A Tutorial on White’s Test for Homoscedasticity in SAS Regression

Authored by

Mohammed looti

November 14, 2025

Understanding Homoscedasticity and the OLS Assumption

When executing regression analysis, particularly through the widely used method of Ordinary Least Squares (OLS), the reliability of the statistical inferences produced is fundamentally dependent upon meeting several core assumptions. The most critical of these assumptions for OLS is **homoscedasticity**. This condition dictates that the variance of the model's error terms—often referred to as the residuals—must remain constant and uniform across all observed levels of the independent variables. Maintaining this stability in variance is essential because it ensures that the OLS estimator retains its optimal statistical property: being the **Best Linear Unbiased Estimator (BLUE)**.

The failure of this assumption introduces a serious statistical issue known as heteroscedasticity. This phenomenon occurs when the systematic spread or scatter of the residuals changes as the predictor values increase or decrease. While the presence of heteroscedasticity typically does not bias the estimated coefficients themselves, it severely compromises the accuracy and consistency of the calculated standard errors. Since these errors are used to construct confidence intervals and calculate test statistics, biased and inconsistent standard errors make traditional hypothesis tests (like t-tests and F-tests) statistically unreliable, preventing analysts from confidently assessing the true significance of predictors in the model.

Consequently, the rigorous identification and subsequent management of non-constant variance is a mandatory step in professional statistical modeling. To formally diagnose this unequal variance, researchers rely on robust diagnostic tools. Among the most popular and versatile tests is White's test. Developed by the renowned econometrician Halbert White, this test provides a general framework for detecting various forms of heteroscedasticity without needing the restrictive assumption that the error terms must be normally distributed. This comprehensive guide details the precise implementation and interpretation of White's test within the powerful SAS statistical software environment, thereby ensuring the foundational validity of your regression analysis.

Setting Up the Sample Data and Regression Model in SAS

Before any diagnostic testing can be performed, two prerequisites must be met: a structured dataset must be prepared, and the specific regression model must be formally defined. To illustrate the process, we will utilize a practical scenario focused on predicting student performance. Our goal is to model a student's final exam score as a function of two predictor variables: the total number of hours they spent studying and the number of preparatory exams they took. This relationship is best analyzed using a multiple linear regression model.

The theoretical structure of this model is mathematically expressed as: **Exam Score =** β**0 +** β**1(hours) +** β**2(prep exams) +** ε. In this equation, β0 represents the intercept, β1 and β2 are the estimated coefficients that quantify the marginal effect of study hours and preparatory exams,

respectively, and ε is the unavoidable random error term. To proceed with the analysis in SAS, we must first construct a sample dataset containing the variables for a group of 20 hypothetical students.

The following SAS code snippet demonstrates the necessary steps to create and verify the data structure. The code utilizes the standard `DATA` step to name the new dataset (`exam_data`) and the `INPUT` statement to define the variables (`hours`, `prep_exams`, `score`). The actual data points are entered directly using the `DATALINES` statement. Following data entry, the `PROC PRINT` command is executed immediately to ensure that the data was loaded accurately and successfully into the SAS environment.

```
/*create dataset*/
data exam_data;
input hours prep_exams score;
datalines;
1 1 76
2 3 78
2 3 85
4 5 88
2 2 72
1 2 69
5 1 94
4 1 94
2 0 88
4 3 92
4 4 90
3 3 75
6 2 90
5 4 90
3 4 82
4 4 85
6 5 90
2 1 83
1 0 62
2 1 76
;
run;

/*view dataset*/
proc print data=exam_data;
```

Executing this preparatory code successfully creates the `exam_data` table, which contains all the necessary observations for our variables. This step is absolutely foundational, as all subsequent regression analysis and advanced diagnostic procedures, including the formal White's test, will rely entirely on the correct definition and population of this initial dataset.

Obs	hours	prep_exams	score
1	1	1	76
2	2	3	78
3	2	3	85
4	4	5	88
5	2	2	72
6	1	2	69
7	5	1	94
8	4	1	94
9	2	0	88
10	4	3	92
11	4	4	90
12	3	3	75
13	6	2	90
14	5	4	90
15	3	4	82
16	4	4	85
17	6	5	90
18	2	1	83
19	1	0	62
20	2	1	76

Implementing White's Specification Test Using PROC REG

With the data successfully loaded and verified in the SAS environment, the next critical step is to fit the specified multiple linear regression model and simultaneously request the specific diagnostic procedure for non-constant variance. In SAS, the robust **`PROC REG`** procedure is the standard mechanism for conducting OLS regression. To specifically request White's test, a powerful but straightforward option must be appended to the **`MODEL`** statement: the **`spec`** option.

The inclusion of the **`spec`** option instructs `PROC REG` to perform a general specification test designed primarily to check for various forms of model misspecification, which critically includes the detection of heteroscedasticity. The technical foundation of White's test relies on generating an

auxiliary regression. In this secondary model, the squared residuals (the estimated errors) obtained from the primary regression are regressed against the original predictor variables, their squared terms, and all possible cross-products between the predictors. The resulting test statistic, which follows a Chi-Square distribution, is derived from the R-squared value of this auxiliary regression, scaled by the number of observations.

To execute this analysis for our student exam score data, analysts must use the following SAS code block. The `MODEL` statement clearly defines the relationship being tested (`score` as the dependent variable, predicted by `hours` and `prep_exams`), and the mandatory inclusion of the `/ spec` keyword ensures that the necessary diagnostic output, specifically including the results of White's general specification test, is generated for interpretation.

```
/*fit regression model and perform White's test*/
proc reg data=exam_data;
model score = hours prep_exams / spec;
run;
quit;
```

Upon execution, `PROC REG` will produce extensive output detailing the primary regression model results, including the estimated coefficient estimates and their associated standard errors. Crucially, a dedicated section within this output, usually labeled "Tests of Model Specification," will present the calculated metrics for White's Test for Heteroscedasticity. Locating this specific table is the next essential step before proceeding to the interpretation phase.

