EXACTA Engenharia de Produção eISSN 1983-9308

Artigos

https://doi.org/10.5585/2023.22928

Normality tests: a study of residuals obtainedon time series tendency modeling

Testes de normalidade: estudo dos resíduos obtidos na modelagem da tendência de uma série temporal

Recebido: 16 set. 2022

Aprovado: 03 abr. 2023

Versão do autor aceita publicada online: 03 abr. 2023

Publicado online: 03 abr. 2023

Como citar esse artigo - American Psychological Association (APA)

Cardoso, F. C., Berri, R. A., Lucca, G., Borges, E. N., & Mattos, V. L. D. (jan./mar. 2025). Normality tests: a study of residuals obtainedon time series tendency modeling. *Exacta, 23*(1), p. 134-158. https://doi.org/10.5585/2023.22928

Submeta seu artigo para este periódico

Processo de Avaliação: *Double Blind Review*

Editor: Dr. Luiz Fernando Rodrigues Pinto

Dados Crossmark

EXACTA Engenharia de Produção eISSN 1983-9308

https://doi.org/10.5585/2023.22928

Processo de Avaliação: *Double Blind Review*
Editor: Prof. Dr. Luiz Fernando Rodrigues Pinto
Recebido em: 16 set. 2022
Aprovado em: 03 abr. 2023

Artigos

Normality tests: a study of residuals obtainedon time series tendency modeling

Testes de normalidade: estudo dos resíduos obtidos na modelagem da tendência de uma série temporal

Fabian Corrêa Cardoso Rafael Alceste Berri Giancarlo Lucca Eduardo Nunes Borges and Viviane Leite Dias de Mattos

Universidade Federal do Rio Grande, Rio Grande, RS, Brasil

Nota dos Autores

Autores declaram que não há conflitos de interesses.
Acknowledgments: I thank the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) for the doctoral scholarship, and to the Postgraduate Program in Computer Modeling from Federal University of Rio Grande that make this work possible. Thanks to Professor Dr. Bruno Lopes Dalmazo for the English language review. This work was partially financed by CNPq and FAPERGS (23/2551-0000126-8).

Abstract

he normality analysis in the distribution of residuals is a determining criterion to verify and validate a model. For example, in modeling financial time series by linear regression, the residuals should be independent of each other, identically distributed, have a normal distribution, and be homoscedastic. Thus, it was aimed to study the performance of some normality tests applied on the residuals obtained from linear regression modeling of time series tendency using polynomials of different degrees. The Jarque-Bera, Anderson-Darling, Kolmogorov-Smirnov-Lilliefors, Doornik-Hansen, and Shapiro-Wilk tests were used, there was agreement in almost all test results, with the exception of the Doornik-Hansen test.

Keywords: normality tests, residual analysis, time series modeling

Resumo

Testes de normalidade: estudo dos resíduos obtidos na modelagem da tendência de uma série temporal

A análise de normalidade na distribuição dos resíduos é um critério determinante

para verificar e validar um modelo. Na modelagem de séries temporais financeiras por regressão linear, por exemplo, os resíduos devem ser independentes uns dos outros, identicamente distribuídos, possuir uma distribuição normal e homocedásticos. Desta forma, este estudo tem como objetivo estudar o desempenho de alguns testes de normalidade aplicados em resíduos obtidos da modelagem por regressão linear da tendência de uma série temporal utilizando polinômios de diferentes graus. Foram utilizados os testes de Jarque-Bera, Anderson-Darling, Kolmogorov-Smirnov Lilliefors, Doornik-Hansen e Shapiro-Wilk, havendo concordância quase que na totalidade dos resultados dos testes, com exceção do teste Doornik-Hansen.

Palavras-chave: testes de normalidade, análise de resíduos, modelagem de séries temporais

1 Introduction

A Time Series (TS) is a set of observations ordered in sequence throughout a time interval (Morettin & Toloi, 2018). Its modeling can be performed by various methods and techniques, including linear regression. When modeling a TS, one way to verify the adequacy of the model found is by analyzing the residuals, which are the difference between the values observed in the TS and the modeled values in the training set. These residuals should be independent, preferably homoscedastic, and normally distributed (Hyndman & Athanasopoulos, 2021).

Simple linear regression expresses the statistical relation between two variables. It can be used to fit a statistical model to a TS for making forecasts. In linear regression, the importance of normality in the residuals is to obtain higher reliability of the model inference (Schmidt & Finan, 2018). Indeed, the assumption of normality is one of the most important assumptions of parametric procedures (Ahmad & Sherwani, 2015; Yap & Sim, 2011). For Das and Imon (2016, p. 11), “it is essential to assess normality of a data before any formal statistical analysis”.

Various procedures, graphics or analytical, have been used to assess the assumption of normality. Although the graphical methods are useful for checking normality of sample data, they don’t provide conclusive evidences (Yap & Sim, 2011), so analytical methods are better.

Tests used to verify the normality of data, including residuals, follow the rules of hypotheses tests of statistical inference that admit a certain margin of error based on evidence. Thus, different tests, and criteria can be used to measure the normality property of the residuals.

Usually, in the TS analysis, the Jarque-Bera test (Jarque & Bera, 1987) is used to confirm, or not, the normality of the residuals. The independence property is usually verified by the Ljung-Box test (Ljung & Box, 1978), while homoscedasticity can be verified by the Autoregressive Conditional Heteroscedasticity (ARCH) test (Engle, 1982). To verify normality, one can also use the Anderson-Darling (Anderson & Darling, 1952), Kolmogorov-Smirnov-Lilliefors (Lilliefors, 1967), Doornik-Hansen (Doornik & Hansen, 1994), and Shapiro-Wilk (Shapiro & Wilk, 1965) tests.

For Yap and Sim (2011, p. 2141), if the distribution is asymmetric, the Shapiro-Wilk is the best test to analyze normality, followed by the Anderson-Darling test. But, if the distribution is asymmetric, D'Agostino test and Shapiro-Wilk test have good power when low kurtosis is present. On the other hand, for high kurtosis, Jarque-Bera, Shapiro-Wilk and Anderson-Darling can be used and have good performance. Ahmad and Sherwani (2015, p. 332) observed that some tests for normality have been affected by changing the sample size, level of significance and alternate distribution. They concluded that Shapiro-Wilk and Shapiro-Francia tests had better results in almost all cases analyzed. On the other hand, Nunes, Mattos and Konrath (2021, p. 177) concluded that, for samples from normality data, Jarque-Bera, Shapiro-Wilk, Anderson-Darling and Kolmogorov-Simirnov-Lilliefors tests have good performance, but for asymmetric data and small sample, Shapiro-Wilk has the better performance.

