A Guide to Testing for Heteroskedasticity with the Breusch-Pagan Test in Stata

Authored by

Mohammed looti

November 8, 2025

RECOMMENDED CITATION

Mohammed looti (2025). *A Guide to Testing for Heteroskedasticity with the Breusch-Pagan Test in Stata*. PSYCHOLOGICAL STATISTICS. Retrieved from https://statistics.arabpsychology.com/?p=13632

The Critical Role of Variance Assumptions in Regression Modeling

Regression analysis stands as a foundational technique in quantitative research, allowing analysts to quantify and model the relationship between a dependent outcome variable and a set of explanatory variables. When employing conventional estimation methods, such as Ordinary Least Squares (OLS), the validity of our conclusions rests heavily upon several theoretical assumptions regarding the underlying data structure. Meeting these assumptions is essential to guarantee that the resulting coefficient estimates are the Best Linear Unbiased Estimators (BLUE). Violations of these core tenets can severely undermine statistical inference, leading to inaccurate conclusions about the variables' true relationships.

Chief among these requirements is the assumption of homoscedasticity--the requirement that the variance of the error terms (residuals) must remain constant across all observations and all levels of the independent variables. When this condition fails, we encounter the problem of heteroscedasticity. Although OLS still produces unbiased coefficient estimates under heteroscedasticity, the calculated Standard errors become biased and inconsistent. This inconsistency introduces significant risk, as it invalidates hypothesis tests and confidence intervals, making it impossible to reliably determine the statistical significance of the explanatory factors.

Heteroscedasticity often manifests when the predictability of the outcome varies systematically with the magnitude of the predictors. For instance, in financial modeling, the volatility of stock returns might increase significantly for higher-priced stocks. Recognizing and diagnosing this non-constant variance is paramount for constructing a robust and trustworthy regression model. To formally assess the presence of this issue, we turn to specific diagnostic tools, most notably the widely respected Breusch-Pagan Test.

Deconstructing the Breusch-Pagan Test Methodology

The Breusch-Pagan Test provides a formal, rigorous statistical framework specifically designed to detect the presence of heteroscedasticity within a linear regression model. Its fundamental goal is to determine if the magnitude of the residual variance is systematically related to the values of the independent variables used in the original OLS estimation. If such a relationship exists, it suggests that the variance is not constant, thereby violating the critical homoscedasticity assumption.

The test operates by executing an auxiliary regression. First, the original OLS model is run, and the residuals are computed. These residuals are then squared. In the auxiliary step, these squared residuals are regressed against the original explanatory variables (or sometimes against the predicted values from the original model). If the explanatory variables collectively possess significant predictive power over the squared residuals, it implies that they are driving the variability of the errors, which is the definition of heteroscedasticity.

The outcome of the test is summarized by a test statistic, typically distributed as a Chi-Square test statistic, along with its associated p-value. The hypotheses underpinning the test are crucial for interpretation: The Null hypothesis (H_0) asserts that the variance is constant (homoscedasticity), meaning the independent variables have no influence on the error variance. Conversely, the alternative hypothesis (H_A) posits that heteroscedasticity is present. We will now demonstrate the precise procedure for implementing this diagnostic test using the powerful statistical software package, **Stata**.

Preparation and Model Specification in Stata

To demonstrate the practical application of the Breusch-Pagan Test, we will utilize the standard, pre-loaded dataset within Stata, known as **auto**. This dataset contains comprehensive information about various automobile characteristics, providing an excellent scenario for a multiple linear regression where variance issues are common.

Step 1: Loading and Reviewing the Dataset

The initial step in any statistical analysis is ensuring that the data is correctly loaded into the environment and its structure is understood. We load the sample data automatically available within **Stata** using the **sysuse** command:

sysuse auto

After loading, it is highly recommended to examine the dataset's contents and structure using the **br** (browse) command. This visual inspection helps confirm data integrity and reminds us of the variable types before specifying our econometric model:

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

For our example, we will focus on modeling the vehicle **price** as the dependent variable, utilizing **mpg** (miles per gallon) and **weight** as our key explanatory variables.

Step 2: Fitting the Regression Model and Generating Residuals

The Breusch-Pagan Test is inherently tied to the residuals of a specific model; therefore, it must be executed immediately following the estimation of the regression equation of interest. We construct a multiple linear regression where vehicle **price** is predicted by **mpg** and **weight**.

The command utilized for running OLS regression in **Stata** is **regress**. The syntax requires the dependent variable first, followed by the list of independent variables. Executing this procedure not only calculates the initial coefficients but critically generates and stores the residuals necessary for the subsequent diagnostic test:

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 resulting output provides the standard regression metrics, including the coefficients and their associated Standard errors. However, we must resist interpreting the statistical significance based on these standard errors until we have formally confirmed the assumption of homoscedasticity. If the Breusch-Pagan Test reveals heteroscedasticity, these initial standard error calculations will be misleading and unreliable for making accurate inferences.

