



Comparative Study of the Finite Sample Properties of Ljung-Box and Box-Pierce in Time Series Econometrics

^a Dr Chakar Khan, ^b Imran Farooq, ^c Dr Mansoor Ahmed, and ^d Dr Manzoor Ahmed

a. Lecturer, Management Sciences Department, University of Balochistan, Quetta. (*Corresponding author*)

Email: chakarjan@yahoo.com

b. National Savings Officer, National Savings, Ministry of Finance, Islamabad.

imranfarooqbugti@yahoo.com

c. Academic Specialist, Higher Education Department, Government of Balochistan, Quetta.

qaumansoor@gmail.com

d. Pro-Vice Chancellor, University of Gwadar, Gwadar

economist.luams@gmail.com

ARTICLE DETAILS

History:

Received: 28-05-2025

Accepted: 04-05-2026

Available Online: 30-06-2026

Keywords:

Finite Sample,
Ljung Box,
Box Pierce,
Time Series

JEL Codes:

62M10,
C12
C15

ABSTRACT

Objective: This paper assesses the finite-sample properties of the Ljung-Box and Box-Pierce tests, specifically their autocorrelation detection and chi-square distribution.

Research Gap: The existing literature checks the asymptotic properties of the Ljung-Box and Box-Pierce tests and finite-sample properties across various sample sizes and autocorrelation levels.

Design/Methodology/Approach: The paper uses Monte Carlo simulations with sample sizes ranging from 15 to 1000 at different autocorrelation levels and reports the power of the test statistics and the Type I error.

The Main Findings: These results indicate that both tests are asymptotically chi-square distributed when sample sizes are small ($n < 100$). The Ljung-Box test follows a chi-square distribution much more closely than the Box-Pierce test. However, when the sample size is less than 150, the Box-Pierce test rejects the null hypothesis, leading to a Type I error. The Ljung-Box test has higher power at lower levels of autocorrelation and with smaller finite sample sizes. Finally, both tests improve in power and Type I error control as sample size increases.

Theoretical / Practical Implications of the Findings: The paper elucidates the finite-sample properties of the tests, including the chi-square distributional behaviour, the power of the test statistics, and the Type I error of both tests. The results help in choosing an appropriate autocorrelation diagnostic test for small and moderate sample sizes.

Originality/Value: This paper compares the small-sample properties of the "portmanteau" tests using Monte Carlo simulation and highlights their weaknesses.

The authors. Published by PJES, IUB. This is an open-access research paper under the Creative Commons Attribution-Non-Commercial 4.0



Recommended Citation:

Khan, C., Farooq, I., Ahmed, M. and Ahmed, M. Comparative Study of the Finite Sample Properties of Ljung-Box and Box-Pierce in Time Series Econometrics. *Pakistan Journal of Economic Studies*, 9(2), 98-104. Available at:

<https://journals.iub.edu.pk/index.php/pjes/article/view/4537>

Corresponding Author's email address: chakarjan@yahoo.com

1. Introduction

One important assumption of the classical linear regression model is that the error terms are uncorrelated. Because of autocorrelation, the ordinary least squares (OLS) assumption is violated, and OLS is no longer the best linear

unbiased estimator under the Gauss-Markov theorem (Plackett, 1949, 1950; Greene, 2018). Autocorrelation is an econometric problem that often leads to spurious results in time-series analysis. Hence, it is important to check whether a model performs well so we can predict accurately and understand how it works. The condition associated with residuals is that they are "white noise"; therefore, portmanteau tests are used to determine whether residuals are "white noise" or not. In large samples, the Box-Pierce and Ljung-Box portmanteau tests perform better under asymptotic theory, but their finite- and small-sample properties are less clear.

