How to Test for Heteroskedasticity with the Breusch-Pagan Test in R

Authored by

stats writer

December 27, 2025

The Breusch-Pagan test is a fundamental diagnostic tool used in statistical modeling, specifically designed to detect the presence of heteroskedasticity within a linear regression model. Heteroskedasticity, often referred to as unequal variance, violates one of the core assumptions of the standard Ordinary Least Squares (OLS) estimation method: that the variance of the error terms (residuals) is constant across all levels of the independent variables. When this assumption, known as homoskedasticity, is violated, the resulting statistical inferences drawn from the model may be compromised.

The goal of the Breusch-Pagan test is to formally assess whether the squared residuals from the regression are systematically related to the predictor variables. If such a relationship exists, it provides strong evidence against the null hypothesis of homoskedasticity. Performing this diagnostic check is crucial for ensuring the reliability of parameter estimates and subsequent hypothesis testing. Failure to detect and correct for heteroskedasticity can lead to incorrect p-value calculations and unreliable confidence intervals, even though the coefficient estimates themselves remain unbiased. This diagnostic is efficiently performed in R using the lmtest package.

The command to install the necessary package is **install.packages("lmtest")**. Once the lmtest package is installed and loaded, the Breusch-Pagan test is executed using the **bptest()** function. This function requires the previously fitted regression model object as its primary input. The output yields a test statistic and a corresponding p-value, which dictates the decision regarding the null hypothesis of constant variance.

Introduction to Heteroskedasticity and the Breusch-Pagan Test

Linear regression models rely on several key assumptions regarding the distribution and behavior of the residuals. Among the most critical is the assumption of constant variance, or homoskedasticity. This condition implies that the spread of the residuals around the regression line remains consistent, regardless of the values taken by the explanatory variables. Conversely, heteroskedasticity occurs when the variance of the error terms changes systematically with the predictors, often resulting in larger errors for larger predicted values or vice versa. Common causes include data aggregation issues, the presence of outliers, or simply modeling phenomena where variability naturally increases with magnitude.

The consequences of undetected heteroskedasticity primarily affect the accuracy of the standard errors. While Ordinary Least Squares (OLS) estimates of the regression coefficients remain unbiased and consistent, their calculated standard errors become inconsistent, leading to inaccurate confidence intervals and hypothesis tests. This inaccuracy translates directly into misleading t-statistics and F-statistics, thus potentially leading to incorrect conclusions about the statistical significance of the predictors. Therefore, employing formal tests like the Breusch-Pagan test is essential for robust econometric and statistical analysis.

The Breusch-Pagan test operates by performing an auxiliary regression where the squared residuals of the original OLS model are regressed on the independent variables. If the independent variables significantly explain the variation in the squared residuals, it suggests that the error variance is not constant. In the context of statistical computing, especially using R, this diagnostic process is streamlined using the lmtest package, which provides immediate results for statistical decision-making regarding the crucial assumption of homoskedasticity.

Prerequisites: Setting up the R Environment

To execute the Breusch-Pagan test effectively in R, users must first ensure they have the necessary computational tools installed and loaded. The core requirement is the lmtest package, which provides the **bptest()** function. Installation is initiated using the command **install.packages("lmtest")** within the R console. This command retrieves the package and integrates it into the local environment, preparing it for use in subsequent analyses.

Once installation is complete, it is crucial to load the library into the current session using **library(lmtest)**. Without this step, R will not be able to locate and execute the diagnostic function, resulting in an error. This loading process is required for every new R session in which the test is needed. Analysts must prioritize maintaining a clean and functional environment to ensure the seamless execution of statistical diagnostics.

The **bptest()** function is straightforward, primarily requiring the fitted OLS regression model object as its main argument. Although the function permits additional arguments, such as specifying the **vcov** (variance-covariance) matrix, standard practice involves utilizing the default studentized version of the test, which provides superior robustness against potential non-normality of the residuals. Therefore, the preparation focuses simply on having a valid **lm** object ready for input.