In this context, the purpose of this paper is to study some normality tests applied to the residuals obtained from linear regression modeling of time series tendency, using polynomials of different degrees.

This paper is organized as follows: Section 2 presents the literature review on normality tests, Section 3 presents the methodology, and Section 4 points out and discusses the results obtained. The concluding remarks are in Section 5.

2 Normality Tests

The normal distribution, also known as Gaussian, is represented by a probability density function that is symmetric and bell-shaped, widely used in statistical inference. The mean is at its center in this distribution, having approximately 99.72% of its values between 3 standard deviations (for more and less) (Lopes, Branco & Soares, 2013). Moreover, the mesokurtic distribution indicates that its tails should not be light or heavy (Akamine & Yamamoto, 2013).

Hypotheses tests are used to verify the presence of normality in a data set, which can be parametric or nonparametric. The first assume that the population follows a specific distribution with a set of parameters (Kaur & Kumar, 2015).

The performance of parametric or nonparametric normality tests searches to confirm the null hypothesis. A hypothesis test, in statistical inference, investigates whether a specific statement is accepted or rejected based on the information obtained from the TS, for example. The null hypothesis (H_0) represents the initial belief (inference), but the results can disprove it, so it is rejected. The alternative hypothesis (H_1) is the one that prevails if H_0 is rejected (Hill, Griffiths & Judge, 2010).

However, the possibility of erroneous inferences must be considered. Thus, there are two possibilities of error: Type I Error: rejecting H_0 when it should be accepted and Type II Error: accepting H_0 when it should be rejected. The p-value is used to decide whether or not to reject H_0, and it represents the probability of obtaining a sample statistic at least as extreme as that resulting from the sample data, under the assumption that the null hypothesis is true. When the p-value of a hypothesis test is less than the chosen value of significance (α), one rejects H_0 (Hill *et al.*, 2010).

2.1 Jarque-Bera Test (JB)

This test evaluates the normality of a random sample through some measures of the distribution shape, having been proposed in Jarque and Bera (1987). According to Bueno (2018, p. 82), in this test, the null hypothesis verified is that the coefficient of asymmetry is null, and the coefficient of kurtosis is equal to 3, e.g., these properties assume the expected values for a normal distribution.

To determine the test statistics from an independent and identically distributed random sample of a random variable, $y_1, y, \dots, y_n$, according to Hill *et al.* (2010, p. 160), we use Eq. (1):

$$JB = n\left[\frac{A^2}{6} + \frac{(k-3)^2}{24}\right], \qquad (1)$$

where n is the number of observations in the TS, A is an estimate of the skewness coefficient, and k is an estimate of the kurtosis coefficient, found by Eq. (2) and Eq. (3), respectively and, according to Seier (2002, p. 2):

$$A = \frac{\sum_{i=1}^{n}\left(\frac{y_i-\overline{y}}{s}\right)^3}{n}, \qquad (2)$$

$$k = \frac{\sum_{i=1}^{n}\left(\frac{y_i-\overline{y}}{s}\right)^4}{n}. \qquad (3)$$

In these equations, $\overline{y}$ and s are the mean and standard deviation of the random sample analyzed, respectively.

The JB test statistics follows a chi-square, χ^2, asymptotic distribution with two degrees of freedom, which quickly determines the p-value. This simple and easy test is widely used in TS, more specifically, to analyze the normality property in the residuals.

2.2 Shapiro-Wilk Test (SW)

Shapiro-Wilk is one of the most widely used tests to evaluate the assumption of normality, being a nonparametric test that allows verifying whether the observed distribution is that expected for the data analyzed (Lopes *et al.*, 2013). As evidenced by Ghasemi and Zahediasl (2012, p. 487), it is based on the correlation between the data and its correspondents in the presence of normality and tests the null hypothesis that the random variable analyzed follows a specific theoretical distribution (which may be the normal one) with unspecified mean and variance (Ahsanullah, Kibria & Shakil, 2014). However, the test statistics used does not have a probability distribution, and the p-values were determined by Monte Carlo simulation (Shapiro & Wilk, 1965).

Let be a random sample of a variable $y_1, y_2, \dots, y_n$, arranged in ascending order. The test statistics, according to Das and Imon (2016, p. 9), is calculated as in Eq. (4) and Eq. (5):

$$W = \frac{\left(\sum_{i=1}^{n} a_i y_i\right)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}, \quad (4)$$

$$a' = (a_1, a_2, \dots, a_n) = \frac{c' V^{-1}}{(c' V^{-2} c)^{1/2}}, \quad (5)$$

where $\overline{y}$ is the sample mean, and according to Pino (2014, p. 23), $c' = (c_1, c_2, \dots, c_n)$ is the vector of expected values for the order statistics, and V is the covariance matrix between these statistics, so $(a_1, a_2, \dots, a_n)$ "can be obtained from the table presented by Shapiro and Wilk (1965)", according to Das and Imon (2016, p. 9).

According to Güner, Frankford and Johnson (2009, p. 1746), the W statistic "can be interpreted as a ratio of two estimates of the variance of the sample", assuming values between zero and one and converging to one as the Gaussian sample size increases. According to Pino (2014, p. 23), it is the most widely used test in regression and correlation analysis in checking for the presence of this property in the residuals. Leotti, Birck and Riboldi (2005, p. 1) consider that the data do not follow a normal distribution and SW is superior to the other tests of adherence to normality.

2.3 The Kolmogorov-Smirnov (KS) Test

The Kolmogorov-Smirnov (KS) test is also nonparametric and can be used to assess the similarity between two cumulative distributions, one empirical and one theoretical, and can be used as a test of adherence with different types of theoretical models. According to Razali and Yap (2011, p. 23), in the null hypothesis of this test, it is considered that the sample data follow a specified distribution (which can be the normal distribution) with predefined mean and variance and, consequently, do not present substantial differences between the two cumulative frequency distributions. In this case, the cumulative distribution of the variable in the sample is expected to be

very close to the cumulative distribution of the theoretical model considered, presenting only minor random deviations.

According to Lilliefors (1967, p. 399), the test statistics focuses on the higher difference between the two cumulative frequency sequences, as presented in Eq. (6):

$$“D = max_x|F^*(x) - S_n(x)| \quad (6)$$

where $S_n(X)$ is the sample cumulative distribution function and $F^*(x)$ is the cumulative normal distribution function” and “if the value of D exceeds the critical value in the table, one rejects the hypothesis that the observations are from a normal population”.

According to Jarque and Bera (1987, p. 168), the KS test is a distance test, specifying mean and variance. According to Lilliefors (1967, p. 399), KS has the advantage that it can be applied to small samples, although the decision is made based on a single value: the highest difference.

