Perform Weighted Least Squares Regression in R

Authored by

Mohammed looti

November 6, 2025

The Problem with Ordinary Least Squares (OLS) Assumptions

Ordinary Least Squares (OLS) regression stands as the cornerstone of many statistical analyses, providing efficient and unbiased coefficient estimates, provided its underlying assumptions are met. However, the reliability of OLS hinges fundamentally on a critical requirement: that the variance of the error term--the difference between the observed and predicted values, often referred to as **residuals**--remains constant across all levels of the independent variables. This ideal state is formally known as **homoscedasticity**.

When real-world data is analyzed, achieving perfect homoscedasticity is rare. Data often exhibit unequal variability, especially in fields like economics or biological sciences where measurements might become less precise at higher values. When this fundamental assumption is violated, the model suffers from **heteroscedasticity**. This condition signifies that the spread or variance of the errors is systematically unequal, often fanning out or tightening in a recognizable pattern as the predictor variable changes.

The presence of heteroscedasticity does not bias the coefficient estimates themselves; they remain unbiased, but they cease to be the most efficient estimates. Crucially, the standard errors of these estimates become unreliable, typically underestimated. This miscalculation leads to inflated t-statistics and potentially misleading p-values, rendering statistical inference--such as hypothesis testing and confidence interval construction--untrustworthy. Consequently, a researcher might mistakenly conclude that a predictor is significant when it is not, or vice versa.

To restore the validity of statistical inference and produce reliable results in the presence of this non-constant variance, advanced estimation techniques are essential. The most powerful and widely accepted method for addressing this issue is the **weighted least squares regression** (WLS). WLS is designed to mitigate the impact of heteroscedasticity by assigning different levels of importance, or weights, to each observation. Observations associated with smaller error variance are deemed more reliable and thus receive greater weight, effectively downplaying the influence of less precise data points.

Step 1: Setting Up the Analysis Environment and Data

This tutorial provides a practical, step-by-step guide on how to correctly implement and interpret weighted least squares regression within the **R** statistical programming environment. R is the industry standard for statistical computing and graphics, offering robust tools for both standard and specialized regression modeling.

To demonstrate the necessity and procedure of WLS, we will first construct a simple, illustrative dataset. This example simulates a common scenario in educational research, where we examine the relationship between the number of hours a student dedicates to studying and their resulting

exam score. Our sample data frame, which we name `df`, includes 16 observations detailing these two variables.

The subsequent code block initializes this data frame in R. This structure serves as the foundation upon which we will build our standard OLS model, diagnose the variance issues, and finally implement the WLS correction. It is important to define the data clearly before proceeding with any statistical modeling.

df <- data.frame(hours=c(1, 1, 2, 2, 2, 3, 4, 4, 4, 5, 5, 5, 6, 6, 7, 8),
score=c(48, 78, 72, 70, 66, 92, 93, 75, 75, 80, 95, 97, 90, 96, 99, 99))

Step 2: Establishing a Baseline with Ordinary Least Squares

Before applying any corrective measures, it is essential to first fit a standard OLS simple linear regression model. This step provides a crucial baseline against which we can compare the corrected WLS results, allowing us to quantify the improvement achieved by accounting for non-constant variance. In this model, `hours` acts as the predictor (independent) variable, and `score` is the dependent variable we aim to predict. We utilize R's powerful built-in function, `lm()`, which is the workhorse for linear modeling.

Upon execution, we immediately review the summary statistics of the OLS model. Key metrics to observe at this stage include the coefficient estimates, the Residual Standard Error (RSE), and the Multiple R-squared value. These initial values will be critically important in Step 5 when we perform a direct comparative analysis to highlight the statistical advantages of the WLS approach. Notice particularly the standard errors associated with the coefficients, as these are the values that heteroscedasticity fundamentally compromises.

#fit simple linear regression model
model <- lm(score ~ hours, data = df)

#view summary of model
summary(model)

Call:
lm(formula = score ~ hours, data = df)

Residuals:
Min 1Q Median 3Q Max
-17.967 -5.970 -0.719 7.531 15.032

Coefficients:

Estimate Std. Error t value Pr(>|t|)
(Intercept) 60.467 5.128 11.791 1.17e-08 ***
hours 5.500 1.127 4.879 0.000244 ***

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

Residual standard error: 9.224 on 14 degrees of freedom
Multiple R-squared: 0.6296, Adjusted R-squared: 0.6032
F-statistic: 23.8 on 1 and 14 DF, p-value: 0.0002438

