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

Authored by

stats writer

March 9, 2026

Introduction to Econometric Validation and the Breusch-Pagan Test

In the field of **econometrics** and quantitative analysis, ensuring the reliability of a **regression analysis** model is paramount. One of the most critical diagnostic steps after fitting an **Ordinary Least Squares** model is checking for the presence of **heteroskedasticity**. This phenomenon occurs when the variance of the **residuals**, or the error terms, is not constant across all levels of the independent variables. To formally identify this issue, researchers frequently employ the **Breusch-Pagan test**, a robust statistical tool designed to detect linear forms of heteroskedasticity. This tutorial provides a comprehensive guide on how to implement this test using **Stata**, a premier software package for data science and statistical modeling.

The **Breusch-Pagan test** serves as a gateway to verifying the **Gauss-Markov assumptions**, which are essential for ensuring that the estimators produced by a regression model are the **Best Linear Unbiased Estimators** (BLUE). When these assumptions are violated, particularly the assumption of homoskedasticity, the **standard error** of the coefficients becomes biased. This bias can lead to misleading **t-statistics** and incorrect **p-values**, potentially causing a researcher to conclude that a variable is statistically significant when it is not, or vice versa. Therefore, mastering the execution and interpretation of the Breusch-Pagan test in **Stata** is a fundamental skill for any serious data analyst.

By utilizing the internal commands within **Stata**, users can efficiently transition from model estimation to diagnostic testing. The software streamlines the process, allowing for a quick evaluation of the relationship between **explanatory variables** and the **response variable** while simultaneously checking for underlying data irregularities. In the following sections, we will explore the theoretical underpinnings of the test, the step-by-step procedural implementation in the **Stata** environment, and the specific remediation strategies one should consider if heteroskedasticity is indeed detected in the dataset.

Understanding the context of your data is just as important as the mathematical execution of the test. **Regression analysis** is more than just drawing a line through a cloud of points; it is about modeling the conditional mean of a distribution. If the "spread" of that distribution changes as you move along the x-axis, your model's predictive power and inferential validity are compromised. The **Breusch-Pagan test** specifically targets this "spread" by examining whether the squared **residuals** can be explained by the independent variables in the original model, providing a formal **chi-squared** statistic to support the findings.

The Theoretical Framework of Heteroskedasticity

Before diving into the software commands, it is essential to define the concept of **heteroskedasticity** in detail. In a standard linear regression model, we assume

homoskedasticity, which means "equal variance." This implies that the **variance** of the error term remains constant regardless of the value of the **independent variable**. For instance, if you are modeling the relationship between income and food expenditure, homoskedasticity would imply that the variability in food spending is the same for both low-income and high-income households. However, in reality, high-income households often exhibit much greater variation in their spending habits, leading to a systematic change in the **variance** of the **residuals**.

When **heteroskedasticity** is present, the **Ordinary Least Squares** (OLS) method still provides unbiased estimates of the **regression coefficients**, but it is no longer the most efficient estimator. Efficiency in statistics refers to the estimator having the smallest possible **variance**. If the errors are heteroskedastic, the **standard error** estimates are incorrect, which invalidates **hypothesis testing**. This is why the **Breusch-Pagan test** is so vital; it acts as a diagnostic check to ensure that the **confidence intervals** and significance tests you report are actually trustworthy and reflect the true nature of the population data.

The **Breusch-Pagan test** specifically tests the **null hypothesis** that the error variances are all equal. The alternative hypothesis, conversely, suggests that the error variances are a function of one or more **explanatory variables**. In **Stata**, the test is typically performed after the regression model has been estimated, as it relies on the **residuals** generated from that specific model. By examining the patterns within these **residuals**, the test determines if there is a statistically significant trend that suggests the presence of non-constant variance.

It is worth noting that **heteroskedasticity** often arises in cross-sectional data where the units of observation (such as individuals, firms, or countries) vary significantly in size or scale. It can also appear in time-series data where volatility changes over time. Regardless of the source, the identification of this issue is the first step toward building a more robust **econometric** model. The **Breusch-Pagan test** is one of the most widely used methods because it is relatively simple to compute and interpret, though it is most effective when the heteroskedasticity is suspected to be a linear function of the **independent variables**.

An Overview of the Breusch-Pagan Statistical Test

The mechanics of the **Breusch-Pagan test** involve a multi-step statistical process. First, the **Ordinary Least Squares** regression is performed, and the **residuals** are calculated. These **residuals** are then squared and regressed against the same **independent variables** (or a different set of variables suspected of causing the variance change). The resulting **test statistic** follows a **chi-squared distribution**. If the **p-value** associated with this statistic is lower than a predetermined **significance level**--typically 0.05--the **null hypothesis** of homoskedasticity is rejected.