2.3.1 KS-Lilliefors Test (KS-L)

The KS-Lilliefors (KS-L) test statistics calculation is performed in the same way as in KS, but the p-values are different because the distribution of the test statistics is constructed from parameter estimates (Öztuna, Elhan & Tüccar, 2006). The main difference between KS and KSL is that, when determining the distributions of the cumulative frequencies expected by the theoretical distribution, the former uses population mean and variance, the latter uses estimates obtained from the values observed in the sample (Abdi & Molin, 2007). When the distribution is not completely identified the probability of type 1 at KS-Lilliefors test tends to be smaller than in this KS test (Lilliefors, 1967 apud Ahmad & Scherwani, 2015).

2.4 Anderson-Darling Test (AD)

This test is a modification of the test proposed by Cramér-Von Mises (1928), and is also nonparametric. It can be used to evaluate whether the distribution of sample data follows some

specific theoretical distribution, and like KS, it works with cumulative distribution functions, and like SW, it works with order statistics. The test checks the null hypothesis that the analyzed random variable follows a specific theoretical distribution (which can be the normal distribution). According to Anderson and Darling (1952, p. 194), the main innovation is incorporating a weight function, which allows it more flexibility. According to Leotti *et al.* (2005, p. 2), this statistic is based on the EDF (Empirical Distribution Function) of the data, belonging to the quadratic class of statistics based on this function.

In this test, according to Ahsanullah *et al.* (2014, p. 47), the null hypothesis states that the data set follows the normal distribution.

To determine the test statistics from *n* ordered sample data, $y_1, y_2, \dots, y_n$, according to Anderson (2010, p. 53), Eq. (7) can be used:

$$A_n^2 = -n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)\left[logu_i + log\left(1 - u_{(n-i+1)}\right)\right], \quad (7)$$

where $u_i = F^0(y_i)$, with $0 \leq u_i \leq 1$ representing the theoretical cumulative distribution function probabilities. For a normal distribution, these values are determined by Eq. (8), according to Arshad, Rasool and Ahmad (2003, p. 86):

$$u_i = P\left(u \leq \frac{y_i - \overline{y}}{s}\right), \quad (8)$$

where $\overline{y}$ and s are the mean and standard deviation of the random sample analyzed and $\frac{y_i - \bar{y}}{s}$is the standard score.

Razali and Yap (2011, p. 31) state that the AD test is not as good as the SW test, but Stephens (1974, p. 733) states that the A^2 statistic is one of the best statistics for assessing departure from normality of an empirical distribution function, agreeing with Grace and Wood (2012).

2.5 Doornik-Hansen Test (DH)

According to Górecki, Horváth and Kokoszka (2020, p. 681), the authors "proposed a simple multivariate normality test based on transformed skewness and kurtosis." Farrel, Salibian-Barrera and Naczk (2007, p. 3) state that "these authors conducted a small simulation study that suggests that their statistics has better properties than other tests based on skewness and kurtosis." But, according to Joenssen and Vogel (2014, p. 17), this test loses performance with increasing dimensionality.

In agreement with Adkins (2018, p. 95), "the Doornik-Hansen test has a χ^2 distribution if the null hypothesis is true", while the alternative hypothesis (H1) considers that the sample does not come from a normal distribution. The use of the statistic χ^2 proves to be computationally a more complex test than Jarque-Bera (Adkins, 2018).

According to Górecki et al. (2020, p. 682), to determine the test statistics, corrected by Wickham (2015) in concordance with Lobato-Velasco (2004), equations 9 to 11 can be used:

$$E_d = \sum_{j=1}^{p}\left(z_{1j}^2 + z_{2j}^2\right), \quad (9)$$

$$\text{where } z_{1j} = f_{1,N}\left(\tau_{Y,j}\right) \quad (10)$$

$$\text{and } z_{2j} = f_{2,N}\left(\kappa_{Y,j}, \tau_{Y,j}\right), \quad (11)$$

where κ* and τ* are the skewness and kurtosis modified by Lobato-Velasco (2004), as can see in equations 12 to 13.

$$\tau_{Y,j}^{*} = \frac{m_{Y,3,j}}{v_{Y,3,j}^{1/2}} \text{ e} \quad (12)$$

$$\kappa_{Y,j}^{*} = \frac{m_{Y,4,j}}{v_{Y,4,j}^{1/2}}. \quad (13)$$

Yigit and Mendes (2014, p. 31) state that Doornik and Hansen (1994) "reported that their test had a good performance in terms of retaining type I error rate at the nominal level and good test power values."

3 Others Properties

3.1 Homoscedasticity: ARCH Test

In the analysis of residuals, homoscedasticity must also be evaluated. It comes from the concept of homogeneous scattering of errors, e.g., with equal variance, because, in this case, there tends to be no outliers and symmetry (Gujarati & Porter, 2011). In a TS, the model errors should have the same variance for any value of the explanatory variables (Alves & Pereda, 2018).

There are several tests to detect heteroscedasticity. However, the ARCH or its variations are usually used when using financial TS due to the high volatility. More details on this test can be seen in Engle (1982).

3.2 Independence of Residuals: Ljung-Box Test(LB)

To verify the independence of the residuals, one can use the Ljung-Box test, which is a variation of Box-Pierce (Hill *et al.*, 2010). This test is set with the null hypothesis that the model does not have a misfit, e.g., autocorrelations up to lag S are equal to zero. If this happens, the residuals are random and independent up to this lag. The null hypothesis says that the residuals are not autocorrelated (Hill *et al.*, 2010).

The test statistics is given, according to Moretin and Toloi (2018, p. 205), by Eq. (14):

$$Q(K) = n(n+2)\sum_{j=1}^{K}\frac{\grave{\rho}_j^2}{(n-j)}, \qquad (14)$$

where n is the total number of observations in the TS, $\grave{\rho}_S$ is the estimated autocorrelation of the TS at lag K, and j is the variance from 1 to K within the TS.

It is necessary to define the maximum lag for the model to calculate the independence of the residuals by Ljung-Box; thus, a value between 10 and 15 is used, according to Morettin and Toloi (2018, p. 205). More details about this test can be seen in Ljung and Box (1978).

4 Methodology

To study the tests of normality were generated four models with different polynomials degrees to the raw data of the VALE3 ON (common) stock daily closing price in the year 2018. To obtain the data set to be analyzed, the website of the brazilian stock exchange (Brasil, Bolsa, Balcão-B3)[a] was accessed. The Vale company was chosen because it represents a high percentage in the Bovespa index, around 15%, and it is a stock of a Brazilian multinational with high turnover and, therefore, high liquidity. It was chosen to use the raw data of the daily closing of the stock of VALE3 ON (common) in the year 2018, totaling 245 observations, for being commonly used data and for the period presenting the challenge of containing initially uptrend and after downtrend.

