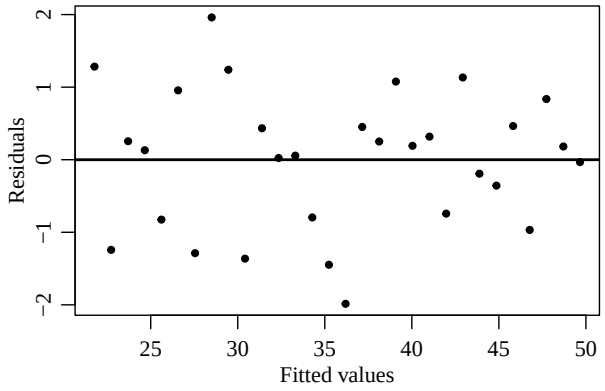

Figure 9: Residuals versus adjusted values.

- Q-Q plot:
  - Check if the errors are approximately normally distributed;
  - Identify *outliers*.

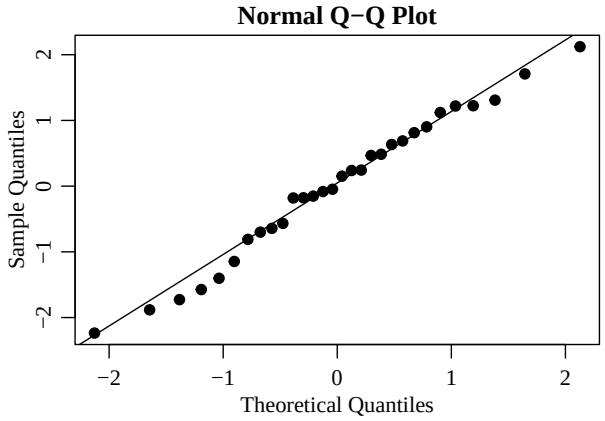

Figure 10: Q-Q plot for residuals.

- Residuals versus collection order:
  - Check the induced correlation by the collection order;

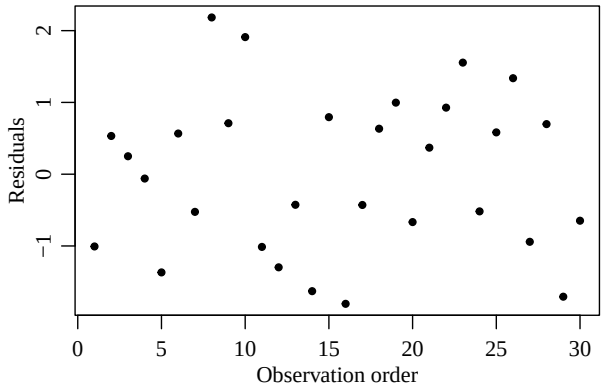

Figure 11: Residuals versus observation collection order.

- Residuals versus omitted variable in the model:
  - Any systematic pattern indicates the need to incorporate the variable into the model.

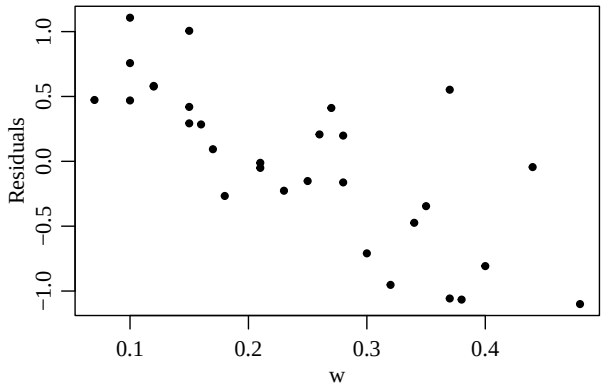

Figure 12: Residuals versus covariate w omitted in the model.

- Residuals versus variable included in the model:
  - Systematic patterns may indicate another relationship between the considered regressor variable and the response variable.
  - Assessment of the homogeneity.

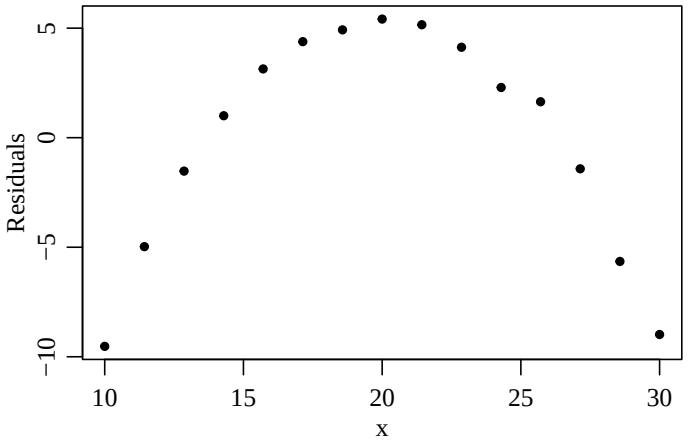

Figure 13: Residuals versus variable x included in the model.

In the Figures 9, 10 and 11 there is a desirable behavior of the residuals when the model is adequate but, on the other hand, Figures 12 and 13 illustrate the inadequacy of the model.

The coefficient of determination R^2 and the adjusted coefficient of determination $\bar{R}^2$ are measures of model adequacy and express how much of the variation of the response variable is explained by the model.

$$R^2 = \frac{SSReg}{SSTotal} \tag{28}$$

$$\bar{R}^2 = 1 - \left(\frac{n-1}{n-p}\right)(1 - R^2) \tag{29}$$

As the number of parameters, p, increases, with n fixed, $\bar{R}^2$ decreases when compared to R^2. This penalty is important because the variation explained by a model with a greater number of parameters will always be greater, which is desirable. On the other hand, the model will be more complex, which is not desirable. Therefore $\bar{R}^2$ is a better measure of suitability than R^2 when the complexity caused by the greater number of parameters is considered.

Example 4. *Let the real part of an electromechanical impedance signature (Ω) be obtained from monitoring a steel plate in an experiment developed at the Gloria campus of the Federal University of Uberlandia. The impedance values were obtained at 3000 frequency points ranging from 40000 to 70000 Hz. Given a large amount of data, only this signature's initial and final values are shown in Table 2.*

Table 2: Initial and end values of the impedance signature.

Frequency	Impedance
40000.00	91.10
40010.00	90.86
40020.01	88.53
40030.01	86.53
⋮	⋮
69969.99	41.58
69979.99	41.39
69990.00	40.76
70000.00	40.71

a) Estimate the parameters of the linearized exponential model (Eq. 6).

Solution

The referred model can be adjusted from the function lm of R with the logarithmic transformation in the base e. The log function does this transformation, and the fitted model is:

$$\widehat{ln(y)} = 5,548 - 2,676 \times 10^{-5} Freq.$$

The following image (Figure 14) illustrates the R code and its output to obtain this model.

```
> fit=lm(log(y)~Freq)
> summary(fit)

Call:
lm(formula = log(y) ~ Freq)

Residuals:
     Min       1Q   Median       3Q      Max
-0.19681 -0.04099 -0.00412  0.03489  0.33350

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.548e+00  6.904e-03   803.6   <2e-16 ***
Freq        -2.676e-05  1.240e-07  -215.8   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.05884 on 2998 degrees of freedom
Multiple R-squared:  0.9395,    Adjusted R-squared:  0.9395
F-statistic: 4.656e+04 on 1 and 2998 DF,  p-value: < 2.2e-16
```

Figure 14: Output of the function *lm* for the linearized exponential model.

b) Assume that the assumptions of normality, homogeneity, and independence of the residuals are met and verify the significance of the model using the F test of the analysis of variance.

Solution

According to the results in item a's figure, the p-value associated with the F test was less than 2.2×10^{-16}. From this, it is concluded that the model is highly significant in explaining the variation in $ln(y)$. The value for the F test presented in Table 1 (Table of analysis of variance) is defined by the probability $P(F > F_0)$ obtained from an F distribution with 1 and 2998 degrees of freedom.

c) Give the value of the coefficient of determination and interpret it.

Solution

The coefficient of determination was 0.9395, which indicates that the fitted model explains 93.95% of the variation in $ln(y)$.

d) Graphically present the signature before and after the trend correction.

Solution

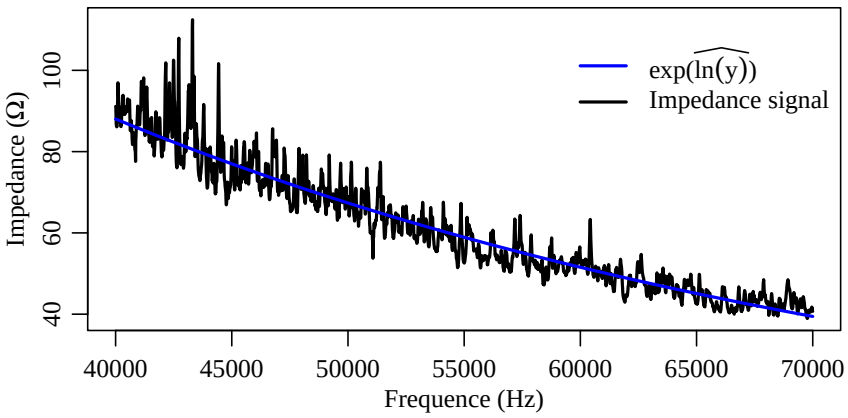

Figure 15: Impedance signature before trend correction.

Since the fitted model satisfactorily models the trend, the impedance signal without trend is obtained by making $y - exp(\widehat{ln(y)})$ *and is shown in Figure 16.*

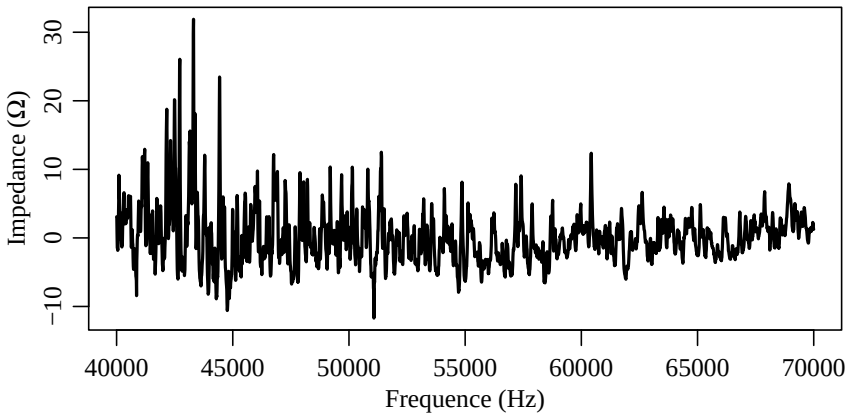

Figure 16: Impedance signature after trend correction.