Existing studies comparing autocorrelation tests began in 1975. L'Esperance and Taylor (1975) compared four autocorrelation tests by using a Monte Carlo simulation. Likewise, Smith (1976) compared six autocorrelation tests, including the Durbin-Watson test, Durbin test, Durbin cumulated periodogram test, Schmidt test and Geary sign test. In the most recent studies, Uyanto (2020) used various lag structures to compare the power of five autocorrelation-detection test statistics. Similarly, Danioko et al. (2022) used a novel technique to compare test power across threshold sample sizes and lags. In contrast, Sun (2023) focused on significance levels and optimal lags in time-series econometrics. Finally, Huang (2024) addresses the type 1 error by using large-sample data across all scenarios for the Ljung-Box test.

However, this study examines the finite- and small-sample properties of the Box-Pierce and Box-Ljung Q statistics. According to Kwan and Sim (2005), the properties exhibited by both of the tests are:

- (1) These tests are asymptotic statistics but require a large sample size to be accurate. For small sample sizes, the Box-Pierce test rejects the null hypothesis too infrequently, and the Ljung-Box test rejects it too often.
- (2) They have an asymptotic chi-square distribution with K degrees of freedom.

We compared and contrasted the Box-Pierce and Box-Ljung Q statistics with respect to the properties above by using a Monte Carlo simulation in MATLAB to generate 56 datasets with varying autocorrelation levels, assess the power of the test statistics at different autocorrelation levels, and evaluate whether the test statistics follow the chi-square distribution. Numerous statistical software packages are available for simulating various tests, such as R, MATLAB, Minitab, EViews and Stata. However, we used MATLAB to conduct simulations to assess the power of both tests across varying significance levels, autocorrelation levels and sample sizes.

2. Description of portmanteau tests

In this section, we discuss both tests and compare them with the literature and its findings.

2.1. Box-Pierce

Box-Pierce developed a portmanteau test in 1970.

$$Q = n \sum_{k=1}^h r_k^2 \quad (2.1)$$

2.2. Ljung-Box

The Ljung-Box test was developed in 1978 to improve the performance by modifying the pioneer portmanteau test.

$$Q = n(n+2) \sum_{k=1}^h \frac{r_k^2}{n-k} \quad (2.2)$$

From both equations 1 and 2, n shows the sample size, h shows the number of lags that are tested and r_k^2 indicates sample autocorrelation at the given lag. Under the null hypothesis, the Q statistic for both tests indicate no autocorrelation and follows a chi-square distribution.

When comparing the two tests, the Box-Pierce test often fails to reject the null hypothesis with small sample sizes, whereas the Ljung-Box test has a higher rejection rate. Both tests are asymptotically chi-squared distributed but require a large sample size to be accurate (Kwan & Sim, 2005). For low levels of autocorrelation, both tests will perform poorly (Uyanto, 2020). Moreover, Danioko et al. (2022) state that the degree-of-freedom problem intensifies when the number of lags (K) exceeds the number of observations (n). The Box-Pierce test has a location-based bias. The normalisation procedure of the box pierce test is not suitable for the iid series with an unknown mean, ultimately leading to poor performance empirically, as expected (Dufour & Roy, 1986).

Hassani and Yeganegi (2019) proposed using optimal lags for the Ljung-Box test. The choice of optimal lags depends on the time series length and the test level. The most commonly used methods do not provide a suitable lag value. Moreover, Uyanto (2020) used a Monte Carlo simulation to compare the statistical power of five tests: Box-Pierce, Ljung-Box, Breusch-Godfrey, Durbin-Watson, and Runs. Results revealed that the Durbin-Watson test outperforms the tests for the dependent variable without lags, and the Breusch-Godfrey test outperforms all other

tests for the dependent variable with lags. Finally, Dalla, Giraitis and Phillips (2020) compared the standard correlogram and the cumulative Ljung-Box tests when the series are not iid (i.e., independent and identically distributed). They introduced a new cumulative test, which remains unaffected by unknown dependence and heteroskedasticity in the series.