The software R Studio (R Core Team, 2019), which is considered to be very reliable, was used to perform the calculations. It has several libraries able to assist this work.

Initially, an exploratory analysis was performed in R Studio, and some summary measures were determined: minimum value, first quartile, median, third quartile, maximum value, asymmetry coefficient, and kurtosis coefficient. For asymmetry and kurtosis calculation the commands skewness and kurtosis, respectively, were used, from the Performance Analytics library (Peterson et al., 2020).

We have also plotted a boxplot for visual assessment of asymmetry and the presence of outliers, a histogram for visual assessment of the shape of the distribution and identification of possible gaps, and a Q-Q Plot for visual assessment of normality.

Polynomial linear regression models, described by Eq. (15), are used to describe the trend of a TS:

$$E(y|t) = \beta_0 + \beta_1 t + \beta_2 t^2 + \cdots + \beta_m t^m. \qquad (15)$$

where y is the variable presented by TS, T is the number of observations with $t = 1,2, \ldots, T$, β_m are the coefficients of the polynomial regression, and m is the degree of the polynomial.

a Raw data can be obtained on https://syr.us/tAQ

In the present study, after performing exploratory analysis of the TS data to evaluate the performance of the normality tests, four models, determined from Eq. (15) ($1 \leq m \leq 4$), were fitted, which generated four data sets formed by their respective residuals.

For model fitting, we used the lstm command, from the forecast library (Hyndman et al., 2022). This is done aiming to analyze the p-value and verify if the models have a good explanatory power. The models fitted analysis of variance measures employed the F statistic, which tests the null hypothesis that all model coefficients are null, and also employed the coefficient of determination (R^2), which is a measure of variation that occurs in a dependent variable, in function of the variations that occur in the independents variables in a linear regression model (Gujarati & Porter, 2011, p. 95).

The residuals, that are the difference between the observed value and the model predicted value, were calculated by the resid command, and after the summary command was used to display its results.

Then the normality tests mentioned above were performed to analyze the residuals and test if the null hypothesis, that the data have a normal distribution, is confirmed: Jarque-Bera, Shapiro-Wilk, Kolmogorov-Smirnov Lilliefors, Doornik-Hansen, and Anderson-Darling.

Jarque-Bera test was executed by the jarque.bera.test command from the tseries library (Trapletti & Hornik, 2020), also the Shapiro-Wilk test was executed by a command from the same library, named shapiro.test.

Anderson-Darling and Kolmogorov-Smirnov Lilliefors tests were executed with the commands ad.test e lillie.test, respectively, both from nortest library (Gross & Ligges, 2015).

Doornik-Hansen test was executed by DH.test command from mvnTest library (Pya et al., 2016).

Significance level of 0.05 was used to perform the linear regression analysis and normality tests.

Tests for independence and homoscedasticity were also performed, complementing the analysis. In the first case, the Ljung-Box test was used, while the ARCH test was used in the second.

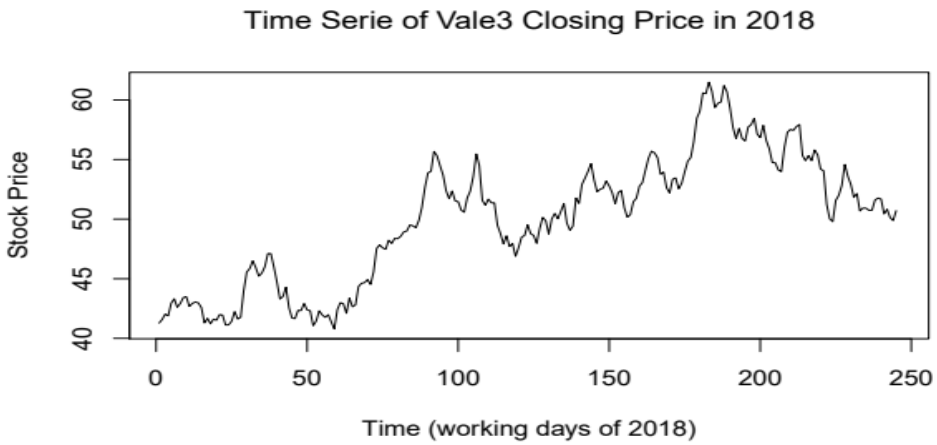
The Ljung-Box test was performed by the command *Box.test*. The ARCH test was executed with the command *ArchTest* from the *FinTS* library (Tsay, 2013).

5 Results

Figure 1 shows the TS curve, where one can see that there was an upward trend (with jumps) in the first 180 working days of the year, which reverted to a downward trend in the last 65 working days.

Figure 1

VALE3 stock closing price time series in 2018

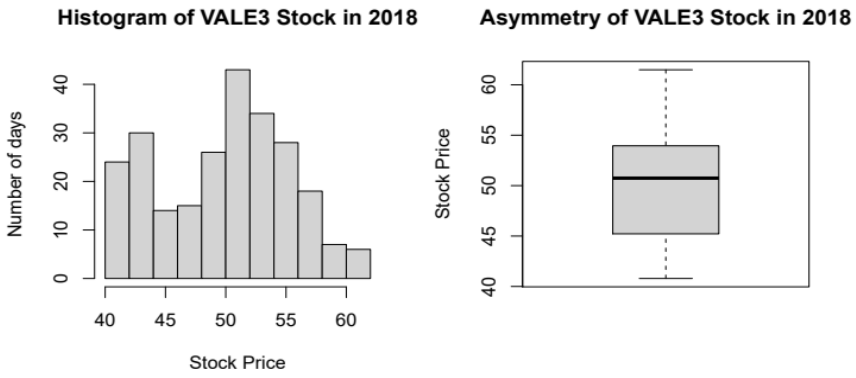


Source: The authors, 2022.

The histogram and boxplot, presented in Figure 2, respectively, visually indicate that the distribution of the TS data is not close to a normal distribution and can be considered symmetrical, corroborated by the asymmetry coefficient that assumed the value of -0.07660403. The kurtosis coefficient provided the value -0.9121602, which reveals a platykurtic curve with light tails.

Figure 2

VALE3 stock in 2018 histogram and boxplot



Source: The authors, 2022.

Table 1 presents the fitting quality of the models. When the p-value is smaller than the level of significance adopted, the model is considered significant, i. e., at least one of the variables included in the model changes the expected value of the response (Mattos, 2004).

Table 1

Quality of fit measures

	First-degree Model	Second-degree Model	Third-degree Model	Fourth-degree Model
F statistic	451,2	315,7	295,2	229,1
Degrees of freedom	243	242	241	240
p-value	$2,2 \times 10^{-16}$	$2,2 \times 10^{-16}$	$2,2 \times 10^{-16}$	$2,2 \times 10^{-16}$
R^2	0,6499	0,7272	0,7887	0,7957
$R^2_{adjusted}$	0,6485	0,7249	0,7861	0,7923

Source: The authors, 2022.