Detailed Example: Implementing the Breusch-Pagan Test in R

We will now walk through a practical example demonstrating the application of the Breusch-Pagan test in R. This process utilizes a standard dataset to fit an initial regression model and then formally tests the residuals for signs of unequal variance. For this illustration, we will use the highly familiar, built-in R dataset, **mtcars**, which is frequently employed for statistical examples.

Our goal is to examine how two critical vehicle characteristics, displacement (**disp**) and gross horsepower (**hp**), influence fuel efficiency, measured by miles per gallon (**mpg**). Before we rely on the statistical significance reported in the initial OLS output, we must confirm that the homoskedasticity assumption holds. If this assumption is violated, the calculated **standard errors** are unreliable, requiring adjustment before interpretation.

The following steps provide the necessary code and output for the complete workflow, starting with

loading the data and fitting the model, and concluding with the execution and review of the Breusch-Pagan diagnostic test. This systematic approach ensures the statistical analysis adheres to accepted robust practices.

Step 1: Constructing the Regression Model

The initial phase involves setting up the data and fitting the linear model that requires diagnostic testing. We use the standard **lm()** function in R to define the relationship. The model fitting step is crucial because the **bptest()** function depends entirely on the residuals generated by this primary model to calculate its test statistic.

In this analysis, **mpg** is the dependent variable, representing the outcome of interest. The variables **disp** and **hp** serve as the independent predictors. Fitting the model yields the standard regression summary, providing coefficient estimates, **standard errors**, and overall goodness-of-fit metrics. If the assumption of constant variance were known to hold, these results would be directly interpretable.

The code block below outlines the necessary commands for data loading, model fitting, and reviewing the initial summary statistics. It is good practice to review the coefficients and R-squared before proceeding to the diagnostic phase, although these estimates should not be fully trusted until the assumption checks are complete:

#load the dataset
data(mtcars)

#fit a regression model
model <- lm(mpg~disp+hp, data=mtcars)

#view model summary
summary(model)

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 30.735904 1.331566 23.083 < 2e-16 ***
disp -0.030346 0.007405 -4.098 0.000306 ***
hp -0.024840 0.013385 -1.856 0.073679 .

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.127 on 29 degrees of freedom
Multiple R-squared: 0.7482, Adjusted R-squared: 0.7309
F-statistic: 43.09 on 2 and 29 DF, p-value: 2.062e-09

Step 2: Executing the Breusch-Pagan Test

After successfully fitting the regression model, the next critical step is to execute the Breusch-Pagan test. This is done using the **bptest()** function from the loaded lmtest package. Recall that the test performs an auxiliary regression of the squared residuals on the predictor variables to determine if their variance is systematically related to the covariates.

The formal hypothesis structure is defined as: the null hypothesis (**H?**) asserts homoskedasticity (error variance is constant). The alternative hypothesis (**H?**) asserts heteroskedasticity (error variance is not constant and varies with predictors). We aim to find evidence strong enough to reject **H?**.

Executing the command is simple, requiring only the **model** object as input. The resultant output, displayed below, provides the test statistic (BP) and the corresponding p-value, enabling the final statistical decision:

#load lmtest library
library(lmtest)

#perform Breusch-Pagan Test
bptest(model)

studentized Breusch-Pagan test

data: model
BP = 4.0861, df = 2, p-value = 0.1296

Interpreting the Test Results

The output provides a **BP statistic** of **4.0861** and a crucial p-value of **0.1296**. Statistical decisions are made by comparing this p-value against a chosen significance level, typically $alpha = 0.05$. If the p-value is less than $alpha$, we reject the null hypothesis; if it is greater than $alpha$, we fail to reject the null hypothesis.

In this case, **0.1296 > 0.05**. Since the calculated p-value is significantly greater than the standard significance level, we **fail to reject the null hypothesis**. This outcome indicates that there is insufficient statistical evidence to conclude that heteroskedasticity is present in the regression model based on the input variables.

The conclusion is that the assumption of homoskedasticity holds for this model. Therefore, the **standard errors** reported in the initial model summary (Step 1) are considered reliable, and the

analyst can proceed with interpreting the statistical significance (t-values and p-values) of the coefficients with confidence, knowing that the variance structure is stable.

