Understanding Heteroscedasticity and the Breusch-Pagan Test with Python

Authored by

Mohammed looti

November 8, 2025

Understanding Heteroscedasticity in Regression Modeling

In the field of regression analysis, particularly when applying the widely used Ordinary Least Squares (OLS) method, understanding the behavior of model errors--or residuals--is paramount. One critical assumption underpinning the reliability of OLS estimates is the concept of homoscedasticity. This term implies that the variance of the error terms is constant across all levels of the independent variables. When this fundamental assumption is violated, we encounter a condition known as Heteroscedasticity.

The term Heteroscedasticity literally translates to "different scatter." It refers to the unequal scatter of residuals, meaning there is a systematic change in the variability or spread of the residuals as the predictor variables change. Specifically, it refers to the case where there is a systematic change in the spread of the residuals over the range of measured values. For instance, in economic models, the variance of errors might be very small for low-value transactions but highly variable for high-value transactions. This non-constant variance suggests that the model's predictive accuracy is reliable for some data ranges but highly unreliable for others, fundamentally compromising the integrity of standard statistical inferences.

Heteroscedasticity is a serious problem because OLS regression assumes that the residuals come from a population that exhibits homoscedasticity. While the presence of heteroscedasticity does not introduce bias into the OLS coefficient estimates--they remain unbiased and consistent--it causes the standard errors of the coefficients to be estimated incorrectly. OLS typically underestimates the true variance of the coefficients, leading to standard errors that are too small. Consequently, calculated test statistics (like t-statistics) become inflated, making us more likely to incorrectly reject a true null hypothesis. When heteroscedasticity is present in a regression analysis, the results of the analysis become hard to trust because the significance tests are flawed.

Diagnosing Heteroscedasticity: The Breusch-Pagan Test

Given the severe implications for statistical inference, formally identifying non-constant variance is a necessary step in robust regression analysis. While graphical methods provide initial clues, a formal statistical test is required for definitive confirmation. One highly effective method to determine if Heteroscedasticity is present in a model is to use the Breusch-Pagan Test.

The Breusch-Pagan Test is a diagnostic tool designed to check if the variance of the residuals from a regression model is systematically related to the independent variables in the model. The underlying mechanism involves performing an auxiliary regression where the squared residuals of the original OLS model are regressed against the predictors. If the predictors significantly explain the variation in the squared residuals, then non-constant variance is inferred.

This tutorial explains how to practically implement and perform a Breusch-Pagan Test in Python

using the **statsmodels** library. By rigorously applying this test, data scientists can gain confidence in the reliability of their OLS standard errors and subsequent hypothesis testing.

Example Setup: Data Preparation in Python

For this practical example, we will use a dataset describing the attributes of 10 basketball players. Our objective is to model how performance statistics (points, assists, and rebounds) influence a player's overall rating. We begin by setting up the data structure using **Pandas** and **NumPy**.

The dataset includes four key variables, and we intend to fit a multiple linear regression model using 'rating' as the response variable and 'points', 'assists', and 'rebounds' as the explanatory variables. The code below loads and displays this sample data:

import numpy as np
import pandas as pd

#create dataset
df = pd.DataFrame({'rating': ,
'points': ,
'assists': ,
'rebounds': })

#view dataset
df

rating points assists rebounds
0 90 25 5 11
1 85 20 7 8
2 82 14 7 10
3 88 16 8 6
4 94 27 5 6
5 90 20 7 9
6 76 12 6 6
7 75 15 9 10
8 87 14 9 10
9 86 19 5 7

With the dataset defined, we are now prepared to fit the multiple linear regression model and then perform a Breusch-Pagan Test to determine if Heteroscedasticity is present in the regression residuals. This two-step approach ensures we first establish the model and then diagnose its compliance with core assumptions.