In Table 1, the coefficient of determination (R^2) is observed, which establishes percentual how much the model explains the variation in the average response of y. This result can be corrected for the loss of degrees of freedom with each term in the model, providing the adjusted R^2. The coefficients of determination assumed values between 0.6499 and 0.7957, ratifying the viability of the models found.

Table 2

Residuals summary analysis

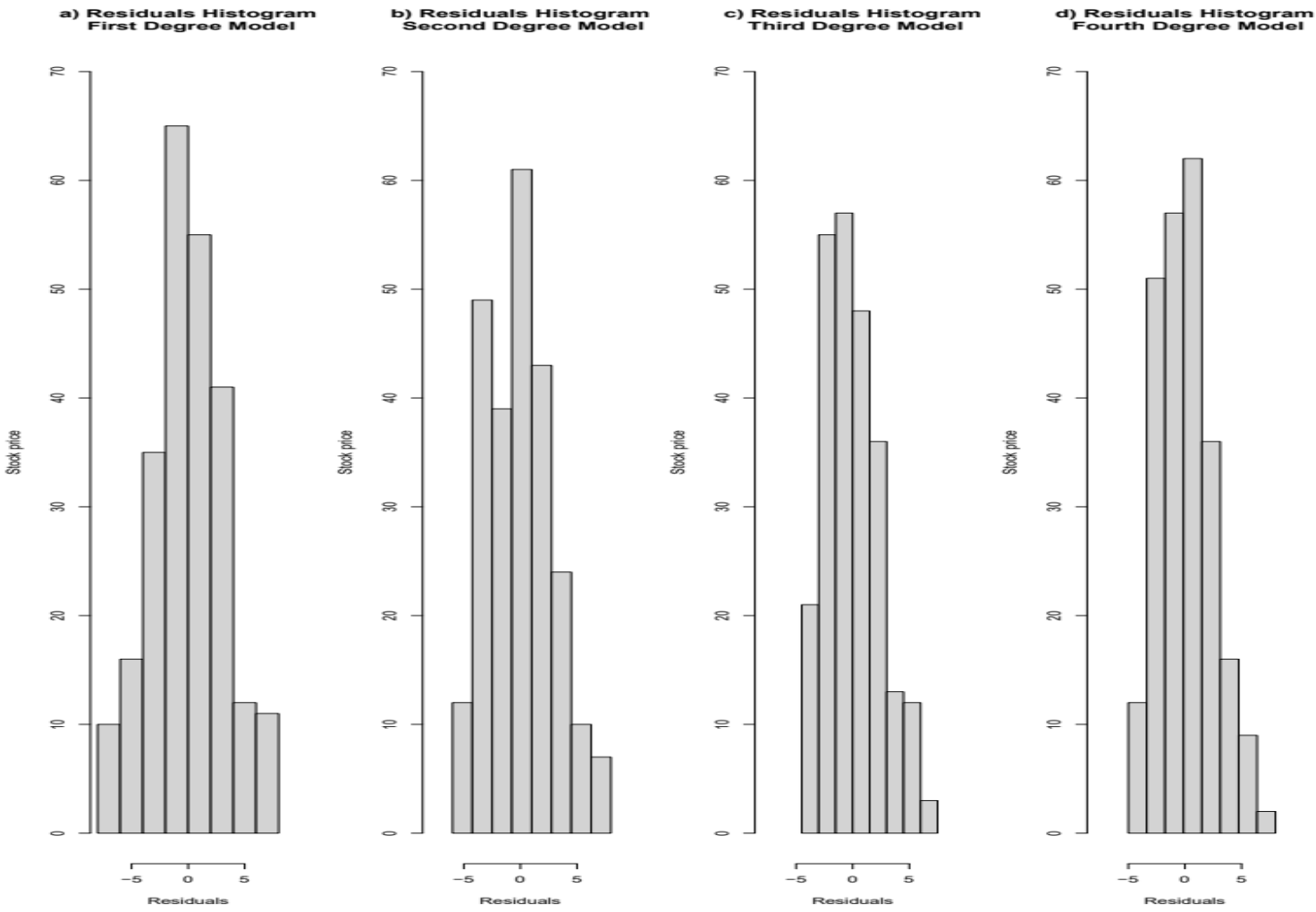
	First Degree Model	Second Degree Model	Third Degree Model	Fourth Degree Model
Minimum	-7,5001	-5,5723	-4,3818	-4,5023
1st quartile	-1,9909	-2,4851	-1,9321	-1,7925
Median	-0,1322	-0,0365	-0,4471	-0,0984
3rd quartile	2,0668	2,0717	1,5450	1,5337
Maximum	7,8424	7,3749	7,4479	7,6539
Asymetry	0,0990	0,3203	0,5681	0,5743
Kurtosis	-0,1267	-0,5367	-0,1689	0,0130

Source: The authors, 2022.

For the diagnostic analysis of the models, the residuals were calculated, which returned the values presented in Table 2. From this table, one can see that the central tendency, represented by the median, closest to zero was in the second-degree model; the first and second-degree models can be considered symmetrical, corroborated by the asymmetry coefficients. The kurtosis coefficients of the first, second, and third-degree models reveal a platykurtic curve, while the fourth-degree model is mesokurtic.

Figure 3

Four models residuals histograms

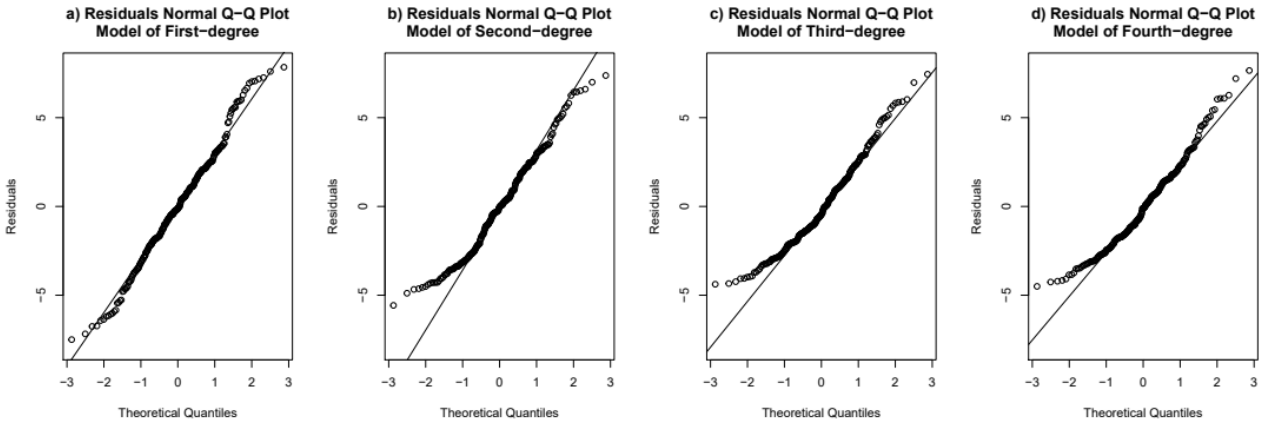


Source: The authors, 2023.