1.5 Linear Regression With Bootstrapping

The bootstrap resampling method is an alternative to classical parametric inference methodologies. This method calculates the empirical distribution of the statistic of interest, mean, variance, model coefficients, etc., over hundreds or thousands of resamples.

There is no need to worry about assuming a distribution for the statistic in question and, therefore, with the inherent assumptions. The estimates in each resampling allow building the empirical distribution of the distribution of interest.

The central idea of this method is to reproduce several scenarios from the original sample by performing other samples with the same amount of elements but with replacement.

The statistics obtained in each resample are enough to summarize all the information from the mean, variance, and confidence interval, among other possible measures.

Therefore, it is enough to assume that the original sample is reliable and to have enough computational resources to calculate a value of the statistic of interest (θ^*) in each new sample.

Each sample obtained from the original is called a *bootstrap* sample of size n. With the same size as the original sample, from θ^* obtained in each of the hundreds or thousands of samples, it models the distribution of *bootstrap*.

The *bootstrap* method is classified into parametric and non-parametric. In the first, it is necessary to postulate a distribution to generate the samples later; in the second, the samples are generated from the empirical distribution. This chapter will use the non-parametric bootstrap since the necessary assumptions for parametric inference are not always met.

In regression, bootstrapping can be performed by resampling the pairs (dependent and independent variables) and the model residuals. In this chapter, pairwise resampling will be used.

Example 5. *Assume that the electromechanical impedance (Ω) is a function of the temperature at a given frequency point (Hz). Consider the following baseline data set on these two variables and estimate the coefficients of the simple linear regression model using the bootstrap technique. Baseline data sets represent the reference state of the specimen and, in this case, are obtained at different temperatures.*

Temperature ($^{\circ}C$)	Impedance (Ω)
29,80	123.53
27.25	122.98
25.70	122.74
24.90	122.41
23.90	122.50
27.25	123.87
30.10	124.46
31.10	124.85
28.20	123.89
26.50	123.86
25.20	122.79
24.20	122.33
27.75	123.60
30.00	124.44
26.40	123.34

Solution

The following image illustrates the execution of the R source code to obtain the estimates of the intercept (β_0), the slope (β_1) and the coefficient of determination (R^2), from the simple linear regression model, $y_i = \beta_0 + \beta_1 x_i + e_i$

```
> B <- 5000
> data<-cbind(y,x) # y <- impedance values and x<-temperature
> Tboot <- matrix(NA, nrow = B,ncol = 3) # 2 coef + R2
> for(i in 1:B){
+ u <- sample(1:nrow(data),nrow(data), replace = T)
+ data2 <- data[u,]
+ y1 <- data2[,1]
+ x1 <- data2[,2]
+
+ fit <- lm(y1 ~ x1)
+  Tboot[i,]<-c(summary(fit)$coefficients[, 1],summary(fit)$r.squared)
+ }
> b0 <- mean(Tboot[,1]);b0
[1] 114.8429
> b1 <- mean(Tboot[,2]);b1
[1] 0.3159266
> r2 <- mean(Tboot[,3]);r2
[1] 0.8049488
> b0.IC <- quantile(Tboot[,1], prob = c(0.025, 0.975));b0.IC
    2.5%    97.5%
113.2060 117.0977
> b1.IC <- quantile(Tboot[,2], prob = c(0.025, 0.975));b1.IC
     2.5%     97.5%
0.2310095 0.3760934
> r2.IC <- quantile(Tboot[,3], prob = c(0.025, 0.975));r2.IC
     2.5%     97.5%
0.5242340 0.9654569
```

Figure 17: R source code executed to get bootstrap estimates.

The estimated model is $\hat{y}_i = 114.8316 + 0.3163x_i$ *with* $R^2 = 80.77\%$ *and from the confidence intervals, it is verified that the null hypothesis of the parameters* β_0 *and* β_1*, is rejected since the intervals* $b0.IC$ *and* $b1.IC$ *do not include the value zero. The value of* R^2 *indicates that 80.77% of the change in impedance is explained by the change in temperature.*

Figure 18 shows the behavior of impedance as a function of temperature at a given frequency point and the main results obtained from bootstrap estimates.

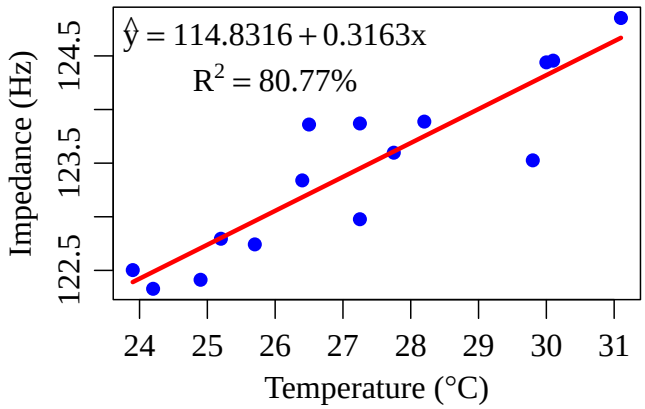

Figure 18: Impedance behavior as a function of temperature.

The bootstrap estimation of the model presented in Eq. 6, for trend correction in the signature, can be obtained similarly to the one shown in the example 5.

Example 6. *Illustrate a case where the assumptions necessary for parametric inference about the example model 4 are not met.*

Solution

```
> fit=lm(log(y)~Freq)
> #summary(fit)
> shapiro.test(residuals(fit))

        Shapiro-Wilk normality test

data:  residuals(fit)
W = 0.97956, p-value < 2.2e-16

> library(lmtest)
> dwtest(modelo)

        Durbin-Watson test

data:  modelo
DW = 0.076243, p-value < 2.2e-16
alternative hypothesis: true autocorrelation is greater than 0

> bptest(modelo)

        studentized Breusch-Pagan test

data:  modelo
BP = 27.069, df = 1, p-value = 1.963e-07
```

Note from the image above that the p-values for the normality (Shapiro-Wilk), independence (Durbin-Watson), and homogeneity of variances (Breusch-Pagan) tests were all less than 0.01. Therefore, the respective tests were significant at 1%, indicating that these assumptions are not met. This finding justifies the use of the bootstrap method for non-parametric inference as applied in the 5 example.

In this chapter, pre-processing of the data set arising from the impedance-based structural health monitoring (SHM) is understood as the investigation and disposal of signatures that are not representative of the actual state of the structure. Matching signatures for trends is also considered a preprocessing technique.

The use of signatures to infer the structure's integrity is classified as post-processing. Therefore, the intermediate step (processing) is defined by using techniques to equate the signatures measured in the baseline acquisition step and the investigation status regarding the effect of temperature. This variable can cause changes in the structure and, if not considered, can lead to erroneous inferences. In the following section, two statistical methodologies can be used for temperature compensation or modeling baseline signatures at a temperature of interest.

2 Data processing

In a sufficiently large number of frequencies, modeling impedance as a function of temperature allows for building reference signatures at temperatures not necessarily observed in the baseline acquisition step. This methodology is called temperature compensation. To apply this methodology, the baseline signatures must be collected in a sufficiently large range of temperatures. In other words, the sampling of temperatures is representative in quantity and amplitude.

The following subsection presents the methodology of polynomial regression, a generalization of simple linear regression, which can be shown in a matrix form as specified in Eq. 11.

2.1 Polynomial regression

The behavior of a dependent variable can be modeled as a function of a single regressor variable by a polynomial model of any order p (Eq. 9). Note that if $p = 1$, we have the simple linear regression model as shown in Eq. 7.

The model shown in Eq. 9 can be written in matrix form as in Eq. 11 and therefore, the vector, least squares estimator for $\boldsymbol{\beta}$, is as shown in Eq. 16.

In polynomial regression, the model matrix, $\boldsymbol{X}$, is composed of columns that are powers of the same variable. Building this matrix implies increasing the correlation or dependence between the columns.

There is multicollinearity when a regressor variable is dependent on another or others. This means that, at least in part, an independent (regressive) variable can be explained by the others. Multicollinearity is, therefore, not a matter of presence or absence but a matter of intensity. This intensity can be quantified by the coefficient of determination, R_k^2, obtained from the adjusted model for *k-th* independent variable as a function of the others. The closer to 1 (100%) the R_k^2 is, the more serious the multicollinearity problem.

The variance inflation factor over the coefficient estimators, VIF_k, is used to measure multicollinearity [Belsley et al., 2005].

$$VIF_k = \frac{1}{1 - R_k^2} \tag{30}$$

Multicollinearity can cause problems such as a high value of the coefficient of determination, non-significance of the parameters (they do not differ significantly from zero), and unexpected (negative/positive) signs for the parameter estimates. Perfect multicollinearity prevents even obtaining least squares estimates since the columns of the matrix $\boldsymbol{X}$, in Eq. 11, will be linearly dependent. Therefore, the classic inverse of the matrix $\boldsymbol{X'X}$ will not exist. This inverse is necessary for the estimation of least squares.

According to Rawlings et al. [1998], the VIF must be below 10 for the multicollinearity problem to be considered unimportant.

A corrective measure is to use data centered, that is, to replace x with $x^* = x - \bar{x}$. Using orthogonal polynomials is a more efficient alternative. Therefore, it should be preferred when this methodology is applicable.