3. Simulation Setup

We extended Uyanto's (2020) simulation methods by using Monte Carlo to generate datasets with sample sizes of 15, 30, 60, 100, 150, 200, 500, and 1000 observations. We constructed eight data sets for each sample size and autocorrelation level: 0, 0.2, 0.4, 0.6, 0.8, 0.9, and 0.99. The observations of a data set consist of a variable "residuals", which is constructed as an Autoregressive (1) series. The "residuals" are created as follows:

$$residuals_t = \rho * residuals_{t-1} + \sqrt{(1 - \rho^2)} * \varepsilon \quad (3.1)$$

ρ denotes the autocorrelation level. The term ε is standard and is normally distributed. $\sqrt{(1 - \rho^2)}$ is a scaling factor that ensures variance stationarity. In this way, we generated 56 datasets, each with a distinct sample size and autocorrelation level. We repeated this simulation 10,000 times. We computed the Box-Pierce Q-statistic and the Ljung-Box Q-statistic for each dataset. For both tests, we set the lag selection criterion to 5 lags and evaluated both statistics at the 95% confidence level. We computed the power of both tests and performed a Wald test to determine whether the difference between them was significant. We also performed power tests at 10%, 2.5%, and 1% significance levels.¹ We also performed a Wald test to determine whether the observed type I error rate is significantly different from 1%, 2.5%, 5%, and 10%. To show that, for small sample sizes, the Ljung-Box test better resembles a chi-square distribution than the Box-Pierce test, we plot the Q-statistics in a histogram. (See Figure 1).

4. Discussion Results

Tables 1 to 4 show the statistical power results. The first column ($\rho = 0$) gives the test's type I error. The autocorrelation levels are discussed horizontally, ranging from 0 to 0.99, while the sample sizes are arranged vertically from smallest to largest, ranging from 5 to 1000. The statistical power of both tests is presented across autocorrelation levels and sample sizes at the 1%, 2.5%, 5%, and 10% significance levels.

First, we discuss simulation results for both the Ljung-Box and Box-Pierce tests, with significance levels decreasing from 5% to 1% as sample sizes increase, to evaluate their finite-sample properties. The expected type I error is 5%, since the table shows results at the 5% significance level. The table shows that the Ljung-Box type I error is closer to 5% at low sample sizes ($n < 100$), indicating that its distribution more closely resembles a chi-square distribution than the Box-Pierce statistic. (As shown by Ljung and Box, 1978). Figure 1 also shows that the Ljung-Box Q-statistic fits a chi-square distribution better than the Box-Pierce statistic. This is in line with the findings that the Box Pierce test theoretical mean is lower than the mean, k , in a chi-square (k) distribution (Ljung & Box, 1978; Kwan and Sim (2005). Although the Ljung-Box test more closely resembles a chi-square distribution, the observed type I error is significantly higher than 5% for sample sizes smaller than 150. This is in line with the literature, which finds that the Ljung-Box test rejects the null hypothesis too often for moderate sample sizes (Danioko, Sidy, et al., 2022).

At the 10% significance level, the Box-Pierce type I error deviates even more from the expected 10%. For small sample sizes, the Ljung-Box test performs better than the Box-Pierce test at the 10% significance level, although its type I error rate remains higher than 10%. At lower significance levels, such as 2.5% and 1%, the Ljung-Box test does not perform better than the Box-Pierce test. The Ljung-Box test rejects the null hypothesis too often, while the Box-Pierce test rejects it too little.

Turning to differences in autocorrelation levels as the sample size increases. If the null hypothesis is false ($\rho > 0$), the Ljung-Box statistic has higher power than the Box-Pierce statistic at low autocorrelation levels and low sample sizes. This is because, by construction, the Ljung-Box Q-statistic is higher than the Box-Pierce test statistic for finite sample sizes. Both tests perform poorly at detecting low autocorrelation levels (below 0.4) in small samples. At an autocorrelation level of 0, the power of the Ljung-Box Q statistic is likely not significantly different from the power of the Box-Pierce test statistic. However, for autocorrelation levels from 0.2 to 0.4, the power of the Ljung-Box Q statistic is significantly different from the power of the Box-Pierce test statistic. Overall, the Ljung-Box test has

¹ The results will be furnished upon request.

significantly higher power than the Box-Pierce test for small sample sizes and low autocorrelations. Uyanto (2020) also observes this. For larger sample sizes, the tests' power increases, consistent with Ljung's (1986) findings. Across significance levels (10%, 5%, 2.5%, and 1%), the main findings remain the same: both tests perform poorly at low autocorrelation levels and with small sample sizes.