Visually, the residual histogram of the first-degree model (Figure 3a) appears to have a distribution closer to normal than the others.

Figure 4

Four models residuals normal Q-Q plot



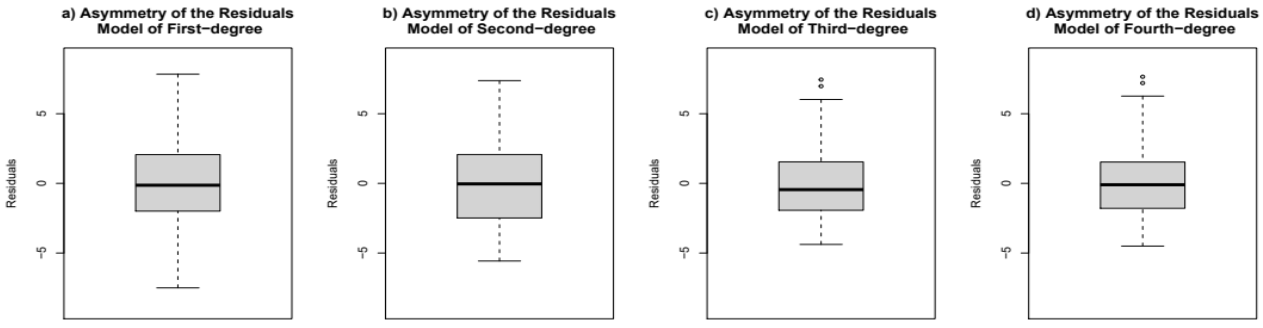
Source: The authors, 2022.

In Figure 4, the normal curves of the quantiles of the residuals for the models studied are plotted. It can be seen that the first one shows a better representation of normality. The others have their beginning and ending a little bit away of the normal quantile-quantile line. Specially the second-degree model is diverging, the others two diverge a little, but, after, converge in the end.

Figure 5 shows the boxplots of the residuals of the analyzed models, showing that the symmetry of the first-degree model is better than the others.

Figure 5

Four models residuals boxplot



Source: The authors, 2022.

The normality tests applied to validate the proposed models were Anderson-Darling, Jarque-Bera, Kolmogorov-Smirnov-Lilliefors, Doornik-Hansen, and Shapiro-Wilk. The results obtained can be seen in Table 3.

It is worth noticing that the Doornik-Hansen test results have provided much higher values and proved to be the most rigorous since all models failed in this test of normality with values very far from the 0.05 level.

Analysing Table 3, the p-value of the normality tests, with a level of significance 0.05, we can infer that the residuals of first-degree polynomial model are in a normal distribution, except for the Doornik-Hansen test. But with the results of the p-values of the other normality tests, the second, third, and fourth degrees polynomial models, we can infer that they are not.

Table 3

Results obtained from normality tests

Degree of polynomial	Measure	Test				
		KSL	JB	SW	AD	DH
1	Test statistics	0,0389	0,5643	0,9908	0,4281	25,5419
	Degrees of freedom	2	2	-	-	-
	p-value	0,4844	0,7542	0,1240	0,3091	$3,91x10^{-5}$
2	Test statistics	0,0691	7,1310	0,9770	1,2919	36,8127
	Degrees of freedom	2	2	-	-	-
	p-value	0,0065	0,0283	0,0005	0,0023	$1,97x10^{-7}$
3	Test statistics	0,0851	13,4720	0,9695	1,9244	48,8110
	Degrees of freedom	2	2	-	-	-
	p-value	0,0002	0,0012	$4,22x10^{-5}$	$6,39x10^{-5}$	$6,39x10^{-10}$
4	Test statistics	0,0785	13,4690	0,9712	1,6056	44,8972
	Degrees of freedom	2	2	-	-	-
	p-value	0,0009	0,0012	$7,28x10^{-5}$	0,0004	$4,18x10^{-9}$

Source: The authors, 2022.
Legend: KS-L = Kolmogorov-Smirnov-Lilliefors, JB = Jarque-Bera, SW = Shapiro-Wilk, AD = Anderson-Darling, DH = Doornik-Hansen.

It is worth noticing to the fact that the sample size can influence the results of the normality tests, as there is a tendency, for larger samples, to reject the null hypothesis, increasing the probability of type I errors. Mainly in tests that use the value of n in their test statistics, such as Jarque-Bera and Anderson-Darling.

According to Demir (2022, p. 397), the Kolmogorov-Smirnov-Lilliefors, Anderson Darling, Shapiro-Wilk, and Jarque-Bera tests are not affected by the sample size, only when the asymmetry and kurtosis coefficients are close to zero.

In the present study, a sample of 245 values was used and only the residuals of the first degree polynomial showed asymmetry (0.0984) and kurtosis (-0.1501) close to zero, satisfying Demir's(2022, p. 397) asymmetry and kurtosis criterion. In the residuals of the second degree polynomial, the asymmetry was relatively low (0.3184), but not the kurtosis (-0.5568). In the residuals of third and fourth degrees polynomials, asymmetries were high, respectively 0.56467 and 0.5708.

According to this author, the results obtained for these measures indicate that the interpretation of normality tests for polynomials of degree two, three, and four may have been influenced by pronounced asymmetry and kurtosis, not showing good performance.As mentioned, it is also essential to check the homoscedasticity and independence of the residuals. To this end, the Ljung-Box and ARCH tests were run, respectively. The Ljung-Box test returned a p-value of 2.2×10^{-16} for all models, therefore, much lower than 0.05, which showed evidence of dependence among residuals. The ARCH test also confirmed the heteroscedasticity effect in all models since the p-value of 2.2×10^{-16} is well below 0.05.

6 Concluding Remarks

The modeling of the TS trend of the closing price of VALE3 stock performed employing the Least Squares Method, fitting polynomial models of up to fourth-degree, found viable models, in which the results of the analysis of variance and the coefficient of determination were reasonable, although some did not meet the requirements of the method used.

The four tests, Jarque-Bera, Shapiro-Wilk, Anderson-Darling, and Kolmogorov-Smirnov-Lilliefors, considered in this analysis presented the same result in the normality evaluation, although they found different p-values. And the Doornik-Hansen normality test was an exception, presenting a

different result in the model of first degree polynomial. Besides, the Doornik-Hansen p-value was always the farthest from the other tests.