The REG Procedure
Model: MODEL1
Dependent Variable: score

Number of Observations Read	20
Number of Observations Used	20

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	937.81311	468.90656	13.62	0.0003
Error	17	585.13689	34.41982		
Corrected Total	19	1522.95000			

Root MSE	5.86684	R-Square	0.6158
Dependent Mean	82.95000	Adj R-Sq	0.5706
Coeff Var	7.07274		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	69.66126	3.07879	22.63	<.0001
hours	1	4.65376	0.98317	4.73	0.0002
prep_exams	1	-0.55943	0.99979	-0.56	0.5831

The REG Procedure
Model: MODEL1
Dependent Variable: score

Test of First and Second Moment Specification		
DF	Chi-Square	Pr > ChiSq
5	3.54	0.6175

Interpreting the Results: Chi-Square Statistic and P-Value

The ultimate objective of running White's test is to arrive at a definitive statistical conclusion regarding the presence or absence of heteroscedasticity within the OLS model. This decision is derived from a careful analysis of the key diagnostic metrics provided in the dedicated output table

generated by `PROC REG`. Specifically, analysts must focus on the calculated Chi-Square test statistic and its corresponding probability value, or p-value.

Based on the sample output image provided above, the calculated Chi-Square test statistic is reported as **3.54**, with 5 degrees of freedom. More critically, the associated p-value is **0.6175**. To provide context for these values, we must explicitly state the formal hypotheses tested by this diagnostic procedure:

Null Hypothesis (H0): The model errors exhibit constant variance; therefore, there is no statistically significant heteroscedasticity present. (The assumption of homoscedasticity holds.)
Alternative Hypothesis (HA): The model errors exhibit non-constant variance, suggesting that heteroscedasticity is a significant problem.

The standard procedure for hypothesis testing requires comparing the calculated p-value against a pre-selected level of significance (α), which is almost universally set at 0.05. The decision rule is straightforward: If the p-value is less than the threshold ($\alpha < 0.05$), we are compelled to reject the Null Hypothesis, concluding that heteroscedasticity is statistically significant. Conversely, if the p-value is greater than or equal to 0.05, we fail to reject the Null Hypothesis.

In this specific student performance example, the obtained p-value of 0.6175 is substantially larger than the standard 0.05 significance level. Consequently, we **fail to reject the null hypothesis**. This outcome is highly favorable, as it provides strong statistical evidence that there is insufficient basis to conclude that non-constant variance is a problem in our student exam score regression model. We can proceed with the model interpretation confidently, knowing that the standard errors of the coefficient estimates are reliable for making valid statistical inferences about the effects of our predictors.

Advanced Strategies for Correcting Detected Heteroscedasticity

The interpretation of White's test dictates the necessary direction for the subsequent analytical phase. When the test yields a non-significant result (as demonstrated above), the crucial assumption of constant variance is upheld, and the calculated standard errors and parameter estimates from the original OLS regression model are deemed efficient and trustworthy. However, a significant result--the statistical rejection of the Null Hypothesis--demands immediate corrective action. A rejected null signals that the reported standard errors are unreliable, which could potentially lead to inaccurate conclusions regarding the statistical significance of the predictor variables.

When heteroscedasticity is confirmed, analysts have several established methodologies to either mitigate the underlying issue or adjust the statistical machinery to account for the presence of non-constant variance. The optimal choice of strategy is often determined by the severity and specific

pattern of the non-constant variance observed in the data. Below are three primary approaches used by econometricians and statisticians to address this violation and restore the integrity of the OLS inferences:

1. Data Transformation of the Response Variable.

One of the simplest and often most effective solutions is applying a mathematical transformation to the dependent variable. The overarching goal of this technique is to stabilize the variance across the entire range of predictor values. Common transformations include using the natural logarithm (e.g., `LOG(Y)`) or the square root (e.g., `SQRT(Y)`). The log transformation is particularly powerful for mitigating positive skewness and can frequently minimize or entirely eliminate observed heteroscedasticity, thereby ensuring the validity of the OLS assumptions.

2. Implementing Weighted Least Squares (WLS) Regression.

The weighted regression approach involves assigning specific weights to each observation that are inversely proportional to the variance of its corresponding error term. The statistical principle here is that observations associated with higher variance (those with larger squared residuals) should inherently exert less influence on the overall model fitting process. By correctly weighting the data points, WLS can produce more efficient coefficient estimates and statistically correct standard errors. However, successfully implementing WLS requires accurately estimating the complex, unknown variance structure of the errors, which represents a significant practical challenge.

3. Employing Heteroscedasticity-Consistent Standard Errors (HCSEs).

Also widely known as robust standard errors, this method is arguably the most preferred and common modern solution. Instead of attempting to fundamentally transform the data or change the core weighting scheme, the HCSE method adjusts the calculation of the standard errors directly to account for the confirmed presence of non-constant variance. HCSEs ensure that resulting hypothesis tests and confidence intervals remain statistically valid even when the homoscedasticity assumption is violated. Most contemporary statistical packages, including SAS, offer built-in options to easily compute robust standard errors, providing a technically sound and straightforward way to maintain reliable inference without having to fundamentally alter the core regression specification.

A sound statistical workflow mandates careful consideration and application of these corrective options when heteroscedasticity is detected, ensuring that the final model yields conclusions that are mathematically robust and substantively meaningful for the research question at hand.