One of the advantages of using **Stata** for this procedure is its ability to handle different versions of

the test. While the original **Breusch-Pagan test** assumed that the error terms followed a **normal distribution**, **Stata** often utilizes the Koenker-Bassett version of the test, which is more robust to violations of normality. This makes the results more reliable when working with real-world data that might be skewed or have heavy tails. The **chi-squared** value produced by the command indicates the strength of the evidence against the **null hypothesis**.

When interpreting the **p-value**, researchers must be disciplined. A **p-value** of 0.04 might lead to a rejection of the **null hypothesis** at the 5% level, suggesting that **heteroskedasticity** is present. However, if the **p-value** is 0.15, the researcher fails to reject the **null hypothesis**, meaning there is insufficient evidence to conclude that the **variance** is non-constant. This doesn't necessarily prove the data is homoskedastic, but it suggests that the **Ordinary Least Squares** assumptions are likely "good enough" for the current model. Understanding this nuance is key to high-level **statistical inference**.

The **Breusch-Pagan test** is often compared to the **White test**, another popular diagnostic for **heteroskedasticity**. While the White test is more general and can detect non-linear forms of variance patterns, the Breusch-Pagan test is generally more powerful when the heteroskedasticity is indeed linear. In practice, many **econometric** practitioners run both tests to ensure a thorough diagnostic profile of their regression models. Within the **Stata** environment, both are easily accessible, allowing for a comprehensive validation workflow.

Preparing the Stata Environment and Loading Datasets

To demonstrate how to perform the **Breusch-Pagan test**, we will use a classic built-in dataset provided by **Stata**. This ensures that the results are reproducible for any user following this guide. The "auto" dataset contains various attributes of different car models from 1978, including price, mileage (mpg), weight, and repair records. This dataset is ideal for **regression analysis** because it contains several continuous variables that are likely to exhibit **heteroskedasticity**, particularly when modeling vehicle prices.

The first step in any **Stata** project is to load the data into the active memory. This is accomplished using the **sysuse** command. Following this, it is good practice to inspect the data visually to ensure there are no obvious errors or missing values that might interfere with the **regression**. The **br** (browse) command opens the Data Editor, providing a spreadsheet-like view of the variables. Observing the scale and range of variables like **price** and **weight** can give you an early indication of whether **data transformation** might be necessary later on.

Step 1: Load and view the data.

Execute the following command in the **Stata** command window to load the dataset:

sysuse auto

Once the data is loaded, you can view the raw structure by typing:

br

	make	price	mpg	rep78	headroom	trunk	weight	length	turn	displacement	gear_ratio	foreign
1	AMC Concord	4,099	22	3	2.5	11	2,930	186	40	121	3.58	Domestic
2	AMC Pacer	4,749	17	3	3.0	11	3,350	173	40	258	2.53	Domestic
3	AMC Spirit	3,799	22	.	3.0	12	2,640	168	35	121	3.08	Domestic
4	Buick Century	4,816	20	3	4.5	16	3,250	196	40	196	2.93	Domestic
5	Buick Electra	7,827	15	4	4.0	20	4,080	222	43	350	2.41	Domestic
6	Buick LeSabre	5,788	18	3	4.0	21	3,670	218	43	231	2.73	Domestic
7	Buick Opel	4,453	26	.	3.0	10	2,230	170	34	304	2.87	Domestic
8	Buick Regal	5,189	20	3	2.0	16	3,280	200	42	196	2.93	Domestic
9	Buick Riviera	10,372	16	3	3.5	17	3,880	207	43	231	2.93	Domestic
10	Buick Skylark	4,082	19	3	3.5	13	3,400	200	42	231	3.08	Domestic
11	Cad. Deville	11,385	14	3	4.0	20	4,330	221	44	425	2.28	Domestic
12	Cad. Eldorado	14,500	14	2	3.5	16	3,900	204	43	350	2.19	Domestic
13	Cad. Seville	15,906	21	3	3.0	13	4,290	204	45	350	2.24	Domestic
14	Chev. Chevette	3,299	29	3	2.5	9	2,110	163	34	231	2.93	Domestic
15	Chev. Impala	5,705	16	4	4.0	20	3,690	212	43	250	2.56	Domestic
16	Chev. Malibu	4,504	22	3	3.5	17	3,180	193	31	200	2.73	Domestic
17	Chev. Monte Carlo	5,104	22	2	2.0	16	3,220	200	41	200	2.73	Domestic
18	Chev. Monza	3,667	24	2	2.0	7	2,750	179	40	151	2.73	Domestic
19	Chev. Nova	3,955	19	3	3.5	13	3,430	197	43	250	2.56	Domestic

By looking at the dataset, we can identify our **response variable** (price) and our **explanatory variables** (mpg and weight). In car markets, it is common to see that as cars become more expensive, the variation in their price relative to their weight or fuel efficiency increases. This is a classic indicator that the **Breusch-Pagan test** will be necessary to validate any subsequent **Ordinary Least Squares** results. Preparing your workspace this way is the foundation of a clean and professional **econometric** workflow.

