

Perform a Ljung-Box Test in Python

Authored by
Mohammed looti

November 7, 2025

RECOMMENDED CITATION

Mohammed looti (2025). *Perform a Ljung-Box Test in Python*. PSYCHOLOGICAL STATISTICS. Retrieved from <https://statistics.arabpsychology.com/?p=12016>

The **Ljung-Box test** is recognized as an indispensable diagnostic instrument within the field of [time series](#) analysis. Its core function is to rigorously evaluate whether a sequence of observations is independently distributed--that is, whether all systematic dependence has been removed--or if there remains a statistically significant level of [autocorrelation](#) across a range of specified lags. This powerful test is most frequently employed post-modeling, specifically applied to the **residuals** of a fitted model, such as an ARIMA or ARMA structure, ensuring that the model has successfully accounted for all non-random, systematic dependencies present in the historical data.

Determining whether significant **autocorrelation** persists in the model's residuals is paramount for validating the overall adequacy of any statistical forecasting structure. If these residuals exhibit structural patterns or correlation, it provides strong evidence that the model itself is incomplete; systematic information remains unexplained, which can lead directly to biased forecasts, inefficient parameter estimates, and ultimately, unreliable statistical inferences. Therefore, passing the Ljung-Box test is a necessary checkpoint before deploying a time series model for prediction.

This comprehensive guide is designed to detail the statistical underpinnings of the **Ljung-Box test** and provide a practical, step-by-step walkthrough for its efficient implementation. We will utilize the robust capabilities of the industry-standard [statsmodels](#) library within the [Python](#) programming environment, illustrating its application through a real-world modeling scenario.

The Foundational Principles of the Q-Test

The Ljung-Box test, often generically referred to as the Q-test, serves as a formal, omnibus test for the absence of serial correlation up to a specified maximum lag (k). Unlike simple diagnostic checks that examine the correlation at only individual lags, the Ljung-Box procedure aggregates the squared autocorrelations across several lags into a single test statistic. This holistic approach provides a more robust and comprehensive view of residual independence. It represents a critical, non-negotiable step in the validation framework following the estimation of any dynamic [time series](#) model.

A central assumption underlying the effectiveness of statistical models like ARIMA is that the model [residuals](#)--which represent the unexplained deviations between the actual observed values and the model's predictions--should be statistically indistinguishable from a [white noise](#) process. By definition, a white noise process exhibits a constant mean, constant variance, and zero autocorrelation at all time lags. The Ljung-Box test provides the statistical evidence needed to confirm whether the model successfully reduced the error term to this ideal white noise state.

Consequently, the primary goal of applying this diagnostic test is to ascertain if the unexplained variance still retains any predictable, structured patterns. If the test statistic suggests that significant correlation remains, it signals a fundamental failure in the model specification. In such cases, the model must be refined, potentially by increasing the order of the Autoregressive (AR) or

Moving Average (MA) components, or by incorporating external explanatory variables that were previously omitted.

Defining the Null and Alternative Hypotheses

As is standard practice across all statistical inference methods, the Ljung-Box test operates within the framework of [hypothesis testing](#). This method requires the formal statement of both a null hypothesis (H_0) and an alternative hypothesis (H_A), which together frame the statistical question regarding the independence of the residual errors being examined.

The formal statements for the Ljung-Box test are defined as follows:

H_0 (Null Hypothesis): The residuals are independently distributed; the autocorrelations are zero for all lags up to the maximum specified lag (k).

H_A (Alternative Hypothesis): The residuals are not independently distributed; they exhibit statistically significant serial correlation at one or more lags up to k .

In the context of model diagnostics, the successful outcome is to **fail to reject the null hypothesis** (H_0). This favorable result indicates that the chosen model has effectively captured and modeled the dependencies in the original data series, leaving behind only uncorrelated noise as the residual error. Conversely, rejecting the null hypothesis implies that the model is inadequate because important information remains correlated within the error term.

The critical decision to reject or retain H_0 is governed by comparing the computed [p-value](#) against a predefined significance level, typically denoted as α (alpha), with a conventional value of 0.05. If the p-value is greater than α , we conclude that there is insufficient evidence to reject the null hypothesis, suggesting the residuals are likely independent. If, however, the p-value is less than α , we must reject H_0 , confirming that significant [autocorrelation](#) persists, thereby necessitating a re-evaluation of the time series model.

Executing the Ljung-Box Test using Python's statsmodels

To implement the Ljung-Box test efficiently within [Python](#), we rely on the dedicated statistical capabilities provided by the [statsmodels](#) library. This package is universally recognized as the standard toolkit for advanced time series analysis and regression modeling. The specific function required for this diagnostic test is `acorr_ljungbox()`, which is located within the library's diagnostics module.