The study of the model's adequacy in meeting the assumptions of normality, independence, and homogeneity of residual variance, as well as the presentation in the 1.4 section, should also be done for polynomial regression. The significance of the model to explain the response variable and the proportion of explained variation are items that must also be investigated in the polynomial regression. A desirable model is one in which the assumptions are met with relatively high and significant R^2.

Example 7. *Consider the data set on the real part of the electromechanical impedance (Ω) obtained at different temperatures ($^{\circ}C$) at a given frequency (Hz). These data come from an experiment developed at the Structural Mechanics Laboratory (LMEst) of the Faculty of Mechanical Engineering (FEMEC) of the Federal University of Uberlandia (UFU).*

Temperature ($^{\circ}C$)	Impedance (Ω)
29.80	631.59
27.25	748.09
25.70	775.73
24.90	775.34
23.90	761.51
27.25	747.76
30.10	605.71
31.10	558.08
28.20	700.31
26.50	767.44
25.20	785.90
24.20	780.12
27.75	719.67
30.00	609.17
26.40	760.08

Fitting the degree 2 polynomial model and checking the multicollinearity problem. In the affirmative case, adopt a strategy to mitigate it. Also, verify that the assumptions of normality, independence, and residual homogeneity are met. Perform all the requested analyses with the help of the R software.

Solution

The following R source code was used to get the questions' required results. In this code, y represents the impedance values and x the respective temperatures.

```
> fit = lm(y~x+I(x^2))
> library(faraway)
> vif(fit)
       x   I(x^2)
736.5912 736.5912
> z=x-mean(x)
> fit2 = lm(y~z+I(z^2))
> vif(fit2)
       z   I(z^2)
1.052605 1.052605
> r=residuals(fit2)
> shapiro.test(r)

        Shapiro-Wilk normality test

data:  r
W = 0.95687, p-value = 0.6382

> require(lmtest)
> dwtest(fit2)

        Durbin-Watson test

data:  fit2
DW = 2.4274, p-value = 0.7653
alternative hypothesis: true autocorrelation is greater than 0

> bptest(fit2)

        studentized Breusch-Pagan test

data:  fit2
BP = 5.5668, df = 2, p-value = 0.06183
```

Figure 19: R source code and output solution for polynomial regression analysis.

The VIF values indicate a severe multicollinearity problem and are the same for the two variables considered. The second is a power of the first. The VIF values' equality is obtained from the regressions of x as a function of x^2 and x^2 as a function of x, which are simple linear regressions. In this case, the R^2 is the square of the Pearson correlation that does not distinguish between the correlations of x and x^2 and between x^2 and x.

On average, the centralization of the explanatory variable x corrected the multicollinearity problem (VIF < 10), and the model adjusted with the new variable, z met the assumptions of normality, independence, and homogeneity according to the Shapiro-Wilk tests, Durbin-Watson and Breusch-Pagan, respectively. The p-values in these tests were greater than 0.05.

The image below shows that the model significantly explains the variation in impedance, according to the F test of the analysis of variance (p-value < 0.01). The percentage of explained variation is high ($R^2 = 98.98\%$), and all coefficients are significantly different from zero according to the t-test. (p-value < 0.01).

```
> summary(fit2)

Call:
lm(formula = y ~ z + I(z^2))

Residuals:
     Min       1Q   Median       3Q      Max
-11.8966  -6.7773   0.8046   5.8502  13.1243

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 743.0317     3.1509  235.81  < 2e-16 ***
z           -28.4358     1.0004  -28.42 2.23e-12 ***
I(z^2)       -5.8123     0.4817  -12.06 4.55e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.279 on 12 degrees of freedom
Multiple R-squared:  0.9898,    Adjusted R-squared:  0.9881
F-statistic: 582.5 on 2 and 12 DF,  p-value: 1.123e-12
```

Figure 20: R source code and output solution for polynomial regression analysis.

The scatter plot with the fitted model are shown in Figure 21.

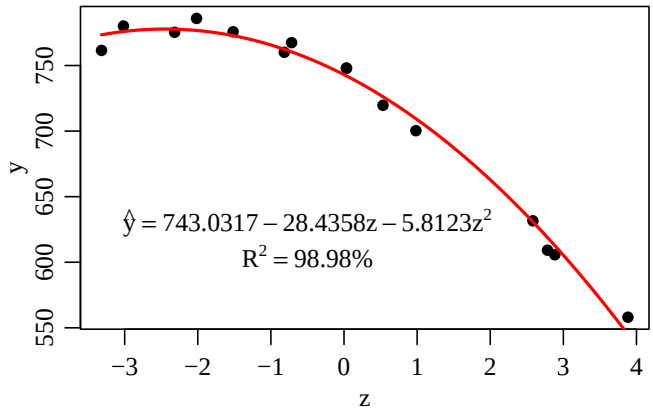

Figure 21: Polynomial regression for the impedance (y) as a function of the temperature centered on the mean ($z = x - \bar{x}$).

The typical characteristic dependence of data set in SHM [Schubert Kabban et al., 2015] motivates the use of statistical methodologies that can consider this autocorrelation in data analysis, for example, in temperature compensation via regression models. Regression models as ARIMA errors is an existing proposal in the literature [Hyndman and Athanasopoulos, 2018] for such an analysis. A summary of this methodology, including the R source codes necessary to obtain results, is presented in the following section.

2.2 Regression with ARIMA errors

The dependence between the impedance values motivates the use of methodologies that can accommodate this characteristic reflected in the residuals of the regression model. The successive nature of the acquisitions carried out by the collection system may justify this dependence.

Regression models with ARIMA errors (Auto-Regressive Integrated Moving Averages Model) are flexible in terms of the residual autocorrelation structure in the sense that it is possible to adopt one among several possibilities (order) for the ARIMA model. The function *auto.arima* of the package *forecast* [Hyndman et al., 2020] of the R software [R Core Team, 2021] is a very useful computational resource that can help in the adjustment and in choosing the most adequate model for each case. This resource will be used in this section to illustrate how to obtain the results in the exemplification analyses.

A polynomial regression with ARIMA errors can be written according to the equation presented in Eq. 31,

$$\begin{aligned} y_t &= \beta_0 + \beta_1 x_t + \beta_2 x_t^2 + \cdots + \beta_k x_t^k + n_t \\ &\text{em que} \quad \phi(B)(1-B)^d n_t = \theta(B) e_t. \end{aligned} \tag{31}$$

In Eq. 31, n_t are the errors or a variable resulting from an AR (Auto-Regressive), MA (Moving Averages), ARMA (Auto-Regressive Moving Averages), and ARIMA (Auto-Regressive Integrated Moving Averages) processes, among others; e_t is called soft noise, that is, a random variable with zero mean and variance σ_e^2 and B is called a delay operator. The index t in Eq. 31 is purposely used to denote the temporal dependence of data acquisition.

The process n_t is said to be moving averages of order q, or $MA(q)$, if,

$$n_t = e_t + \delta_1 e_{t-1} + \delta_2 e_{t-2} + \cdots + \delta_q e_{t-q}, \tag{32}$$

Using the delay operator, Eq. 32 can be rewritten as

$$n_t = (1 + \delta_1 B + \delta_2 B^2 + \cdots + \delta_q B^q) e_t = \theta(B) e_t, \tag{33}$$

On the other hand, if

$$n_t = \alpha_1 n_{t-1} + \alpha_2 e_{t-2} + \cdots + \alpha_p e_{t-p} + e_t \tag{34}$$

then n_t is called an autoregressive process of order p, or $AR(p)$. Using the delay operator, Eq. 34 is:

$$(1 - \alpha_1 B - \alpha_2 B^2 - \cdots - \alpha_p B^p) n_t = e_t \text{ ou } \phi(B) n_t = e_t \tag{35}$$

and therefore the $AR(p)$ process can be rewritten as:

$$n_t = \phi(B)^{-1} e_t = \psi(B) e_t \tag{36}$$

Another possibility for time-dependent data modeling is the autoregressive moving average models, $ARMA(p, q)$. The expression for a $ARMA(p, q)$ model is:

$$n_t = \alpha_1 n_{t-1} + \alpha_2 n_{t-2} + \cdots + \alpha_p n_{t-p} + e_t + \delta_1 e_{t-1} + \delta_2 e_{t-2} + \cdots + \delta_q e_{t-q} \tag{37}$$

In general, representations through models as described in Eq. 37 requires a smaller number of parameters when compared to representations through models as in Eq. 32 or in Eq. 34. Using the delay operator, the model in Eq. 37 is as in Eq. 38:

$$(1 - \alpha_1 B - \alpha_2 B^2 - \cdots - \alpha_p B^p) n_t = (1 + \delta_1 B + \delta_2 B^2 + \cdots + \delta_q B^q) e_t \tag{38}$$

or

$$\phi(B) n_t = \theta(B) e_t \tag{39}$$

The models described in Eq. 32, 34 and 37 are suitable for stationary time series. The trend is a characteristic that, if present, characterizes a lack of stationarity and can be removed by taking one or more differences between consecutive values of the series.

An ARMA model, where n_t is replaced by its *d-th* difference $\Delta^d n_t$, can accommodate some kind of non-stationarity and the differentiated series will be denoted by the expression as shown in Eq. 40.

$$W_t = \Delta^d n_t = (1 - B)^d n_t \tag{40}$$

and the integrated autoregressive process of moving averages, $ARIMA(p, d, q)$ is written according to equation Eq. 41 or Eq. 42,

$$W_t = \alpha_1 W_{t-1} + \alpha_2 W_{t-2} + \cdots + \alpha_p W_{t-p} + e_t + \delta_1 e_{t-1} + \delta_2 e_{t-2} + \cdots + \delta_q e_{t-q} \tag{41}$$

or, equivalently according to Eq. 42:

$$\phi(B)(1 - B)^d n_t = \theta(B) e_t. \tag{42}$$

After establishing the model, either as defined in Eq. 31, 32, 34, 37 or in Eq. 41 the next step is to estimate its parameters. Considering a polynomial regression model with ARIMA errors, the number of parameters will be $k+1+p+q+1+1$ where $k+1$ (Eq. 31) is the number of coefficients of the polynomial model, $p + q + 1$ the number of parameters associated with the ARIMA model (Eq. 41) and 1 more parameter to be estimated is the variance associated with white noise.