Constructing a Multiple Linear Regression Model

With the data loaded and inspected, the next logical step is to perform a **multiple linear regression**. This allows us to quantify the relationship between multiple **independent variables** and a single **dependent variable**. In our example, we want to see how much of a car's price can be explained by its fuel efficiency (mpg) and its weight. The **regress** command in **Stata** is the primary tool for this task, providing a comprehensive output that includes **coefficients**, **standard errors**, and **R-squared** values.

Step 2: Perform multiple linear regression.

Type the following command to initiate the **regression analysis**:

regress price mpg weight

```
. regress price mpg weight
```

Source	SS	df	MS	Number of obs	=	74
				F(2, 71)	=	14.74
Model	186321280	2	93160639.9	Prob > F	=	0.0000
Residual	448744116	71	6320339.67	R-squared	=	0.2934
				Adj R-squared	=	0.2735
Total	635065396	73	8699525.97	Root MSE	=	2514

price	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
mpg	-49.51222	86.15604	-0.57	0.567	-221.3025	122.278
weight	1.746559	.6413538	2.72	0.008	.467736	3.025382
_cons	1946.069	3597.05	0.54	0.590	-5226.245	9118.382

The output table generated by **Stata** provides several key pieces of information. The **coefficients** represent the estimated change in the **dependent variable** for a one-unit change in the **independent variable**, holding all other variables constant. For example, the **coefficient** for weight will tell us how much the price increases for every additional pound of vehicle weight. However, we cannot fully trust the **standard errors** or the **p-values** in this table until we have confirmed that the assumption of homoskedasticity is satisfied.

If **heteroskedasticity** is present, the **Ordinary Least Squares** estimators for the **variance** of the **coefficients** will be biased. This means the **t-test** for each variable might indicate significance when it shouldn't, leading to a **Type I error**. By proceeding to the **Breusch-Pagan test** immediately after the regression, we are performing due diligence to ensure that our model's findings are statistically sound and defensible in a professional or academic setting.

Executing the Breusch-Pagan Test via the hettest Command

After fitting the regression model, **Stata** stores the **residuals** in its temporary memory, allowing for immediate diagnostic testing. To perform the **Breusch-Pagan test**, we use the **hettest** command. This command is an abbreviation for "heteroscedasticity test." By default, when run without additional arguments, it performs the version of the Breusch-Pagan test that evaluates whether the **variance** of the **residuals** is a function of the fitted values of the **dependent variable**.

Step 3: Run the Breusch-Pagan Test.

Simply type the following command into **Stata**:

hettest

```
. hettest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of price

         chi2(1)      =    14.78
         Prob > chi2  =   0.0001
```

This command triggers **Stata** to calculate the squared **residuals** and perform the auxiliary **regression analysis** necessary to generate the **test statistic**. The resulting output is concise but contains all the necessary information to make a determination about your model's **homoskedasticity**. It specifically reports a **chi-squared** value and the corresponding **p-value**, which are the primary metrics for **hypothesis testing** in this context.

One of the strengths of the **hettest** command is its flexibility. While the default usage is standard, you can also specify particular variables if you suspect that **heteroskedasticity** is only related to one of the **explanatory variables** rather than the overall model. For instance, if you believed that only vehicle weight caused variance issues, you could refine the command to focus specifically on that variable. This level of control is what makes **Stata** an industry standard for **econometric** diagnostics.

Detailed Interpretation of the Test Output

Understanding how to read the **Breusch-Pagan test** output in **Stata** is essential for correct **statistical inference**. The output contains several labels that correspond to the mathematical components of the test. Below is a breakdown of how to interpret each element of the results window:

Ho: This represents the **null hypothesis**. In the case of the **hettest** command, the **null hypothesis** is always that there is constant variance (homoskedasticity) among the **residuals**.
Variables: This identifies the **response variable** or the fitted values used in the auxiliary **regression analysis**. In our example, it refers to the predicted values of **price**.
chi2(1): This is the **chi-squared** test statistic. The number in parentheses (1) represents the **degrees of freedom**. In our example, the statistic is 14.78, which is a relatively high value.