Table 1: Comparison of the test power of the Box-Pierce Test and the Ljung-Box Test for Autocorrelation at 1% significance level

Autocorrelation Tests	Autocorrelation Levels	Sample Size							
		15	30	60	100	150	200	500	1000
BP	0	0.0038	0.0066	0.0094	0.0111	0.0095	0.0088	0.0087	0.0097
LB		0.0205	0.014	0.015	0.0133	0.0124	0.0096	0.0091	0.0102
BP	0.2	0.0016	0.0173	0.0478	0.1056	0.1984	0.3003	0.8484	0.9971
LB		0.0218	0.035	0.0613	0.1206	0.213	0.313	0.8524	0.9974
BP	0.4	0.006	0.0898	0.3553	0.6982	0.9157	0.9814	1	1
LB		0.0433*	0.1371*	0.3975*	0.7204*	0.9226*	0.9828	1	1
BP	0.6	0.0246	0.3599	0.8631	0.9929	1	1	1	1
LB		0.1107*	0.4486*	0.8838*	0.9939	1	1	1	1
BP	0.8	0.0877	0.7383	0.9933	1	1	1	1	1
LB		0.2424*	0.7966*	0.9952	1	1	1	1	1
BP	0.9	0.1426	0.8656	0.9988	1	1	1	1	1
LB		0.3126*	0.9003*	0.999	1	1	1	1	1
BP	0.99	0.1941	0.9225	0.9999	1	1	1	1	1
LB		0.3849*	0.9431*	0.9999	1	1	1	1	1

Note: The autocorrelation levels are given horizontally, then followed by the specific autocorrelation test type

The sample sizes are expressed vertically from the smallest sample size to the largest sample size

BP shows the Box-Pierce test

LB shows the Ljung-Box Test

* on the power of the Test indicates the significant difference between LB test statistics and BP test statistics at the 1% significance level

Source: Authors' Compilation

Table 2: Comparison of the test power of the Box-Pierce Test and the Ljung-Box Test for Autocorrelation at 2.5% significance level

Autocorrelation Levels	Autocorrelation Tests	Sample Size							
		15	30	60	100	150	200	500	1000
0	BP	0.0087	0.0131	0.0207	0.0247	0.0243	0.0222	0.0228	0.0254
	LB	0.0395	0.0307	0.0294	0.0306	0.0302	0.0254	0.0239	0.0263
0.2	BP	0.0065	0.0356	0.0835	0.1731	0.2903	0.4183	0.9094	0.9993
	LB	0.0444	0.0638	0.1043	0.189	0.3094	0.4327	0.9109	0.9993
0.4	BP	0.0177	0.1452	0.4641	0.7941	0.9531	0.9918	1	1
	LB	0.0841*	0.2081*	0.5101*	0.8141*	0.9564	0.993	1	1
0.6	BP	0.0593	0.4683	0.9121	0.9968	1	1	1	1
	LB	0.1801*	0.5484*	0.9263*	0.9974	1	1	1	1
0.8	BP	0.1739	0.8129	0.9969	1	1	1	1	1
	LB	0.3413*	0.8562*	0.9976	1	1	1	1	1
0.9	BP	0.2404	0.9095	0.9994	1	1	1	1	1
	LB	0.4198*	0.9306*	0.9996	1	1	1	1	1
0.99	BP	0.3098	0.9484	0.9999	1	1	1	1	1
	LB	0.49*	0.9625*	0.9999	1	1	1	1	1

Note: on the power of the Test indicates the significance of the difference between LB test statistics and BP test statistics at the 2.5% significance level