The structure and syntax for calling this function are highly intuitive, requiring the input data series (which must typically be the array of residuals derived from a fitted model) and the specification of the maximum number of lags to be included in the test calculation.

The function signature is defined as:

```
acorr_ljungbox(x, lags=None)
```

The crucial arguments are defined as follows:

x: This mandatory argument represents the data series undergoing the test. When performing model validation, this input must always be the array of [residuals](#) generated by the time series model.

lags: This argument specifies the maximum number of lags (or a list of specific lags) at which the combined test statistic should be computed. The selection of an appropriate lag value is fundamental to the test's sensitivity and is addressed in detail in the final section.

The output generated by the `acorr_ljungbox()` function yields two essential metrics: the calculated test statistic (often labeled `lb_stat`) and the corresponding [p-value](#) (`lb_pvalue`). It is this p-value that serves as the basis for making the final conclusion regarding the underlying **hypothesis testing** procedure.

Case Study: Diagnosing ARMA Model Residuals

To demonstrate the practical application of the Ljung-Box test, we will utilize the built-in "SUNACTIVITY" dataset, accessible through the [statsmodels](#) library, which tracks annual sunspot activity. The test is always performed on residuals, so our initial step involves fitting a time series model--here, a simple [ARMA model](#)--to generate the error terms we need to diagnose.

The subsequent code snippet illustrates the process: loading the time series data, fitting a straightforward ARMA(1,1) model, and then executing the Ljung-Box test on the resulting residuals. For this initial test run, we arbitrarily select a maximum lag value of 5, which instructs the test to analyze the collective autocorrelation across the first five time periods.

```
import statsmodels.api as sm
```

```
#load data series
```

```
data = sm.datasets.sunspots.load_pandas().data
```

```
#view first ten rows of data series
```

```
data
```

```
YEAR SUNACTIVITY
```

```
0 1700.0 5.0
```

```
1 1701.0 11.0
```

```
2 1702.0 16.0
```

```
3 1703.0 23.0
```

```
4 1704.0 36.0
```

```
#fit ARMA model to dataset
```

```
res = sm.tsa.ARMA(data, (1,1)).fit(dispatch=-1)
```

```
#perform Ljung-Box test on residuals with lag=5
```

```
sm.stats.acorr_ljungbox(res.resid, lags=, return_df=True)
```

```
lb_stat lb_pvalue
```

```
5 107.86488 1.157710e-21
```

The output for the lag 5 test provides a significant test statistic (`lb_stat`) of **107.86488** and, more critically, a minuscule **p-value** (`lb_pvalue`) of 1.157710×10^{-21} . Given that this p-value is astronomically small--far below the standard significance threshold of $\alpha=0.05$ --we are compelled to **reject the null hypothesis** (H_0).

This diagnostic failure confirms that the residuals generated by the simple ARMA(1,1) model are not independent; they retain highly significant serial correlation. This result suggests that the model is fundamentally inadequate for accurately capturing the complex, cyclical behavior inherent in the Sunspot activity data, and a far more sophisticated structure (such as a higher-order ARIMA or a seasonal SARIMA model) is required to eliminate the remaining correlation.

The Critical Role of Lag Selection

The choice of the maximum lag value (k) is arguably the most critical operational decision when performing the Ljung-Box test. This choice represents a trade-off between sensitivity and power. If k is set too small, the test may possess insufficient power to detect important correlations that exist at longer time horizons. Conversely, if k is excessively large, the test may become overly sensitive to random, non-systematic variation, reducing its reliability as a genuine diagnostic tool.

While various guidelines exist, a frequently cited rule of thumb for general data series suggests setting k to be approximately $\min(\frac{N}{4}, 10)$, where N is the total number of observations. However, when the Ljung-Box test is used specifically for model diagnostics on residuals, it is often recommended to choose k slightly larger than the sum of the orders of the fitted model components (e.g., for an ARMA(p, q) model, k should be chosen such that $k > p+q$).

To illustrate the impact of this choice, we can rerun the Ljung-Box test on the same residuals, but this time using a higher lag value, such as 20:

```
#perform Ljung-Box test on residuals with lag=20  
sm.stats.acorr_ljungbox(res.resid, lags=, return_df=True)
```

```
lb_stat lb_pvalue  
20 343.634016 9.117477e-61
```

By increasing the lag to 20, the test statistic escalates dramatically to **343.634016**, and the corresponding [p-value](#) shrinks even further to 9.117477×10^{-61} . This outcome strongly reinforces the initial diagnosis: the residual [autocorrelation](#) is deeply significant and spans a wide range of time periods. The persistence of extremely low p-values across multiple, reasonable lag choices serves as the most compelling evidence that significant, predictable structure remains locked within the residual series, confirming the need for comprehensive model refinement.