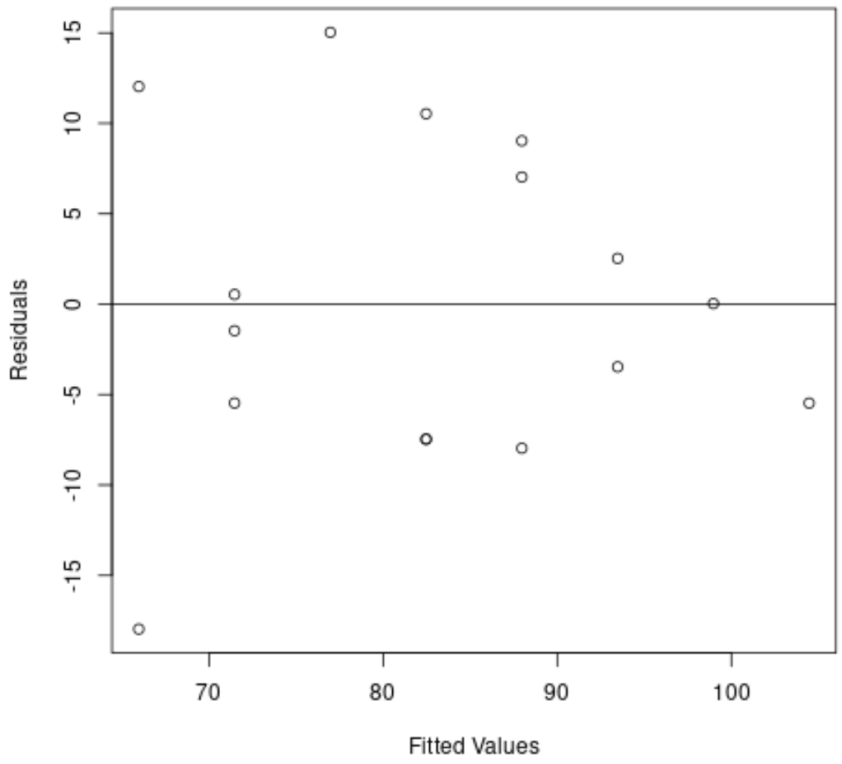
Step 3: Formal Diagnosis of Heteroscedasticity

The validity of using weighted least squares relies entirely on confirming the presence of non-constant error variance in the OLS model. We employ a two-pronged diagnostic approach: visual inspection followed by a rigorous statistical test. We begin by plotting the **residuals** of the initial OLS model against its fitted (predicted) values. This plot is perhaps the quickest and most intuitive way to spot variance issues.

In a model adhering to the homoscedasticity assumption, the residuals should scatter randomly around the horizontal zero line, forming a roughly rectangular band of constant width. Any systematic pattern, such as a broadening (fanning out) or narrowing (funneling in) of the vertical spread as the fitted values increase, serves as a strong visual indicator of unequal variance. The R code below generates this diagnostic plot, including a reference line at zero for easy interpretation.

#create residual vs. fitted plot
plot(fitted(model), resid(model), xlab='Fitted Values', ylab='Residuals')

#add a horizontal line at 0
abline(0,0)



As the resulting plot clearly illustrates, the scatter of the **residuals** does not form a horizontal band. Instead, the points distinctly fan out, forming a recognizable "cone" or triangular shape. This pattern is the classic visual signature confirming that the variance of the error term is systematically increasing as the predicted exam score increases. This strong visual evidence confirms the presence of **heteroscedasticity**, necessitating a statistically sound correction.

To move beyond visual confirmation and formally confirm this statistical violation, we perform the **Breusch-Pagan test**. This test is a highly reliable method that assesses whether the error variance is related to the independent variables. It requires loading the `lmtest` package in R. The test operates under the following hypothesis framework:

#load lmtest package
library(lmtest)

#perform Breusch-Pagan test
bptest(model)

studentized Breusch-Pagan test

data: model
BP = 3.9597, df = 1, p-value = 0.0466

Null Hypothesis (H0): The model exhibits **Homoscedasticity** (constant error variance).
Alternative Hypothesis (HA): The model exhibits **Heteroscedasticity** (non-constant error variance).

Given a p-value of **0.0466**, which is less than the conventional significance level of 0.05, we possess sufficient statistical evidence to reject the null hypothesis. Both the visual inspection and the formal statistical test converge on the same conclusion: the error variance is non-constant, and standard OLS inference would be unreliable. This statistical finding mandates the transition to a more robust estimation technique like WLS.

Step 4: Implementing Weighted Least Squares Regression (WLS) in R

Since our diagnostic steps confirmed that the OLS assumptions are violated, we now implement **weighted least squares regression**. The fundamental goal of WLS is to transform the heteroscedastic problem into a homoscedastic one by assigning weights (or inverse variance estimates) to each observation. Observations that are less variable (more precise) are given larger weights, while observations with greater variability (less precise) are effectively down-weighted, thereby minimizing their influence on the final coefficient estimates.