Step-by-Step Implementation in Python

Step 1: Fit a Multiple Linear Regression Model

The first prerequisite for the Breusch-Pagan test is having a fitted OLS model from which residuals can be extracted. We use the **statsmodels.formula.api** to specify and fit the model relating the rating to the three predictor variables.

```
import statsmodels.formula.api as smf

#fit regression model
fit = smf.ols('rating ~ points+assists+rebounds', data=df).fit()

#view model summary
print(fit.summary())
```

This process generates the model object `fit`, which contains all necessary components for the diagnostic test, including the calculated residuals (`fit.resid`) and the design matrix of the independent variables (`fit.model.exog`).

Step 2: Perform the Breusch-Pagan Test

Next, we utilize the `sms.het_breuschpagan` function from the **statsmodels** library. This function calculates the necessary statistics for the test, allowing us to assess the presence of non-constant variance.

```
from statsmodels.compat import lzip
import statsmodels.stats.api as sms

#perform Bresuch-Pagan test
names =
test = sms.het_breuschpagan(fit.resid, fit.model.exog)

lzip(names, test)
```

The output presents four key metrics: the Lagrange multiplier statistic, its associated p-value, the F-statistic, and its associated p-value. These values are essential for hypothesis testing.

Interpreting the Test Results and Drawing Conclusions

A Breusch-Pagan Test uses a formal hypothesis structure to evaluate the constant variance

assumption:

The Null Hypothesis (H0): Homoscedasticity is present (the variance of the residuals is constant).

The Alternative Hypothesis (Ha): Heteroscedasticity is present (the variance of the residuals is related to the independent variables).

The decision rule requires comparing the calculated p-value against a chosen significance level ($alpha$), typically 0.05. If the p-value is less than 0.05, we reject H_0, concluding that heteroscedasticity exists. If the p-value is greater than 0.05, we fail to reject H_0.

In this example, the Lagrange multiplier statistic for the test is approximately **6.004** and the corresponding p-value is **0.1114**. Because this p-value (0.1114) is not less than the standard significance level of 0.05, we **fail to reject the null hypothesis**. This means we do not have sufficient statistical evidence to say that Heteroscedasticity is present in this specific regression model. Thus, for this dataset, we can proceed with confidence that the standard errors derived from the OLS estimation are reliable.

Strategies for Addressing Heteroscedasticity

In cases where the Breusch-Pagan Test leads to the rejection of the null hypothesis--meaning heteroscedasticity is detected--corrective measures must be taken. Addressing this issue is crucial for ensuring valid statistical inference, particularly for hypothesis testing and confidence interval construction. There are three primary strategies commonly employed to remedy the situation:

Transform the Dependent Variable. One of the most straightforward ways to mitigate heteroscedasticity is by applying a variance-stabilizing transformation to the dependent variable. If the spread of the residuals increases with the predicted value, taking the logarithm (log) of the dependent variable is a common and often effective transformation. Log transformations tend to compress the scale, making the variance of the residuals more homogeneous. Other power transformations, such as square root, may also be appropriate depending on the specific relationship observed in the residual plots.

Redefine the Dependent Variable. Another effective method is to redefine the dependent variable to naturally reduce scale variability. This usually involves switching from modeling a raw value to modeling a rate or a ratio. For example, instead of modeling the total count of an event, one might model the rate of that event per unit time or population size. Using a rate for the dependent variable helps standardize the error term relative to the size of the observation, often solving the scale-related heteroscedasticity.

Use Weighted Regression. The most sophisticated statistical solution is to use Weighted Least Squares (WLS) or, more broadly, Generalized Least Squares (GLS). This approach assigns a

weight to each data point based on the inverse of the variance of its fitted value. Essentially, WLS gives less weight to observations associated with larger error variance (which are less reliably estimated) and more weight to observations with smaller error variance. When the proper weights are successfully estimated and applied, WLS provides the most efficient and unbiased linear estimators, resolving the issue of incorrect standard errors caused by heteroscedasticity.

These methods allow researchers to continue their regression analysis while respecting the necessary assumptions of the OLS framework, leading to reliable conclusions and robust models.