Making $\beta = (\beta_0, \beta_1, \cdots, \beta_k)$, $\alpha = (\alpha_0, \alpha_1, \cdots, \alpha_p)$, $\delta = (\delta_0, \delta_1, \cdots, \delta_q)$ and considering $\xi = (\beta, \alpha, \delta, \sigma_e^2)$, the likelihood function for ξ will be $L(\xi|D)$, where D represents the sample data. The maximum likelihood estimators (MLE) of ξ will be the values that maximize L or $l = log(L)$.

After the adjustment, the study of the model's suitability must be carried out. One possibility is to insert more parameters, obtaining a new model. The significance of the parameter(s) inserted allows verifying whether the more complex model improves the fit of the simpler model.

Among the criteria for selecting models, and in particular the regression models with ARIMA errors, the criteria based on the maximum likelihood function are the most used, most often the Akaike Information Criterion (AIC) [Akaike, 1974], the Akaike Information Corrected Criterion (AICc) [Bozdogan, 1987] and the Bayesian Schwarz Criterion (BIC) [Schwarz, 1978].

The adequacy of the models adjusted via the maximum likelihood method can be quantified using pseudo coefficients of determination and, in the case of time series models, also using the precision of the forecasts.

The adequacy of the model must also be done through the analysis of residuals. If the model is adequate, it is expected that the residuals or white noise are normally distributed, randomly around zero, with constant variance and uncorrelated.

In time series analysis, he traditionally uses the Lilliefors [Thode, 2002a] test to verify the hypothesis of normality, the Box-Ljung [Ljung and Box, 1978] or Breusch-Godfrey [Breusch, 1978, Godfrey, 1978] tests to confirm the independence hypothesis and the Breusch-Pagan test [Breusch and Pagan, 1979] to verify the homogeneity of variance.

Example 8. *Consider the data in the table below from an experiment carried out on the Gloria campus from the Federal University of Uberlandia (UFU).*

Table 3: Real part of electromechanical impedance signature (Ω) at 36 temperatures ($^{\circ}C$).

Temperature	Impedance	Temperature	Impedance
18.43	84.36	17.14	88.60
17.95	85.66	18.43	86.09
17.95	86.28	18.27	87.26
19.08	83.71	17.63	88.80
17.14	90.26	17.95	87.73
16.66	91.14	19.24	84.06
16.66	91.28	18.92	84.59
17.47	90.11	17.95	87.32
17.31	90.42	18.27	87.34
16.82	90.72	19.72	84.85
15.86	95.33	19.56	85.10
17.14	94.28	18.92	86.46
16.02	95.76	19.24	85.98
15.53	97.28	20.04	84.95
16.02	96.16	19.56	85.58
17.31	90.72	18.59	87.42
17.14	89.94	18.92	87.01
16.50	88.20	18.76	87.04

a) Fit the simple linear regression model for the impedance as a function of temperature and show that the assumptions of normality and homogeneity are met. Still, the same is not valid for the assumption of independence.

Solution

The model was fitted with y representing the impedance values and x the respective temperatures. In the following image, it can be seen that the p-values for the Lilliefors test (lillie.test) and the Breusch-Pagan test (bptest) were greater than 0.05, and therefore the hypotheses of normality and homogeneity are not rejected at 5% significance. On the other hand, in the Box-Ljung test (checkresiduals(fit2,test="LB")), a p-value lower than 0.05 was obtained, and, therefore, the independence assumption is not met.

The two measures R^2 and mean squared error (mean squared error - mse) were calculated and are helpful for decision making regarding the model's adequacy and for comparing it with another proposed model. The values of these measures were respectively 79.21% and 2.6889.

```
> if(!require(nortest)) install.packages('nortest')
> if(!require(forecast)) install.packages('forecast')
> if(!require(Metrics)) install.packages('Metrics')
> fit1=lm(y~x)
> lillie.test(residuals(fit1))

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  residuals(fit1)
D = 0.0745, p-value = 0.8813

> checkresiduals(fit1,test="LB")

        Ljung-Box test

data:  Residuals
Q* = 33.277, df = 5, p-value = 3.315e-06

Model df: 2.   Total lags used: 7

> bptest(fit1)$p.value
       BP
0.1281203
> cor(y,fitted(fit1))^2
[1] 0.7920986
> mse(y,fitted(fit1))
[1] 2.688923
```

Figure 22: R source code and output solution for fitting the simple linear regression model.

b) Fit the simple linear regression model with an ARIMA structure for the residuals. This can be done with the function auto.arima of the package forecast of R language. Show that the assumptions are met in this case; therefore, the parametric inference from this model is valid.

Solution

With the function auto.arima, the autoregressive structure of order one was selected for the residuals, ARIMA(1,0,0) and the assumptions of normality (lillie.test), of independence (checkresiduals(fit2,test="LB")) and homogeneity (bptest) were met because the p-values of the respective tests were greater than 0.05. The value of the pseudo R^2 *was* 91.60%*, and the month was equal to* 1.0969*.*

Then, from the evaluation of the assumptions, the pseudo R^2*, and the mse, it is concluded that the regression model with AR(1) structure for the residuals is more suitable than the simple linear regression model.*

```
> fit2=auto.arima(y,xreg=cbind(x))
> summary(fit2)
Series: y
Regression with ARIMA(1,0,0) errors

Coefficients:
         ar1  intercept        x
      0.7993   130.0195  -2.3301
s.e.  0.1022     4.3660   0.2369

sigma^2 = 1.197:  log likelihood = -53.26
AIC=114.51   AICc=115.8   BIC=120.84

Training set error measures:
                    ME     RMSE       MAE       MPE      MAPE      MASE      ACF1
Training set 0.09035373 1.047317 0.7300343 0.0870417 0.8156447 0.4754428 0.1086486
> lillie.test(residuals(fit2))

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  residuals(fit2)
D = 0.1389, p-value = 0.07666

> checkresiduals(fit2,test="LB")

        Ljung-Box test

data:  Residuals from Regression with ARIMA(1,0,0) errors
Q* = 3.7959, df = 4, p-value = 0.4343

Model df: 3.   Total lags used: 7

> bptest(residuals(fit2) ~ fitted(fit2))$p.value
        BP
0.06984022
> cor(y,fitted(fit2))^2
[1] 0.915992
> mse(y,fitted(fit2))
[1] 1.096872
```

Figure 23: R source code and output solution for fitting the simple linear regression model with ARIMA structure to residuals.

Polynomial models with ARIMA structure for residuals (example 8) appear as an excellent alternative to traditional polynomial models (example 7) when they do not satisfy the necessary assumptions for inference under the parametric optics. These models can predict electromechanical impedance at temperatures not necessarily observed in the baseline data collection step but in the structure investigation steps. The recurrent use of these models at different frequencies (Hz) allows for modeling or predicting a baseline signature.

In this chapter, the processing is understood as the matching of a signature in the investigation of the structure with its respective reference signature regarding the effect of the temperature variable. The post-processing stage of statistical analysis, the inference about the structure's status, is called post-processing. Some possible statistical techniques at this stage are presented in the 3 section below.

3 Data post-processing

3.1 Comparison tests on damage metrics

3.1.1 Hypothesis test on more than two means

The F test, in the analysis of variance, is used to verify whether the null hypothesis, defined as the equality of groups or populations, on average, should be rejected or not. The set of possibilities complementary to the null hypothesis represents the alternative hypothesis. The expression "at least one population differs from the others on average" clearly illustrates the alternative hypothesis. In this case, it is also assumed that the groups are independent.

One-way analysis of variance consists of breaking down the total variance of the data or the response variable into variance between and within factor levels. There is a statistical difference between them, on average. Otherwise, the evidence will be weak, and the different hypotheses will not be supported.

The test statistic used in the analysis of variance is given by:

$$F = \frac{\frac{SS_{between}}{J-1}}{\frac{SS_{within}}{J(I-1)}} \tag{43}$$

where $SS_{Between}$ is the sum of squares between levels, SS_{Whitin} is the sum of squares within levels, J is the number of levels or groups, and I is the number of values or repetitions in each group. Assuming that the assumptions of homogeneity of variance and normality of the residuals are met, the variable defined in (43) follows a distribution F of *Snedecor* with $J-1$ and $J(I-1)$ degrees of freedom. When the number of repetitions is not the same in all groups, the amount $J(I-1)$ must be calculated considering this fact by making $\sum_{j=1}^{J}(I_j-1)$ in where I_j is the number of repetitions at each factor level.

The hypothesis test (test F) of the analysis of variance is a right-sided test because the null hypothesis will only be rejected if the observed value of F is located at the extreme right of the F distribution of *Snedecor*. The decision rule or critical value is determined from the F distribution with $J-1$ and $J(I-1)$ degrees of freedom and a given significance; if the observed value of F is greater than the critical value, it is said that it is strong enough evidence to reject the hypothesis of equality between the means, that is, the null hypothesis, and it is concluded that at least one mean differs from the others for the established significance.

The analysis of variance is linked to a model that can be defined as follows:

$$y_{ij} = \mu + f_j + e_{ij} \tag{44}$$

y_{ij} is the i-th observation of the response variable at the j-th level of the factor f, μ is constant (generally the average of all observations), and e_{ij} is the error associated with the i-th repetition of the j-th level. For the inferences to be valid, it is necessary that the assumption of normality, independence, and homogeneity of the residuals (constant variance) be valid, in notation $e_{ij} \sim N(0, \sigma^2)$, These assumptions can be verified using the Shapiro-Wilk tests [ROYSTON, 1982] and the Bartlett test [BARTLETT, 1937]. The validity of Bartlett's test is dependent on the normality hypothesis. When this hypothesis is not met, the Levene test [LEVENE, 1960] is an alternative that only requires the existence of any continuous distribution for the variable in question. For the independence of the residuals, we have the Durbin-Watson test [DURBIN and WATSON, 1950].