Determining the appropriate weights is the most crucial step in WLS. While the true functional form of the heteroscedasticity is often unknown, a pragmatic and frequently effective approach, particularly when the variance appears proportional to the fitted values (as suggested by our cone-shaped plot), is to set the weights inversely proportional to the variance of the errors. We can estimate this variance using the squared residuals obtained from the initial OLS model. Specifically, the weights (`wt`) are calculated as the inverse of the squared fitted values derived from a secondary regression of the absolute OLS residuals on the OLS fitted values.

The first line of the R code below performs this necessary weight calculation. Once the weights are successfully calculated, we re-run the linear model using the same formula as the OLS model but introduce the `weights` argument within the `lm()` function. This subtle addition tells R to perform the minimization process using the defined weights, thereby correcting for the non-constant error variance structure.

```
#define weights to use (based on the inverse squared fitted residuals from OLS)
wt <- 1 / lm(abs(model$residuals) ~ model$fitted.values)$fitted.values^2

#perform weighted least squares regression
wls_model <- lm(score ~ hours, data = df, weights=wt)

#view summary of model
summary(wls_model)
```

```
Call:
lm(formula = score ~ hours, data = df, weights = wt)

Weighted Residuals:
Min 1Q Median 3Q Max
-2.0167 -0.9263 -0.2589 0.9873 1.6977

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 63.9689 5.1587 12.400 6.13e-09 ***
hours 4.7091 0.8709 5.407 9.24e-05 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.199 on 14 degrees of freedom
Multiple R-squared: 0.6762, Adjusted R-squared: 0.6531
F-statistic: 29.24 on 1 and 14 DF, p-value: 9.236e-05
```

The resulting output for the WLS model demonstrates several key changes compared to the initial OLS results. Firstly, the coefficient estimates have shifted; the intercept increased from 60.467 to 63.9689, and the slope for *hours* decreased from 5.500 to 4.7091. This shift reflects the fact that the WLS estimates are now statistically more efficient, as they have been optimized by prioritizing the lower-variance observations. Secondly, and perhaps more importantly, the overall model fit metrics show significant improvement, confirming the efficacy of the weighting scheme in stabilizing the error variance and providing a more accurate representation of the true underlying relationship.

Step 5: Assessing the Performance Gains of WLS

The final and most instructive step involves a direct comparison of the key performance indicators between the initial OLS model and the corrected WLS model. This comparison confirms the statistical necessity and efficacy of the **weighted least squares regression** approach when dealing with heteroscedastic data. The primary objective of WLS is to enhance the efficiency of the coefficient estimates, which translates directly into increased precision and a reduction in the model's overall error.

We focus specifically on two crucial metrics: the Residual Standard Error (RSE) and the Multiple R-squared value. The improvements observed in these statistics are substantial:

The Residual Standard Error (RSE) for the weighted least squares model was dramatically reduced to **1.199**. This figure represents a massive improvement over the RSE of **9.224** reported

by the original simple linear regression model. The RSE is an estimate of the standard deviation of the error term, measured in the units of the response variable (exam score). A lower RSE signifies that, on average, the predicted values generated by the WLS model are significantly closer to the actual observed scores than those produced by OLS.
The R-squared value, which measures the proportion of variance in the dependent variable explained by the predictor, increased from **0.6296** in the OLS model to **0.6762** in the WLS model. This increase suggests that after correcting for the unequal variance, the WLS model is capable of explaining a greater percentage of the total variability in the students' exam scores, yielding a statistically superior fit.

In summary, the dramatic reduction in RSE, combined with the higher R-squared, provides compelling evidence that the WLS model is superior. By correctly incorporating the inverse of the estimated error variance through weights, the WLS methodology successfully stabilized the error structure, yielding more precise and reliable coefficient estimates. This robust fit ensures that any subsequent statistical conclusions drawn from the model, such as the estimated increase in score per hour studied (4.7091), are statistically valid and not compromised by the presence of **heteroscedasticity**.

Further Exploration of Robust Regression Techniques

While weighted least squares provides an excellent solution for dealing with error structures where the variance is related to the predictor values, it is not the only robust estimation technique available. Researchers often encounter varying degrees and forms of assumption violations, prompting the need for a diverse toolkit of statistical methods.

For those looking to deepen their understanding of how to manage complex error structures and outliers in regression analysis, further exploration of these related concepts in R is highly recommended. These alternative approaches offer different ways to achieve efficiency and robustness when OLS fails to meet its strict requirements:

Generalized Least Squares (GLS): This technique is a broader generalization of OLS and WLS, which is particularly useful when the errors exhibit both heteroscedasticity and serial or spatial correlation.
Robust Regression (e.g., using M-estimators): These methods are designed primarily to be resistant to outliers in the response variable, offering an alternative when WLS's assumption about the variance structure is difficult to define precisely.
Methods for Transforming Variables to Achieve Homoscedasticity: Techniques such as logarithmic or square-root transformations can sometimes stabilize the variance of the errors, allowing a standard OLS model to be used successfully on the transformed data.