Step 3: Executing the Diagnostic Test with hettest

Immediately following the successful estimation of the regression model using the **regress** command, the relevant residuals are automatically retained by **Stata**. We can then proceed directly to executing the Breusch-Pagan Test using the dedicated hettest command, which serves as **Stata**'s built-in procedure for checking heteroscedasticity.

Crucially, the standard Breusch-Pagan procedure requires no additional arguments when run immediately after **regress**, as **Stata** is programmed to automatically use the recently calculated residuals and the original explanatory variables for the auxiliary regression:

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
```

Interpreting the Breusch-Pagan Test Results

The concise output generated by the **hettest** command contains all the necessary metrics for a definitive decision regarding the constancy of the error variance. A meticulous review of these statistics is essential for drawing statistically valid conclusions about the model's fitness and determining the necessity of corrective measures.

Null Hypothesis (Ho): This line explicitly states the assumption under test: that the variance of the residuals is constant (homoscedasticity). Our analytical goal is ideally to fail to reject this Null hypothesis, which would affirm the validity of the OLS assumption.

Variables: This confirms that the test is utilizing the residuals derived from the regression of the dependent variable, **price**.

chi2(1): This metric presents the calculated value of the Chi-Square test statistic, which quantifies the evidence against the constant variance assumption. In our example, the statistic is calculated as **14.78**.

Prob > chi2: This is the corresponding p-value, which represents the probability of observing a test statistic as extreme as 14.78 if the Null hypothesis were true. The p-value here is extremely small: **0.0001**.

To finalize the diagnostic test, we compare the calculated **Prob > chi2** (0.0001) against our chosen significance level, typically $alpha = 0.05$. Since 0.0001 is substantially lower than 0.05, the result indicates strong statistical significance. We must therefore definitively reject the Null hypothesis (H_0). The rejection of H_0 provides compelling evidence that severe heteroscedasticity is present in the data, confirming that the initial Standard errors and subsequent t-statistics produced by the OLS regression are invalid and require immediate adjustment.

Remedial Strategies for Addressing Heteroscedasticity

Once the Breusch-Pagan Test has confirmed the presence of non-constant variance, corrective action is mandatory to ensure that the statistical inferences drawn from the model are sound. If the test had resulted in a failure to reject the null hypothesis, the homoscedasticity assumption would hold, and we could confidently interpret the original OLS output. However, since we rejected H_0 in our example, one of the following three major remedial measures must be applied to obtain reliable coefficient estimates and trustworthy Standard errors.

Transforming the Response Variable

One fundamental remedy involves applying a mathematical transformation to the dependent variable. Non-constant variance frequently arises because the inherent scale of the response variable leads to larger errors at higher values. Standard transformations, such as the natural logarithm (log) or the square root, work to compress the scale of the variable, often successfully stabilizing the variance of the model's residuals.

For instance, in our example, we could re-run the regression using **log(price)** instead of the raw **price** variable. Log transformations are particularly effective in socioeconomic models involving variables like prices or income, where variability tends to increase exponentially with magnitude. An added benefit is that the interpretation of the coefficients shifts from absolute units to percentage changes, which is often more meaningful in economic contexts.

Employing Weighted Least Squares (WLS)

Weighted regression, specifically known as Weighted Least Squares (WLS), is a powerful technique that directly mitigates heteroscedasticity by adjusting the influence of individual observations on the model fit. This methodology requires assigning a specific weight to every data point, where the weight is ideally inversely proportional to the known or estimated variance of its error term.

By giving observations associated with higher variance--the very source of the heteroscedasticity problem--smaller weights, their impact on the squared residuals is diminished. When the correct weighting scheme is successfully determined and applied, WLS provides efficient and reliable estimates, thereby restoring the validity of the OLS assumptions. However, determining the correct weights often requires strong theoretical grounding or an iterative process based on careful modeling of the residual structure.

Utilizing Robust Standard Errors (Huber-White)

The most common and generally simplest pragmatic solution, particularly favored in applied **econometrics** and data science, is the use of **robust standard errors** (often referred to as Huber-White standard errors). Instead of attempting to modify the variance structure through

transformation or complex weighting, this approach simply adjusts the calculation of the Standard errors themselves, making them inherently "robust" to any misspecification of the error variance function.

Robust standard errors provide an accurate measure of the true variability of the regression coefficients, even when the underlying errors are known to be heteroscedastic. This adjustment ensures that the reported t-statistics and associated p-values are statistically trustworthy for inference, allowing researchers to maintain the original model specification without complex data transformations. In **Stata**, this solution is implemented effortlessly by appending the **robust** option to the **regress** command (e.g., **regress price mpg weight, robust**). This method is widely accepted as the fastest and least disruptive path to reliable inference when heteroscedasticity is detected.