The hypotheses involved in the analysis of variance do not allow us to verify which levels or groups differ from each other. They only allow us to reject or not the hypothesis of equality of all

means. Therefore, in situations where the decision is for at least one mean different from the others (rejecting the null hypothesis), there is a need for an additional test for two-by-two comparisons or to contrast groups of means. The Student-Newman-Keuls (SNK) [NEWMAN, 1939] and Tukey tests [TUKEY, 1953], among others, are indicated for two-by-two comparisons. The Scheffé Test [SCHEFFÉ, 1953] is flexible as it allows comparing any contrast between means. The two groups (contrasts) involved can have different amounts of means and each mean can come from a different number of repetitions or observations.

Example 9. *Consider the following data about the CCD metric under baseline conditions (B), and damage 1, 2, and 3 (D1, D2, and D3). These data come from electromechanical impedance signals obtained from a steel plate from the petrochemical industry.*

Table 4: CCD by condition and repetition.

Condition	Replications									
	1	2	3	4	5	6	7	8	9	10
B	0.0234	0.0305	0.0393	0.0474	0.0385	0.0208	0.0243	0.0408	0.0342	0.0206
D1	0.1739	0.1577	0.1840	0.1894	0.1586	0.1999	0.1856	0.1579	0.1623	0.1654
D2	0.2767	0.2659	0.2554	0.2564	0.2718	0.2761	0.2957	0.2546	0.2746	0.2875
D3	0.3500	0.3749	0.3570	0.3693	0.3841	0.3810	0.3624	0.3786	0.3517	0.3637

a) Check the model assumptions, normality of residuals, homogeneity of variances, and independence of residuals at 5% of significance, using the Shapiro-Wilk, Bartlett, and Durbin-Watson tests.

Solution

```
> shapiro.test(residuals(lm(dados1~danos),data=dados2))

        Shapiro-Wilk normality test

data:  residuals(lm(dados1 ~ danos), data = dados2)
W = 0.94708, p-value = 0.06021

> bartlett.test(dados1~danos,data = dados2)

        Bartlett test of homogeneity of variances

data:  dados1 by danos
Bartlett's K-squared = 2.0527, df = 3, p-value = 0.5616

> durbinWatsonTest(lm(dados1~danos),data=dados2)
 lag Autocorrelation D-W Statistic p-value
   1      -0.1033981       2.192324    0.91
 Alternative hypothesis: rho != 0
```

Figure 24: R source code for verification of assumptions.

From the results presented in Figure 24 there is no significant evidence to reject the hypotheses of normality (p-value = 0.06021), of homogeneity of variances (p-value=0.5616) and of independence of residuals (p-value = 0.91). Therefore, the assumptions about the residuals of the model are satisfied.

b) Do the analysis of variance for the model, conclude on the F test and apply Tukey's test for comparison of means, if applicable.

Solution

```
> dic(danos,dados1, quali = TRUE,mcomp = "tukey", sigT = 0.05, sigF = 0.05)
------------------------------------------------------------------------
Quadro da analise de variancia
------------------------------------------------------------------------
           GL      SQ       QM     Fc      Pr>Fc
Tratamento  3 0.61534 0.205112 1237.4 2.3526e-36
Residuo    36 0.00597 0.000166
Total      39 0.62130
------------------------------------------------------------------------
CV = 6.1 %

------------------------------------------------------------------------
Teste de normalidade dos residuos
Valor-p:  0.06020767
De acordo com o teste de Shapiro-Wilk a 5% de significancia, os residuos podem ser considerados normais.
------------------------------------------------------------------------

------------------------------------------------------------------------
Teste de homogeneidade de variancia
valor-p:  0.5615511
De acordo com o teste de bartlett a 5% de significancia, as variancias podem ser consideradas homogeneas.
------------------------------------------------------------------------

Teste de Tukey
------------------------------------------------------------------------
Grupos Tratamentos Medias
a        dano3   0.36727
 b       dano2   0.27147
  c      dano1   0.17347
   d     danob   0.03198
------------------------------------------------------------------------
```

Figure 25: R source code for analysis of variance and multiple comparison.

The result of the F test presented in Figure 25 allows us to conclude that there is significant evidence to reject the hypothesis of equality of means (p-value < 0.05), and at least one differs from the others. Using the Tukey test, all experimental conditions vary, two by two, regarding the average CCD metric. Thus, the methodology effectively-identified structure changes since all damages differed from baseline.

3.1.2 Wilcoxon-Mann-Whitney Test

This test is an alternative to the t test when the normality assumption is not met in at least one of the populations involved. Its application can be made even when there is no information about the distributions, but the variables involved have a measurement scale that is at least ordinal. The two samples must be independent. The interest in applying this test is to verify if a population tends to have higher values than the other or if they have the same median.

The Wilcoxon-Mann-Whitney test is based on the ranks of the values obtained by combining the two samples. The values are ordered from smallest to largest, regardless of which population each value comes from. Let R_1 be the sum of ranks in sample 1 and R_2 be the sum of ranks in sample 2, then define the test statistic as follows:

$$W_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1; W_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2 \tag{45}$$

where n_1 and n_2 are respectively the sizes of samples 1 and 2. Make W equal to the smallest value between W_1 and W_2 as the value of the test statistic. For values of $n < 20$, the Mann-Whitney U table is used to find the critical values necessary for the decision to reject the null hypothesis. The normal distribution can be approximated if n is greater than 20.

Let R be the sum of the ranks in sample 1:

$$z = \frac{R - \mu_R}{\sigma_R} \tag{46}$$

where $\mu_R = \frac{n_1(n_1+n_2+1)}{2}$, $\sigma_R = \sqrt{\frac{n_1 n_2(n_1+n_2+1)}{12}}$, n_1 and n_2 are respectively the sizes of samples 1 and 2. The variable defined in Eq. 49 has an approximately standard normal distribution and therefore, the critical values needed for decision making are determined from this distribution.

Example 10. *Considering the following data set on the CCD metric (baseline and damage), calculated from electromechanical impedance signatures obtained from steel sheet from the petrochemical industry.*

Table 5: CCD metric for damage and repetition.

Condition	Replications									
	1	2	3	4	5	6	7	8	9	10
Baseline	0.0054	0.0469	0.0447	0.0145	0.0162	0.0197	0.0195	0.0048	0.0261	0.0066
Damage	0.1255	0.1051	0.2240	0.4486	0.1718	0.1134	0.1199	0.1064	0.4540	0.3349

a) Check normality assumptions for populations at 5% significance.

Solution

```
> shapiro.test(ccdbase)

        Shapiro-Wilk normality test

data:  ccdbase
W = 0.867, p-value = 0.09221

> shapiro.test(ccddano)

        Shapiro-Wilk normality test

data:  ccddano
W = 0.78514, p-value = 0.009561
```

Figure 26: R source code and its output solution for the normality test.

According to the results (Figure 26), there is no significant evidence to reject the hypothesis of normality (p-value = 0.09221) for the baseline population. Still, the same does not occur for the damaged population (value -p = 0.0009561); Still, a non-parametric alternative, such as the Mann-Whitney test, should be used to compare the two conditions.

b) Using the Mann-Whitney test, comparing baseline and damage CCD metric populations at 5% significance.

Solution

$$R_1 = 55; W = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1 = 100 + \frac{10(11)}{2} - 55 = 100 \quad (47)$$

$$R_2 = 155; W = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2 = 100 + \frac{10(11)}{2} - 155 = 0 \quad (48)$$

Thus, the smallest value for W is zero ($W = 0$). The value of the Mann-Whitney table with $\alpha = 0.05$ and $n_1 = n_2 = 10$ is 23, so as $W < 23$ the null hypothesis is rejected.

Another alternative to solve this question is through the function wilcox.test, in the R language, as shown in Figure 27.

```
> wilcox.test(ccdbase,ccddano, correct=FALSE, alternative="two.sided")

        Wilcoxon rank sum test

data:  ccdbase and ccddano
W = 0, p-value = 1.083e-05
alternative hypothesis: true location shift is not equal to 0

> median(ccdbase)
[1] 0.01785
> median(ccddano)
[1] 0.14865
>
```

Figure 27: R source code and output solution for comparison of baseline and damaged groups.

According to the results presented in Figure 27, there is significant evidence to reject the hypothesis that the populations are centered in the same position (p-value <0.05), that is, the two populations differ, with the median damaged CCD (0.14865) higher than baseline (0.01785).

3.1.3 Kruskal-Wallis Test

The Kruskal-Wallis test is an alternative to conventional analysis of variance when the assumptions of normality and homogeneity of variances are not met.

The procedure for implementing the test consists of sorting the data set from all samples or (K) levels as a single sample and assigning ranks to all values. For each sample or level, determine the sum of ranks R_k with $k = 1, 2, 3, \cdots, K$. The following test statistic follows a *chi-square* distribution with $K - 1$ degrees of freedom.

$$H = \frac{12}{N(N+1)} \sum_{k=1}^{K} \left(\frac{R_k^2}{n_k} \right) - 3(N+1) \quad (49)$$

In Eq. 49, N is the total number of observations and n_k is the size or number of observations at the factor level k. If the ranks are similarly distributed among the sample groups, then H will be a relatively small number. The Kruskal-Wallis test is right-sided; that is, the hypothesis of equality between the groups will be rejected if the test statistic, H, is a value located at the extreme right of the *chi-square distribution* . Based on the significance of the test and the degrees of freedom, the critical value for decision-making is determined.

If the necessary assumptions are met, the use of analysis of variance is preferable since, in general, parametric tests are more powerful than non-parametric tests.