Prob > chi2: This is the **p-value**. It tells us the probability of observing a **test statistic** as extreme as 14.78 if the **null hypothesis** were true.

In our specific output, the **p-value** is 0.0001. Because this value is significantly lower than the standard **significance level** of 0.05, we must reject the **null hypothesis**. This rejection provides strong evidence that **heteroskedasticity** is present in the data. Consequently, we cannot rely on the **standard error** estimates from our original **Ordinary Least Squares** regression, as they are likely inaccurate. This finding necessitates further action to "fix" the model before the results can be used for decision-making or publication.

Failing to reject the **null hypothesis** (e.g., if the **p-value** were 0.25) would have meant that we could proceed with our original **regression analysis** without major concerns about **heteroskedasticity**. However, since we did reject it, we are now in the diagnostic and remediation phase of the **econometric** process. Recognizing this distinction is what separates a basic user from an expert in **Stata**.

Strategic Remediation: Data Transformations

Once **heteroskedasticity** has been confirmed by the **Breusch-Pagan test**, the first and often most effective strategy is **data transformation**. Transforming the **response variable** can stabilize the **variance** across the range of the **independent variables**. The most common technique is the **log transformation**. By taking the natural logarithm of the **dependent variable** (e.g., using log(price) instead of price), you often compress the scale of the data, which naturally reduces the magnitude of the **residuals** at higher values.

The logic behind a **log transformation** is rooted in the fact that many economic variables exhibit exponential growth or multiplicative errors rather than additive ones. When you apply a log, these multiplicative relationships become additive, often satisfying the **homoskedasticity** requirement. Another alternative is the **square root transformation**, which is particularly useful when the data follows a Poisson-like distribution where the **variance** is proportional to the mean. **Stata** makes these transformations easy with the **generate** command.

After applying a **data transformation**, it is imperative to re-run the **regression analysis** and the **Breusch-Pagan test**. The goal is to see if the **p-value** of the **hettest** command rises above the 0.05 threshold. If the transformation was successful, the new model will have homoskedastic **residuals**, making the **t-statistics** and **standard error** estimates reliable once again. This iterative process of testing and transforming is a standard part of refining a high-quality **econometric** model.

However, users should be aware that transforming the **dependent variable** changes the interpretation of the **regression coefficients**. In a log-level model, the **coefficient** represents the

percentage change in the **response variable** for a one-unit change in the **independent variable**. While this is often a very useful interpretation in economics, it is a significant shift from the original "dollar amount" interpretation of the car price. Researchers must balance the need for statistical validity with the need for clear, actionable insights.

Advanced Statistical Adjustments: Weighted Regression and Robust Errors

If **data transformation** does not resolve the issue, or if the research context requires the variables to remain in their original units, more advanced methods are available. One such method is **weighted regression**, also known as **Weighted Least Squares** (WLS). In this approach, **Stata** assigns a weight to each observation based on the inverse of the **variance** of its fitted value. Observations with higher **variance** are given less weight, effectively "punishing" the data points that are less certain. This can effectively neutralize the impact of **heteroskedasticity** on the model's efficiency.

Another highly popular solution in modern **econometrics** is the use of **robust standard errors**, often called Huber-White **standard errors** or "sandwich estimators." Instead of trying to eliminate the **heteroskedasticity**, this method adjusts the **standard error** calculations to account for it. In **Stata**, this is done by simply adding the **vce(robust)** option to the end of your **regress** command. This does not change the **coefficients**, but it provides a more accurate **standard error** that is "robust" to the presence of non-constant variance.

Using **robust standard errors** is often the preferred choice when the exact functional form of the **heteroskedasticity** is unknown. It provides a safeguard that ensures your **hypothesis testing** is valid regardless of the variance structure. Many academic journals now require **robust standard errors** as a default for any published **regression analysis**. By understanding how to move from a **Breusch-Pagan test** to implementing **vce(robust)**, you ensure that your work meets the highest standards of **statistical** rigor.

In summary, performing a **Breusch-Pagan test** in **Stata** is a vital diagnostic step for any researcher using **linear regression**. By identifying **heteroskedasticity**, interpreting the **chi-squared** statistics, and applying remediation techniques like **data transformation** or **robust standard errors**, you can produce models that are both accurate and reliable. **Stata** provides all the tools necessary to navigate this process, from initial data loading to advanced **econometric** adjustment, ensuring the integrity of your statistical conclusions.