Source: Authors' Compilation

Table 3: Comparison of the test power of the Box-Pierce Test and the Ljung-Box Test for Autocorrelation at 5% significance level

Autocorrelation Levels	Autocorrelation Tests	Sample Sizes							
		15	30	60	100	150	200	500	1000
0	BP	0.0182	0.0268	0.0406	0.0453	0.0465	0.0459	0.0465	0.0509
	LB	0.0646	0.0558	0.0557	0.0531	0.0531	0.0521	0.0478	0.052
0.2	BP	0.0165	0.0589	0.1273	0.2376	0.3894	0.5233	0.9455	0.9996
	LB	0.0696*	0.0958*	0.1527*	0.2632*	0.4073*	0.5366*	0.947*	0.9996
0.4	BP	0.0374	0.2076	0.5679	0.863	0.9746	0.9964	1	1
	LB	0.1201*	0.2817*	0.6081*	0.8773*	0.9777	0.9965	1	1
0.6	BP	0.1085	0.5763	0.9497	0.9986	1	1	1	1
	LB	0.2501*	0.6443*	0.9585*	0.9988	1	1	1	1
0.8	BP	0.3659	0.9116	0.9992	1	1	1	1	1
	LB	0.5466*	0.9366*	0.9994	1	1	1	1	1
0.9	BP	0.5855	0.9754	0.9997	1	1	1	1	1
	LB	0.7259*	0.9818*	0.9997	1	1	1	1	1
0.99	BP	0.7545	0.9925	1	1	1	1	1	1
	LB	0.8412*	0.9942*	1	1	1	1	1	1

Note: * on the power of the Test indicates a significant difference between LB test statistics and BP test statistics at the 5% significance level

Source: Authors' Compilation

Table 4: Comparison of the test power of the Box-Pierce Test and the Ljung-Box Test for Autocorrelation at 10% significance level