It was observed that the sample size can be a limitation to the inference of conclusions on the tests results. So, it is recommend use several tests to better analysis this desirable property when making a TS forecast.

It was observed too that the Jarque-Bera test always presented a higher p-value than the others, while the Doornik-Hansen test presented a lower one. This suggests that the Jarque-Bera test presents a tendency not to reject the null hypothesis associated with normality, resulting in a higher probability of occurrence of type 2 error. On the other hand, the Doornik-Hansen test would be showing a tendency to reject the null hypothesis, which could represent a higher probability of type 1 error.

It should be noted that failure to meet this requirement still allows good predictions to be obtained Bueno (2018). Because of this, if the goal of the TS analysis is to make predictions, the Jarque-Bera test seems to be the most suitable.

Finally, we observe that ARIMA (Autoregressive Independent Moving Average) or GARCH (Generalized Autoregressive Conditional Heteroscedasticity) models would better fit for this TS because it is a financial TS and quite volatile.

As future work, we consider studying the performance of normality tests with different sample sizes, as well as different intensities of asymmetry and kurtosis coefficients.

References

Abdi, H.; Molin, P. (2007). Lilliefors/Van Soest test of normality. *Encyclopedia of measurement and statistics*. pp. 540-544.

Adkins, L. C. (2018). *Using gretl for Principles of Econometrics*, Version 1.0411. 5ª. Oklahoma State University, Oklahoma, USA. Recuperada em 23 fevereiro, 2022 de http://www.learneconometrics.com/gretl/poe5/using_gretl_for_POE5.pdf.

Ahmad, F., & Sherwani, R. A. K. (2015). A power comparison of various normality tests. *Pakistan*

Journal of Statistics and Operation Research, pp. 331-345. https://doi.org/10.18187/pjsor.v11i3.845

Ahsanullah, M.; Kibria, B. M. G., & Shakil, M. (2014). *Normal and Student's t Distributions and Their Application*, Vol. 4. Paris: Atlantic Press.

Alves, D., & Pereda, P. C. (2018). *Econometria Aplicada*. Rio de Janeiro: Elsevier. Recuperada em 08 fevereiro, 2022 de https://integrada.minhabiblioteca.com.br/#/books/9788595156012/

Anderson, T. W. (2010). Anderson-Darling Tests of Goodness-of-Fit. *International encyclopedia of statistical science*, Vol. 1, pp. 52-54. https://doi.org/10.1007/978-3-642-04898-2_118

Anderson, T. W., & Darling, D. A. (1952). Asymptotic Theory of Certain "Goodness of Fit" Criteria Based on Stochastic Processes. *Annals of Mathematical Statistics*, Vol. 23, pp. 193-212. https://doi.org/10.1214/aoms/1177729437

Arshad, M., Rasool, M. T., & Ahmad, M. I. (2003). Anderson Darling and Modified Anderson Darling Tests for Generalized Pareto Distribution. *Pakistan Journal of Applied Sciences*, Vol. 3, nº. 2, pp. 85-88. https://doi.org/10.3923/jas.2003.85.88

Bueno, R. de L. da S. (2018). *Econometria de séries temporais*. 2ª. São Paulo: Cengage Learning. Recuperada em 08 fevereiro, 2022 de https://integrada.minhabiblioteca.com.br/#/books/9788522128259/

Cramér, H. On the composition of elementary errors. Skand Aktuarietids, v. 11, p. 13-74, 1928.

Das, K. R., & Imon, A. H. M. R. (2016). A Brief Review of Tests of Normality. *American Journal of Theoretical and Applied Statistics*. Vol. 5, nº. 1, 2016, pp. 5-12. https://doi.org/10.11648/j.ajtas.20160501.12

Demir, S. Comparison of normality tests in terms of sample sizes under different skewness and Kurtosis coefficients. *International Journal of Assessment Tools in Education*, Vol. 9, nº. 2, 2022, pp. 397-409.

Doornik, J. A., & Hansen, H. (1994). An Omnibus Test for Univariate and Multivariate Normality. *Transformation*, Vol. 2, pp. 2-17. https://doi.org/10.1111/j.1468-0084.2008.00537.x

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica: Journal of the econometric society*, pp. 987-1007. https://doi.org/10.2307/1912773

Farrell, P. J., Salibian-Barrera, M., & Naczk, K. (2007). On tests for multivariate normality and associated simulation studies. *Journal of statistical computation and simulation*, Vol. 77, nº. 12, pp. 1065-1080. https://doi.org/10.1080/10629360600878449

Ghasemi, A., & Zahediasl, S. (2012). Normality Tests for Statistical Analysis: A Guide for Non-Statisticians. *International Journal of Endocrinology and Metabolism*, Vol. 10, nº. 2, pp. 486-489. https://doi.org/10.5812/ijem.3505

Górecki, T., Horváth, L., & Kokoszka, P. (2020). Tests of normality of functional data. *International Statistical Review*, Vol. 88, nº. 3, pp. 677-697. https://doi.org/10.1111/insr.12362

Grace, A. W., & Wood, I. A. (2012). Approximating the tail of the Anderson–Darling distribution. *Computational Statistics and Data Analysis*. Vol. 56, nº. 12, pp. 4301–4311. https://doi.org/10.1016/j.csda.2012.04.002

Gross, J., & Ligges, U. *Package 'nortest'*. (2015) Recuperada em 07 fevereiro, 2022 de https://cran.r-project.org/web/packages/nortest/nortest.pdf.

Gujarati, D. N., & Porter, D. C. (2011). *Econometria Básica*. Tradução Denise Durante, Mônica Rosemberg, Maria Lúcia G. L. Rosa. 5ª. Porto Alegre: McGraw Hill Bookman. Recuperada em 15 setembro, 2021 de https://integrada.minhabiblioteca.com.br/#/books/9788580550511/.

Güner, B., Frankford, M. T., & Johnson, J. T. (2009). A study of the Shapiro–Wilk test for the detection of pulsed sinusoidal radio frequency interference. *IEEE transactions on Geoscience and Remote Sensing*, Vol. 47, nº. 6, pp. 1745-1751. https://doi.org/10.1109/TGRS.2008.2006906

Hill, R. C., Griffiths, W. E., & Judge, G. G. (2010). *Econometria*. Tradução Alfredo Alves de Farias. 3ª. São Paulo: Saraiva. Recuperado em 08 fevereiro, 2022 de https://integrada.minhabiblioteca.com.br/#/books/9788502109735/.