Consequences of Detecting Heteroskedasticity

Had we rejected the null hypothesis (i.e., if the p-value was less than 0.05), it would confirm the presence of heteroskedasticity, signaling a violation of a critical OLS assumption. While OLS coefficient estimates remain unbiased, the calculated **standard errors** are inaccurate and inconsistent, severely impacting the reliability of inferential statistics.

Often, heteroskedasticity leads OLS to underestimate the true variability of the coefficients, resulting in t-statistics that are artificially large and p-values that are artificially small. This inflated confidence can cause the analyst to mistakenly conclude that a predictor variable is statistically significant when it is not, a serious Type I error. Conversely, less common instances can lead to overestimation of variance, resulting in Type II errors.

Therefore, when the Breusch-Pagan test confirms unequal variance, remedial measures are mandatory. Relying on the original OLS output without correction invalidates subsequent hypothesis testing, confidence interval construction, and prediction intervals, rendering the statistical conclusions unreliable and misleading.

Remediation Strategies for Heteroskedasticity

When the Breusch-Pagan test confirms the presence of heteroskedasticity, analysts must employ strategies to obtain robust inferences. These corrective measures fall into two broad categories: altering the data structure (transformation) or modifying the estimation procedure (robust standard errors or weighting).

The choice between these methods depends on the data context and the analyst's priorities. Data transformation (e.g., taking the log) successfully stabilizes the error variance, but it often complicates the interpretation of the coefficients because they no longer represent the effect on the original scale of the response variable. Conversely, methods that adjust the estimation process maintain the original model structure, preserving interpretability while ensuring the **standard errors** are robustly calculated.

The most common and practical solutions available to address this inferential problem include transforming the response variable or utilizing Weighted Least Squares (WLS) regression or Heteroskedasticity-Consistent Standard Errors (HCSEs). The following sections detail the approaches for correction.

Method 1: Transformation of the Response Variable

Data transformation is often the first, easiest remedial step. The objective is to stabilize the variance by altering the scale of the response variable so that the spread of the residuals becomes more uniform across the range of fitted values. This strategy is particularly effective when the magnitude of the error variance appears proportional to the magnitude of the response variable.

The most widely utilized approach involves the **natural logarithm transformation**. If the response variable is Y, it is replaced by $ln(Y)$. This transformation tends to compress larger values more aggressively than smaller values, effectively reducing positive skewness and stabilizing variance. The log transformation is ideal when the theoretical relationship is multiplicative or when data naturally displays exponential growth patterns.

Other transformations include the **square root transformation** ($sqrt{Y}$), useful for count data or when variance is proportional to the mean, and the **reciprocal transformation** ($1/Y$). While transformations are mathematically sound and often restore homoskedasticity, the primary drawback is the altered interpretation of the resulting coefficients, which now describe the change in the transformed variable rather than the original units of measurement.

Method 2: Utilizing Weighted Least Squares (WLS) Regression

A more sophisticated econometric approach is the use of Weighted Least Squares (**WLS**) regression. Instead of changing the variables, WLS changes the estimation procedure. It assigns weights to each observation inversely proportional to the estimated variance of its error ($sigma^2_i$). This means observations with high variance--those contributing most to the heteroskedasticity problem--receive lower weights, reducing their influence on the parameter estimates.

If the exact functional form of the heteroskedasticity is correctly specified, WLS provides coefficient estimates that are not only unbiased but also the Most Efficient Linear Unbiased Estimators (BLUE), thereby yielding the most precise estimates possible. The challenge lies in accurately estimating the weight structure, which often requires a preliminary step of regressing squared residuals on potential variance predictors.

Alternatively, when the specific form of heteroskedasticity is unknown, analysts can use **Heteroskedasticity-Consistent Standard Errors** (HCSEs), commonly known as robust standard errors. These standard errors, implemented in R using packages like **sandwich** alongside **coeftest** from the lmtest package, adjust the variance calculation without altering the OLS coefficient estimates. This preserves the original model interpretation while ensuring that the calculated **standard errors**--and thus the t-statistics and p-values--are asymptotically valid, providing a powerful and non-invasive solution to the problem of unequal variance.