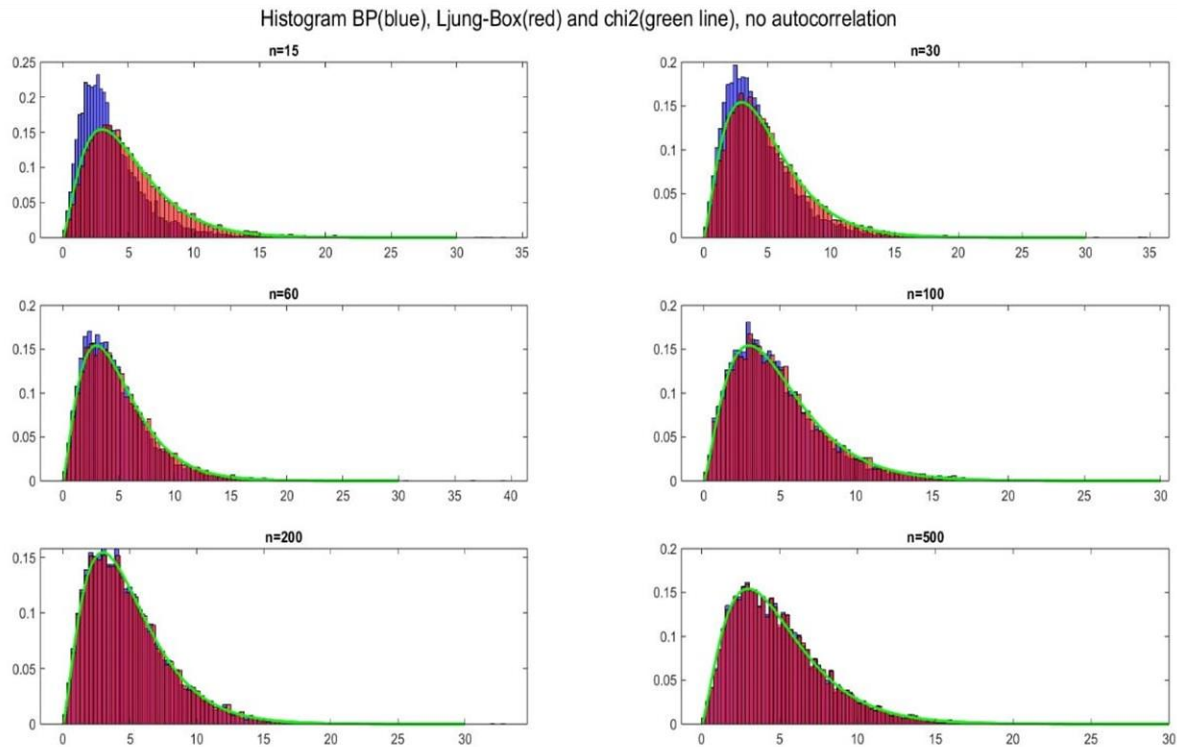
Autocorrelation Tests	Autocorrelation Levels	sample size							
		15	30	60	100	150	200	500	1000
BP	0	0.0366	0.0575	0.0803	0.087	0.0908	0.094	0.095	0.0975
LB		0.1169*	0.1033*	0.1053*	0.1044*	0.1021*	0.1011*	0.0978	0.0991
BP	0.2	0.0402	0.1044	0.2046	0.3479	0.5122	0.6419	0.9715	1
LB		0.1268*	0.1661*	0.2409*	0.3721*	0.5304*	0.6525*	0.9724	1
BP	0.4	0.0837	0.3108	0.6792	0.9174	0.9872	0.9989	1	1
LB		0.2073*	0.3946*	0.7163*	0.9271*	0.9884	0.999	1	1
BP	0.6	0.1902	0.6619	0.9684	0.9993	1	1	1	1
LB		0.3492*	0.7315*	0.973*	0.9993	1	1	1	1
BP	0.8	0.371	0.9077	0.9991	1	1	1	1	1
LB		0.5456*	0.9314*	0.9994	1	1	1	1	1
BP	0.9	0.4542	0.9586	0.9998	1	1	1	1	1
LB		0.617*	0.9696*	0.9998	1	1	1	1	1
BP	0.99	0.527	0.9774	0.9999	1	1	1	1	1
LB		0.6723*	0.9835*	1	1	1	1	1	1

Note: * on the power of the Test indicates a significant difference between LB test statistics and BP test statistics at the 10% significance level.

Source: Authors' Compilation

First, we look at the histogram given below: the sample size is 5; the Ljung-Box statistic follows a chi-square distribution, but the Box-Pierce statistic does not. Let's increase the sample size to 30 and 60; the results match the distribution for a sample size of 5. Increasing the sample size improves the fit of the chi-square distribution in the Box-Pierce test. For a sample size of 100, the Box-Pierce statistic is approximately chi-square distributed. Finally, for sample sizes of 200 and 500, the Box-Pierce statistic follows a chi-square distribution. In conclusion, increasing sample size improves the chi-square distribution of the Box-Pierce statistic, but the Ljung-Box statistic follows a chi-square distribution by construction; thus, increasing sample size does not affect its chi-square distribution.

Figure 1: Histogram of the Q-statistic of the Ljung Box test with 5 lags (in Red) and a histogram of the Box-Pierce Q statistic with 5 lags are plotted against the Chi2(5) probability density function.



Source: Authors' Compilation

5. Conclusion

The research findings indicate that the Ljung-Box (LB) test outperforms the Box-Pierce (BP) test for small, finite samples because its distribution aligns better with the chi-square distribution. However, both tests exhibit distinct biases: for small sample sizes, the Box-Pierce test tends to reject the null hypothesis too infrequently, whereas the Ljung-Box test rejects it more often at small and moderate sample sizes (below 150).