Hyndman, R. J., & Athanasopoulos, G. (2021) *Forecasting: principles and practice*. 3ª. Melbourne:

OTexts. Recuperado em 07 agosto, 2022 de https://otexts.com/fpp3.

Hyndman, R. J., Athanasopoulos, G., Bergmeir, C., Caceres, G., Chhay, L., O'hara-Wild, M., Petropoulos, F., Razbash, S., Wang, E., & Yasmeen, F. (2022). *forecast: Forecasting functions for time series and linear models*. R package version 8.16, Recuperado em 07 fevereiro, 2022 de https://pkg.robjhyndman.com/forecast/.

Jarque, C. M., & Bera, A. K. (1987). A Test for Normality of Observations and Regression Residuals. *International Statistical Review*, Vol. 55, nº. 2, pp.163-172. https://doi.org/10.2307/1403192

Joenssen, D. W., & Vogel, J. (2014). A power study of goodness-of-fit tests for multivariate normality implemented in R. *Journal of Statistical Computation and Simulation*, Vol. 84, nº. 5, pp. 1055-1078. https://doi.org/10.1080/00949655.2012.739620

Kaur, A., & Kumar, R. (2015). Comparative Analysis of Parametric and Nonparametric Tests. *Journal of Computer and Mathematics Sciences*, Vol. 6, pp. 336-342.

Leotti, V. B., Birck, A. R., & Riboldi, J. (2005). Comparação dos Testes de Aderência à Normalidade Kolmogorov-smirnov, Anderson-Darling, Cramer-Von Mises e Shapiro-Wilk por Simulação. *Anais do 11º Simpósio de Estatística Aplicada à Experimentação Agronômica*.

Lilliefors, H. W. (1967). On the Kolmogorov-Smirnov Test for Normality with Mean and Variance Unknown. *Journal of the American statistical Association*, Vol. 62, nº. 318, pp. 399-402. https://doi.org/10.1080/01621459.1967.10482916

Ljung, G. M., & Box, G. E. P. (1978). On a measure of lack of fit in time series models. *Biometrika*, Vol. 65, nº. 2, pp. 297-303. https://doi.org/10.1093/biomet/65.2.297

Lobato, I. N., & Velasco, C. (2004). A simple test of normality for time series. *Econometric Theory*, Vol. 20, nº. 4, pp. 671-689. https://doi.org/10.1017/S0266466604204030

Lopes, M. de M., Branco, V. T. F. C., & Soares, J. B. (2013). Utilização dos testes estatísticos de Kolmogorov-Smirnov e Shapiro-Wilk para verificação da normalidade para materiais de pavimentação. *Revista dos Transportes*, Vol. 21, nº. 1, pp. 59-66. https://doi.org/10.4237/transportes.v21i1.566

Mattos, V. L. D. de. (2004). *Identificação de Efeitos de Dispersão em Experimentos com Poucas Replicações*. Tese de doutorado. Universidade Federal de Santa Catarina.

Morettin, P. A., & Toloi, C. M. C. (2018). *Análise de séries temporais: Modelos lineares univariados*. 3ª. São Paulo: Edgard Blucher.

Nunes, G. S., Mattos, V. L. D. de, & Konrath, A. C. (2021). Considerações sobre teste de normalidade utilizados pelo software GRETL. In Gonçalves, M. C. S., & Jesus, B. G. (Org.). *Contemporânea*, Vol. 22, pp. 174-178.

Öztuna, D., Elhan, A. H., & Tüccar, E. (2006). Investigation of Four Different Normality Tests in Terms of Type 1 Error Rate and Power under Different Distributions. *Turkey Journal of Medical Science*, Vol. 36, nº. 3, pp. 171-176.

Peterson, B. G., Carl, P., Boudt, K., Bennett, R., Ulrich, J., Zivot, E., & Cornilly, D. (2020). *Package PerformanceAnalytics*. Recuperado em 15 setembro, 2021 de https://cran.r-project.org/web/packages/PerformanceAnalytics/PerformanceAnalytics.pdf.

Pino, F. A. (2014). A Questão da Não-Normalidade: uma revisão. *Revista de Economia Agrícola*, Vol. 61, nº. 2, pp. 17-33.

Pya, N., Voinov, V., Makarov, R., & Voinov, Y. (2016). *Package 'mvnTest'*. Recuperado em 21 fevererio, 2022 de https://cran.r-project.org/web/packages/mvnTest/mvnTest.pdf.

Razali, N. M., & Yap, B. W. (2011). Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. *Journal of Statistical Modeling and Analytics*, Vol. 2, nº. 1, pp. 21-33.

R Core Team (2019). R: *A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Recuperada em 08 fevereiro, 2022 de https://www.R-project.org/.

Schmidt, A. F., & Finan, C. (2018). Linear regression and the normality assumption. *Journal of clinical epidemiology*, Vol. 98, pp. 146-151. https://doi.org/10.1016/j.jclinepi.2017.12.006

Seier, E. (2002). Comparison of tests for univariate normality. *InterStat Statistical Journal*, Vol. 1, pp.

1-17.

Shapiro, S. S., & Wilk, M. B. (1965). An Analysis of Variance Test for Normality (complete samples). *Biometrika*, Vol. 52, pp. 591-611. https://doi.org/10.2307/2333709

Stephens, M. A. (1974). EDF statistics for goodness of fit and some comparisons. *Journal of the American statistical Association*, Vol. 69, nº. 347, pp. 730-737. https://doi.org/10.1080/01621459.1974.10480196

Trapletti, A., & Hornik, K. (2020) *tseries: Time Series Analysis and Computational Finance*. Recuperada em 15 setembro, 2021 de https://CRAN.R-project.org/package=tseries.

Tsay, R. (2013). *Package 'FinTS'*. Recuperada em 15 setembro, 2021 de http://www2.uaem.mx/r-mirror/web/packages/FinTS/FinTS.pdf.

Wickham, P. (2015). *Package 'normwhn.test'*. Recuperada em 23 fevereiro, 2022 de https://cran.r-project.org/web/packages/normwhn.test/normwhn.test.pdf.

Yap, B. W., & Sim, C. H. (2011). Comparisons of various types of normality tests. *Journal of Statistical Computation and Simulation*, Vol. 81, n. 12, pp. 2141-2155. https://doi.org/10.1080/00949655.2010.520163

Yiğit, S., & Mendeş, M. (2016). Usage of multidimensional scaling technique for evaluation performances of multivariate normality tests. *British Journal of Applied Science & Technology*, Vol. 16, pp. 1-8.

Zeileis, A., & Hothorn, T. (2002). Diagnostic Checking in Regression Relationships. R News, Vol. 2, Issue 3, pp. 7–10. Recupearada em 15 setembro, 2021 de https://CRAN.R-project.org/doc/Rnews/