Example 11. *Let the following values from the baseline CCD metric (B) and damage 1, 2, and 3 (D1, D2, and D3) conditions, be calculated from electromechanical impedance signatures for steel sheet from the petrochemical industry.*

Table 6: CCD for damage and repetitions

Condition	Replications									
	1	2	3	4	5	6	7	8	9	10
B	0.0383	0.0364	0.0480	0.0385	0.0232	0.0387	0.0434	0.0476	0.0213	0.0496
D1	0.1804	0.1797	0.1792	0.1698	0.1663	0.1926	0.1836	0.1719	0.1995	0.1660
D2	0.2540	0.2671	0.2841	0.2601	0.2508	0.2765	0.2786	0.2539	0.2924	0.2727
D3	0.3534	0.1143	0.4307	0.4080	0.2019	0.2710	0.1899	0.3106	0.3743	0.4862

a) Show that the assumptions of normality of residuals, homogeneity of variances, and independence of residuals, at 5% significance, are not met. Use the Shapiro-Wilk, Bartlett, and Durbin-Watson tests implemented in R language.

Solution

```
> shapiro.test(residuals(lm(dados1~danos),data=dados2))

        Shapiro-Wilk normality test

data:  residuals(lm(dados1 ~ danos), data = dados2)
W = 0.79106, p-value = 4.434e-06

> bartlett.test(dados1~danos,data = dados2)

        Bartlett test of homogeneity of variances

data:  dados1 by danos
Bartlett's K-squared = 73.809, df = 3, p-value = 6.522e-16

> durbinWatsonTest(lm(dados1~danos),data=dados2)
 lag Autocorrelation D-W Statistic p-value
   1      -0.08284929         1.939   0.462
 Alternative hypothesis: rho != 0
```

Figure 28: Normality, homogeneity and independence tests in R language.

From the results presented in Figure 28, it appears that there is significant evidence to reject both the hypothesis of normality and homogeneity of residues (p-value < 0.05). The independence hypothesis is not rejected (p-value = 0.462). Being of interest, the groups must be compared using non-parametric tests.

b) Comparing the groups using the Kruskal-Wallis test, at 5% significance.

Solution

```
> comparison<-kruskal(dados1,danos,console=TRUE,alpha=0.05,group=TRUE)

Study: dados1 ~ danos
Kruskal-Wallis test's
Ties or no Ties

Critical Value: 30.56927
Degrees of freedom: 3
Pvalue Chisq  : 1.047413e-06

danos,  means of the ranks

      dados1  r
dano1   16.7 10
dano2   29.0 10
dano3   30.8 10
danob    5.5 10

Post Hoc Analysis

t-Student: 2.028094
Alpha    : 0.05
Minimum Significant Difference: 5.131162

Treatments with the same letter are not significantly different.

      dados1 groups
dano3   30.8      a
dano2   29.0      a
dano1   16.7      b
danob    5.5      c
```

Figure 29: R source code and solution for the Kruskal-Wallis test application.

From the results presented in Figure 29, it can be concluded that there is significant evidence to reject the hypothesis that all groups are equal (p-value < 0.05). The Kruskal-Wallis test is a global test, and an additional examination is needed for specific comparisons such as two-by-two comparisons. For this, the modified t test can be used. This test verified that all experimental conditions present significantly different CCD metrics (Figure 30) and ascend from baseline condition to damage 3.

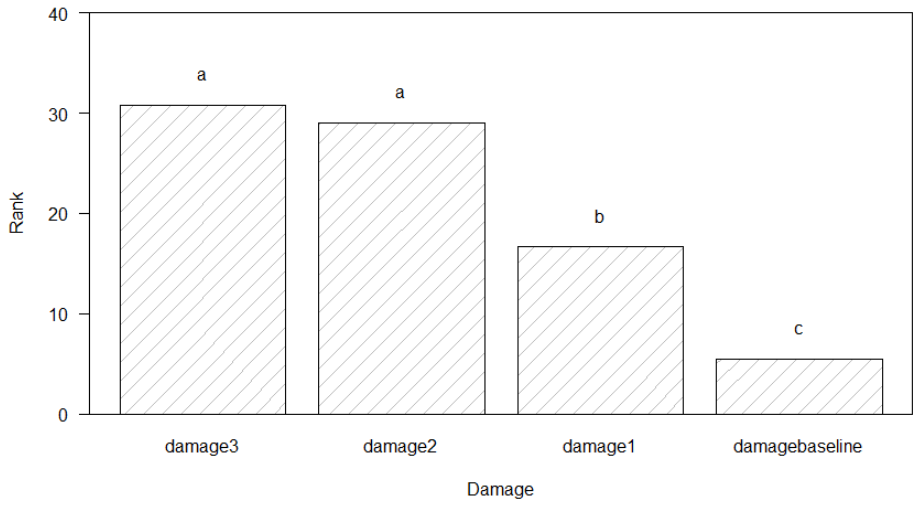

Figure 30: Rankings and Kruskal-Wallis test results for the four experimental conditions.

3.2 Threshold definition

The decision on a significant change in the structure, based on some damage metric such as the CCD metric, for example, is not a deterministic issue, given the inherent variations in data generation. In this way, there will be variation between the CCD values calculated using impedance signatures collected in the baseline measurements and between the CCD values calculated at times of damage investigation even under the same experimental conditions.

These uncertainties must be considered in the decision-making regarding the change in the structure of the baseline measurement step.

The procedure presented in this section consists of establishing a threshold or threshold from the confidence intervals for the mean and standard deviation. The interval estimates for these two parameters include the error associated with the forecast of each one of them.

The threshold is defined by adding to the upper limit of the confidence interval for the mean three times the upper limit of the confidence interval for the standard deviation. A CCD value greater than the threshold will indicate a change in the structure of the pristine condition. On the other hand, a CCD value less than the threshold will suggest that the structure has not changed, at least expressively, when compared to the reference status.

3.2.1 Parametric Threshold

The parametric intervals with $1-\alpha$ confidence for the mean and the variance of the metric are obtained from Eq. (50) and (51). It should be noted that these intervals are obtained from the baseline data. The denomination parametric intervals is due to the fact that normal distribution is assumed for the sample mean and *chi-square* for the variable $\chi^2 = \dfrac{(n-1)S^2}{\sigma^2}$, from which the confidence interval for σ^2 is established.

$$\left[\bar{X} - t_{(v,\frac{\alpha}{2})}\frac{S}{\sqrt{n}}; \bar{X} + t_{(v,\frac{\alpha}{2})}\frac{S}{\sqrt{n}}\right] \tag{50}$$

$$\left[\frac{vS^2}{\chi^2_{(v,\frac{\alpha}{2})}}; \frac{vS^2}{\chi^2_{(v,1-\frac{\alpha}{2})}}\right] \tag{51}$$

where, $\bar{X}$ and S^2 are respectively the mean and variance calculated from the metric values obtained in the pristine condition, n is the sample size, $v = n-1$ are the degrees of freedom, $t_{(v,\alpha)}$ is a quantile of the t distribution of *Student* with v degrees of freedom and probability $\frac{\alpha}{2}$ accumulated to its right, $\chi^2_{(v,\frac{\alpha}{2})}$ and $\chi^2_{(v,1-\frac{\alpha}{2})}$ are quantiles of the distribution of *Chi-Square* with probabilities $\frac{\alpha}{2}$ and $1-\frac{\alpha}{2}$ rolled to the right, respectively. The confidence interval for the standard deviation is determined by calculating the square root of the confidence interval limits for the variance Eq. (51), as shown in the equation Eq. (52).

$$\left[\sqrt{\frac{vS^2}{\chi^2_{(v,\frac{\alpha}{2})}}}; \sqrt{\frac{vS^2}{\chi^2_{(v,1-\frac{\alpha}{2})}}}\right] \tag{52}$$

The threshold or threshold for the decision rule is then determined through Eq. (53)

$$\textit{Threshold} = \mu_{\max} + 3\sigma_{\max} \tag{53}$$

where $\mu_{\max}$ and $\sigma_{\max}$ are respectively the upper limits for the confidence intervals for the mean and standard deviation built from the baseline data set.

The inference from the interval estimates calculated through Eq. (50) and (52) is valid if the normality hypothesis about the variable in question, the CCD metric, for example, is not rejected. This hypothesis can be verified using the Shapiro-Wilk [Royston, 1995] test, using the experimental data.

Example 12. *Let the following CCD damage metric values be calculated from electromechanical impedance signatures for a pristine condition of the steel plate.*

0,0223	*0,0453*	*0,0700*	*0,0632*	*0,0183*
0,0523	*0,0550*	*0,0290*	*0,0375*	*0,0057*
0,0259	*0,0194*	*0,0359*	*0,0611*	*0,0372*
0,0098	*0,0426*	*0,0206*	*0,0311*	*0,0317*

a) Checking the hypothesis of normality, at 5% of significance, using the Shapiro-Wilk test.

Solution

```
> y=c(0.0223, 0.0453, 0.0700, 0.0632, 0.0183, 0.0523, 0.0550,
+     0.0290, 0.0375, 0.0057, 0.0259, 0.0194, 0.0359, 0.0611,
+     0.0372, 0.0098, 0.0426, 0.0206, 0.0311, 0.0317)
> shapiro.test(y)

        Shapiro-Wilk normality test

data:  y
W = 0.97289, p-value = 0.8144
```

The p-value associated with the Shapiro-Wilk test is greater than 0.05; Then, the normality hypothesis is not rejected at 5% of significance.

b) Building the interval for the mean with 95% confidence.

Solution

```
> t.interval = function(data, conf.level){
+   n = length(data)
+   df = n - 1
+   t = qt((1 - conf.level)/2, df, lower.tail = FALSE)
+   xbar = mean(data)
+   s = sd(data)
+   c(xbar - t * s/sqrt(n), xbar + t * s/sqrt(n))
+ }
> t.interval(y,0.95)
[1] 0.02731178 0.04407822
```

Figure 31: Confidence interval obtained via parametric inference for the mean of the CCD metric.

With 95% confidence, the range from 0.02731178 to 0.04407822 contains the mean of the CCD variable (Figure 31).

c) Obtaining the range for the standard deviation with 95% confidence.

Solution

```
> var.interval = function(data, conf.level){
+  n = length(data)
+  df = n - 1
+  chi_lower = qchisq((1 - conf.level)/2, df)
+  chi_upper = qchisq((1 - conf.level)/2, df, lower.tail = FALSE)
+  v = var(data)
+  c((df*v)/chi_upper, (df*v)/chi_lower)
+  }
> var.interval(y,0.95) # Conf. interval for the variance
[1] 0.0001855628 0.0006844617
> sqrt(var.interval(y,0.95)) # Conf. interval for the sd
[1] 0.01362214 0.02616222
>
```

Figure 32: Confidence interval obtained via parametric inference for the standard deviation of the CCD metric.