References

- Box, G. E., & Pierce, D. A. (1970). Distribution of residual autocorrelations in autoregressive-integrated moving average time series models. *Journal of the American statistical Association*, 65(332), 1509–1526.
<https://www.tandfonline.com/doi/abs/10.1080/01621459.1970.10481180>
- Danioko, S., Zheng, J., Anderson, K., Barrett, A., & Rakovski, C. S. (2022). A novel correction for the adjusted box-pierce test. *Frontiers in Applied Mathematics and Statistics*, 8, 873746.
<https://doi.org/10.3389/fams.2022.873746>
- Dufour, J. M., & Roy, R. (1986). Generalized portmanteau statistics and tests of randomness. *Communications in Statistics-Theory and Methods*, 15(10), 2953-2972.
https://www.tandfonline.com/doi/abs/10.1080/03610928608829288?casa_token=KZx0EtzNIUsAAAAA:4UZbGTQTOUljsAGFvawzMcb5ZqY_ns2EubCLF45Cy9H9DqIX6D9-9gKiG34bkjdqcfXh4NVYjKrV
- Giraitis, L. (2020). Robust Tests for White Noise and Cross-Correlation. *Econometric Theory*. Cambridge University Press (CUP)
- Greene, W. H. (2018). *Econometric Analysis*. Pearson Education, Inc., New York, N.Y., 8th edition
- Hassani H, Yeganegi MR (2019) Sum of squared ACF and the Ljung-Box statistics. *Physica A* 520:81–86
<https://www.sciencedirect.com/science/article/abs/pii/S0378437118315413>
- Huang (2024). A Novel Correction for the Multivariate Ljung-Box Test. [Published Master's thesis] Chapman University.
- Kwan, A. C., Sim, A. B., & Wu, Y. (2005). A comparative study of the finite-sample performance of some portmanteau tests for randomness of a time series. *Computational statistics & data analysis*, 48(2), 391-413.
<https://doi.org/10.1016/j.csda.2004.01.003>

- L'Esperance, W. L. & Taylor, D. (1975). The power of four tests of autocorrelation in the linear regression model. *Journal of Econometrics*, 3(1):1 –21.
<https://www.sciencedirect.com/science/article/abs/pii/0304407675900627>
- Ljung, G. M. (1986). Diagnostic testing of univariate time series models. *Biometrika*, 73(3), 725-730.
<https://doi.org/10.1093/biomet/73.3.725>
- Ljung, G. M., & Box, G. E. (1978). On a measure of lack of fit in time series models. *Biometrika*, 65(2), 297-303.
<https://doi.org/10.1093/biomet/65.2.297>
- Oukhouya, H., & Himdi, K. E. (2023). A comparative study of ARIMA, SVMs, and LSTM models in forecasting the Moroccan stock market. *International Journal of Simulation and Process Modelling*, 20(2), 125-143.
<https://www.inderscienceonline.com/doi/abs/10.1504/IJSPM.2023.136481>
- Plackett, R. L. (1949). A historical note on the method of least squares. *Biometrika*, 36(3/4):458–460.
<https://www.jstor.org/stable/2332682>
- Plackett, R. L. (1950). Some theorems in least squares. *Biometrika*, 37(1/2):149–157.
<https://www.jstor.org/stable/2332158>
- Smith, V. K. (1976). The estimated power of several tests for autocorrelation with non-first-order alternatives. *Journal of the American Statistical Association*, 71(356), 879-883.
<https://www.tandfonline.com/doi/abs/10.1080/01621459.1976.10480962>
- Sun, Q. (2023). Comparative Study of the Ljung-Box Test and Its Improved Methods. *Advances in Applied Mathematics*, 12, 1884.
https://pdf.hanspub.org/AAM20230400000_30847497.pdf
- Uyanto, S. S. (2020). Power comparisons of the five most commonly used autocorrelation tests. *Pakistan Journal of Statistics and Operations Research*, 16(1), 119-130.
<http://dx.doi.org/10.18187/pjsor.v16i1.2691>

Acknowledgments:

The authors thank everyone who supported and provided continuous guidance in this study.

Disclosure statement:

The author does not report any potential conflict of interest.

Disclaimer:

Any views and opinions which are expressed in this research paper are those of the author alone and also do not reflect the views of any institute.