With 95% confidence, the range from 0.01362214 to 0.02616222 contains the standard deviation of the CCD variable (Figure 32).

d) Determine the parametric threshold from the previous results.

Solution

```
> Thershold = 0.04407822 + 3*0.02616222
> Thershold
[1] 0.1225649
>
```

Figure 33: Threshold obtained via parametric inference.

Therefore, CCD values lower than 0.1226 (Figure 33) do not suggest significant changes in the integrity of the structure at investigation moments after the baseline condition.

The purpose of implementing the functions in the R language for calculating the confidence intervals for the mean and variance is that they can be used in other situations where the interest is to obtain these interval estimates.

The statistical methodology for the construction of confidence intervals in Eq. (50) and (51) assumes that the hypothesis of normality of the variable in question is not rejected. This hypothesis was tested based on the available sample data set. If there is no normality, the threshold can be obtained from the non-parametric approach via *bootstrap* resampling estimation.

3.2.2 Non parametric threshold

As mentioned, this method calculates the empirical distribution of the statistic of interest, mean, variance, etc., over hundreds or thousands of resamples. Then, no need to worry about assuming a distribution for the statistic in question and with the inherent assumptions.

With the values of the statistic of interest obtained in each resampling, it is enough to summarize all the information from the mean, variance, and confidence interval, among other possible measures.

The following example calculates the threshold (Eq. 53) in which the intervals for the mean and the standard deviation are defined using the *bootstrap* method.

Example 13. *The CCD values presented below were calculated from 16 impedance signatures measured from an experiment developed at the Structural Mechanics Laboratory of the Faculty of Mechanical Engineering of the Federal University of Uberlandia. The data set was adapted considering the signatures only in the frequencies indicated in Table 7. The median signature of the group was considered as baseline status. Therefore, the CCD metric was calculated using the pairs formed by each signature and the median signature. The variation in this baseline metric is due only to random or uncontrolled factors and not structural changes. The CCD values are as shown below.*

0,0202	*0,0202*	*0,0201*	*0,0202*	*0,0201*	*0,0201*	*0,0201*	*0,0201*
0,0201	*0,0202*	*0,0201*	*0,0223*	*0,0209*	*0,0209*	*0,0208*	*0,0207*

a) Checking the hypothesis of normality of the variable via the Shapiro-Wilk test.

Solution

```
> ccd=c(0.0202 ,0.0202, 0.0201, 0.0202, 0.0201, 0.0201, 0.0201, 0.0201,
+         0.0201, 0.0202, 0.0201, 0.0223, 0.0209, 0.0209, 0.0208, 0.0207)
> shapiro.test(ccd)

        Shapiro-Wilk normality test

data:  ccd
W = 0.64392, p-value = 4.336e-05
```

Figure 34: Normality test for the CCD variable.

The sample data provide enough evidence to reject the normality hypothesis about the CCD variable at 0.5% of significance (Figure 34).

b) Calculating bootstrap confidence intervals for the mean of the CCD metric.

Solution

```
> boot.ic.mean <- function(data, R,conf.level) {
+   resamples <- lapply(1:R, function(i) sample(data, replace=T))
+   r.mean <- sapply(resamples, mean)
+   ci = quantile(r.mean,probs=c((1-conf.level)/2,(1 + conf.level)/2))
+   c(ci)
+ }
> boot.ic.mean(data=ccd,R=10000,conf.level=0.95)
      2.5%      97.5%
0.02020625 0.02076250
>
```

Figure 35: Bootstrap confidence interval for the mean.

With 95% confidence, the range from 0.0202 to 0.0208 contains the mean of the CCD variable (Figure 35).

c) Calculating bootstrap confidence intervals for the standard deviation of the CCD.

Solution

```
> boot.ic.sd <- function(data, R,conf.level) {
+   resamples <- lapply(1:R, function(i) sample(data, replace=T))
+   r.sd <- sapply(resamples, sd)
+   ci = quantile(r.sd,probs=c((1-conf.level)/2,(1 + conf.level)/2))
+   c(ci)
+ }
> boot.ic.sd(data=ccd,R=100000,conf.level=0.95)
        2.5%        97.5%
0.0002272114 0.0008650385
>
```

Figure 36: Bootstrap confidence interval for the standard deviation.

With 95% confidence, the range from 0.0002 to 0.0008 contains the population standard deviation of the CCD metric (Figure 36).

d) Determining the threshold from the results obtained in the previous items via the non-parametric method of bootstrap.

Solution

```
> Thershold = 0.02075625 + 3*0.0008650385
> Thershold
[1] 0.02335137
>
```

Figure 37: Threshold obtained via bootstrap method.

In later health conditions than the baseline, obtaining a CCD value lower than 0.0233 (Figure 37) indicates that the structural integrity hypothesis is not rejected.

4 Concluding Remarks

The quality of the data set to infer the structure's condition is of principal importance under penalty of erroneous decisions. Thus, atypical signatures should be disregarded in the baseline and posterior signature measuring step. Signatures associated with the CCD metric's outliers should also be considered outliers and discarded.

The trend in the impedance curve along the frequency domain, i.e., a characteristic not necessarily atypical, can affect the results if its effect is different in the different impedance signatures. In this chapter, this characteristic was modeled using the linearized exponential model.

The effect of temperature on impedance was another application explored in this chapter, using regression polynomials. The residual dependence attributed, at least in large part, to the consecutive form of data collection was also considered in the modeling. The status of the structure's

integrity was compared by comparing the CCD metrics obtained for pristine and investigated cases. Thresholds were also determined to verify statistically significant changes in the structure's behavior.

In conclusion, parametric inference techniques cannot always be applied; therefore, both this kind of inference and alternative non-parametric methods were presented. Parametric inference should be preferred as it is more powerful when the conditions for its application are met.

Acknowledgements

The authors are thankful to Petrobras for the financial support via the research project (R&D).

References

H. Akaike. A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723, 1974.

M. S. BARTLETT. Propriedades de suficiência e testes estatísticos. *Anais da Royal Society de Londres. Série A - Ciências Matemáticas e Físicas*, 160(901):268–282, 1937.

D. A. Belsley, E. Kuh, and R. E. Welsch. *Regression diagnostics: Identifying influential data and sources of collinearity*. John Wiley & Sons, 2005.

H. Bozdogan. Model selection and akaike's information criterion (aic): The general theory and its analytical extensions. *Psychometrika*, 52(3):345–370, 1987.

T. S. Breusch. Testing for autocorrelation in dynamic linear models. *Australian economic papers*, 17(31):334–355, 1978.

T. S. Breusch and A. R. Pagan. A simple test for heteroscedasticity and random coefficient variation. *Econometrica: Journal of the Econometric Society*, pages 1287–1294, 1979.

J. Durbin. Tests for serial correlation in regression analysis based on the periodogram of least-squares residuals. *Biometrika*, 56(1):1–15, 1969.

J. DURBIN and G. S. WATSON. Testing for serial correlation in least squares regression. *I. Biometrika, London*, 37:409–428, 1950.

B. Efron and R. Tibshirani. An introduction to the bootstrap/bradley efron and robert j, 1993.

J. Fox and S. Weisberg. *An R companion to applied regression*. Sage publications, 2018.

L. G. Godfrey. Testing against general autoregressive and moving average error models when the regressors include lagged dependent variables. *Econometrica: Journal of the Econometric Society*, pages 1293–1301, 1978.

J. Gross and U. Ligges. *nortest: Tests for Normality*, 2015. URL `https://CRAN.R-project.org/package=nortest`. R package version 1.0-4.

R. Hyndman, G. Athanasopoulos, C. Bergmeir, G. Caceres, L. Chhay, M. O'Hara-Wild, F. Petropoulos, S. Razbash, E. Wang, and F. Yasmeen. *forecast: Forecasting functions for time series and linear models*, 2020. URL `http://pkg.robjhyndman.com/forecast`. R package version 8.12.

R. J. Hyndman and G. Athanasopoulos. *Forecasting: principles and practice*. OTexts, 2018.

N. Khatun et al. Applications of normality test in statistical analysis. *Open Journal of Statistics*, 11(01):113, 2021.

E. Langford. Quartiles in elementary statistics. *Journal of Statistics Education*, 14(3), 2006.

H. LEVENE. Robust tests for equality of variances. *In: I. Olkin, et al., Eds., Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling, Stanford University Press, Palo Alto*, 1960.

G. M. Ljung and G. E. Box. On a measure of lack of fit in time series models. *Biometrika*, 65(2): 297–303, 1978.

M. H. A. NEWMAN. Elementos da topologia de conjuntos de pontos planos. *Cambridge*, 1939.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021. URL `https://www.R-project.org/`.

J. O. Rawlings, S. G. Pantula, and D. A. Dickey. *Applied regression analysis: a research tool*. Springer, 1998.

R. A. Reddy. *Applied data analysis and modeling for energy engineers and scientists.* Springer, Nova Iorque; Dordrecht; Heidelberg; Londres., 2011.

J. Royston. Shapiro-wilk normality test and p-value. *Applied statistics*, 44(4):547–551, 1995.

J. P. ROYSTON. An extension of shapiro and wilk's w test for normality to large samples. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 31(2):115–124, 1982.

H. A. SCHEFFÉ. A method for judging all contrasts in the analysis of variance. *Biometrika*, 40 (1-2):87–110, 1953.

C. M. Schubert Kabban, B. M. Greenwell, M. P. DeSimio, and M. M. Derriso. The probability of detection for structural health monitoring systems: Repeated measures data. *Structural Health Monitoring*, 14(3):252–264, 2015.

G. Schwarz. Estimating the dimension of a model. *The annals of statistics*, pages 461–464, 1978.

J. R. Taylor. *Introdução à análise de erros.* Bookman, Porto Alegre., 2012.

H. Thode. Testing for normality marcel dekker. *Inc. New York*, pages 99–123, 2002a.

H. Thode. Testing for normality marcel dekker. *Inc. New York*, pages 99–123, 2002b.

J. W. TUKEY. The problem of multiple comparisons. *Unpublished manuscript, Princeton University*, 1953.

J. W. Tukey. *Exploratory data analysis*, volume 2. Reading, Mass., 1977.

A. Zeileis and T. Hothorn. Diagnostic checking in regression relationships. *R News*, 2(3):7–10, 2002. URL `https://CRAN.R-project.org/doc/Rnews/`.

A Baseline data set from an experiment carried out on steel sheets at the Structural Mechanics Laboratory (LMEst) of the Faculty of Mechanical Engineering (FEMEC) of the Federal University of Uberlandia (UFU).

Table 7: Real part of electromechanical impedance signatures (Ω) by frequency (Hz), in 16 samples and median.

Freq,	Impedance Signatures																
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	Mediana
40000.00	25.75	25.93	25.93	25.40	25.77	25.77	25.94	25.94	25.94	25.75	25.76	25.92	25.75	25.21	25.57	25.75	25.77
40010.00	26.26	26.26	26.07	25.90	26.08	26.09	26.26	26.09	26.26	26.26	26.26	26.23	26.06	25.71	26.06	26.08	26.09
40020.01	26.20	26.20	26.21	26.20	26.21	26.38	26.20	26.37	26.20	26.02	26.20	26.02	26.01	26.01	25.83	25.82	26.20
40030.01	26.99	26.98	27.00	26.99	26.99	26.99	27.00	26.98	27.00	26.98	26.99	26.60	26.61	26.60	26.60	26.78	26.98
40040.01	29.17	29.01	29.00	29.01	29.00	29.01	29.01	29.02	29.02	29.01	29.01	28.62	28.95	28.80	28.79	28.78	29.01
40050.02	32.66	32.66	32.66	32.66	32.67	32.66	32.67	32.67	32.67	32.69	32.69	32.64	32.81	32.81	32.63	32.80	32.67
40060.02	36.62	36.46	36.63	36.46	36.44	36.63	36.46	36.63	36.46	36.44	36.62	36.26	36.62	36.63	36.61	36.59	36.60
40070.02	37.53	37.53	37.54	37.54	37.54	37.54	37.54	37.54	37.54	37.54	37.53	36.99	37.35	37.53	37.53	37.53	37.53
40080.03	37.66	37.49	37.66	37.67	37.67	37.85	37.67	37.66	37.66	37.67	37.67	37.25	37.47	37.46	37.62	37.48	37.66
40090.03	38.48	38.45	38.45	38.28	38.27	38.45	38.46	38.46	38.28	38.45	38.27	38.20	38.44	38.25	38.25	38.25	38.36
40100.03	37.44	37.63	37.65	37.66	37.64	37.63	37.66	37.63	37.63	37.63	37.64	37.80	37.81	37.79	37.81	37.78	37.64
40110.04	35.24	35.25	35.24	35.24	35.24	35.25	35.42	35.44	35.25	35.25	35.25	35.61	35.42	35.60	35.42	35.61	35.25
40120.04	32.91	32.72	32.91	32.92	32.72	32.91	32.91	32.91	32.91	33.10	32.92	32.91	32.92	32.92	32.92	32.94	32.91
40130.04	30.35	30.36	30.18	30.54	30.36	30.36	30.54	30.55	30.35	30.55	30.55	30.18	30.18	30.19	30.19	30.01	30.36
40140.05	27.56	27.36	27.56	27.56	27.36	27.37	27.56	27.56	27.56	27.55	27.56	27.19	26.98	26.99	27.01	26.98	27.46
40150.05	25.21	25.21	25.21	25.21	25.21	25.21	25.21	25.21	25.21	25.21	25.21	24.84	24.82	24.84	24.83	24.83	25.21
40160.05	24.53	24.53	24.53	24.35	24.34	24.34	24.34	24.34	24.35	24.53	24.35	24.14	23.96	24.14	24.32	24.15	24.34
40170.06	24.40	24.40	24.60	24.40	24.40	24.58	24.42	24.58	24.58	24.40	24.58	24.03	24.21	24.20	24.21	24.20	24.40
40180.06	25.38	25.55	25.38	25.56	25.56	25.37	25.38	25.38	25.39	25.56	25.39	25.18	25.36	25.36	25.36	25.18	25.38
40190.06	27.45	27.63	27.45	27.45	27.45	27.46	27.46	27.46	27.45	27.44	27.46	27.08	27.26	27.43	27.43	27.26	27.45
40200.07	30.36	30.35	30.36	30.36	30.35	30.36	30.54	30.36	30.35	30.36	30.36	29.99	30.17	30.17	30.35	30.18	30.36

40210.07	32.25	32.28	32.25	32.08	32.26	32.26	32.26	32.26	32.09	32.26	32.08	31.87	31.72	31.71	31.72	31.90	32.17
40220.07	30.33	30.16	30.16	30.15	30.33	30.16	30.16	30.16	30.14	30.16	30.16	29.94	29.75	29.76	29.75	29.75	30.16
40230.08	27.03	27.04	27.05	27.04	27.04	27.05	27.05	27.05	27.05	27.23	27.05	26.66	27.00	27.00	27.00	27.00	27.04
40240.08	24.75	24.75	24.93	24.75	24.75	24.95	24.94	24.95	24.75	24.77	24.95	24.54	24.91	25.10	25.09	24.91	24.91
40250.08	24.08	23.91	24.08	23.91	23.92	24.09	23.90	24.09	23.72	24.10	23.91	23.86	23.91	24.26	24.09	23.90	23.91
40260.09	24.00	24.00	24.00	24.00	24.00	24.18	24.01	24.00	24.01	24.00	24.01	24.17	23.98	23.98	23.98	23.99	24.00
40270.09	24.30	24.48	24.31	24.31	24.32	24.32	24.32	24.31	24.32	24.32	24.32	24.15	24.12	24.13	24.12	24.13	24.31
40280.09	24.29	24.30	24.31	24.48	24.48	24.48	24.31	24.31	24.30	24.48	24.31	23.78	24.11	24.10	24.12	24.12	24.30
40290.10	23.91	23.91	23.91	23.91	23.91	23.91	23.91	23.91	23.91	24.09	23.92	22.99	23.54	23.53	23.53	23.71	23.91
40300.10	23.47	23.64	23.46	23.64	23.47	23.47	23.66	23.49	23.49	23.48	23.48	22.92	23.10	23.27	23.10	23.27	23.47
40310.10	23.91	23.92	23.74	23.92	23.91	23.74	23.92	23.91	23.91	23.92	23.74	23.00	23.54	23.54	23.54	23.53	23.82
40320.11	24.57	24.57	24.57	24.59	24.56	24.59	24.57	24.59	24.56	24.59	24.58	23.65	24.54	24.54	24.55	24.37	24.57
40330.11	25.25	25.23	25.24	25.23	25.23	25.23	25.24	25.24	25.23	25.24	25.23	24.64	25.38	25.20	25.20	25.22	25.23
40340.11	26.39	26.38	26.39	26.39	26.39	26.39	26.40	26.40	26.22	26.40	26.22	26.32	26.53	26.36	26.35	26.35	26.39
40350.12	28.80	28.78	28.79	28.79	28.80	28.80	28.79	28.79	28.79	28.80	28.62	29.23	29.10	28.93	28.93	28.76	28.80
40360.12	32.38	32.39	32.39	32.22	32.37	32.40	32.40	32.22	32.37	32.38	32.40	33.08	32.55	32.55	32.36	32.57	32.39
40370.12	35.15	35.15	34.98	35.14	34.98	35.14	35.14	34.97	35.15	34.98	34.97	35.39	34.99	34.97	34.96	34.97	34.98
40380.13	35.32	35.32	35.33	35.31	35.31	35.31	35.32	35.30	35.32	35.12	35.31	34.61	34.95	35.13	35.12	34.94	35.31
40390.13	34.08	33.90	33.90	33.90	33.90	33.89	33.90	33.89	33.90	33.89	33.89	33.31	33.69	33.70	33.69	33.70	33.89
40400.13	32.02	32.02	32.03	32.02	32.02	32.02	32.03	32.04	32.03	32.02	32.02	31.46	31.46	31.46	31.46	31.47	32.02
40410.14	29.12	29.12	29.12	28.94	29.12	29.30	29.11	29.11	29.12	29.12	29.11	28.73	28.73	28.72	28.54	28.54	29.11
40420.14	27.03	27.03	27.03	26.85	26.85	26.84	27.02	26.85	26.85	27.04	27.03	26.64	26.64	26.64	26.64	26.64	26.85
40430.14	25.85	25.86	25.86	25.86	25.86	25.86	25.86	25.86	25.86	25.85	25.86	25.66	25.48	25.66	25.47	25.66	25.86
40440.15	25.43	25.42	25.61	25.42	25.42	25.43	25.43	25.61	25.44	25.44	25.62	25.41	25.41	25.22	25.40	25.41	25.43
40450.15	25.39	25.22	25.21	25.22	25.21	25.40	25.40	25.40	25.40	25.39	25.40	25.21	25.03	25.03	25.02	25.01	25.22
40460.15	24.66	24.85	24.67	24.85	24.84	24.66	24.85	24.85	24.67	24.85	24.67	24.47	24.48	24.48	24.47	24.48	24.67
40470.16	23.91	23.91	23.92	23.91	23.91	23.91	24.09	23.92	23.91	23.91	24.10	23.71	23.70	23.54	23.72	23.72	23.91
40480.16	23.29	23.29	23.29	23.31	23.30	23.31	23.31	23.30	23.30	23.31	23.30	23.10	23.10	22.92	22.92	23.10	23.30
40490.16	23.55	23.55	23.55	23.72	23.73	23.73	23.55	23.55	23.55	23.55	23.55	23.35	23.52	23.35	23.35	23.34	23.55