

How to Perform Weighted Least Squares Regression in R?

Authored by
stats writer

December 14, 2025

RECOMMENDED CITATION

stats writer (2025). *How to Perform Weighted Least Squares Regression in R?*.
PSYCHOLOGICAL SCALES. Retrieved from <https://scales.arabpsychology.com/?p=107392>

Understanding the OLS Assumption: Homoscedasticity

One of the fundamental assumptions underpinning standard **Ordinary Least Squares (OLS) regression** is that the random errors, or **residuals**, are distributed with equal variance across all levels of the predictor variables. This critical assumption is formally termed **homoscedasticity**. When data satisfy this condition, the OLS estimator provides the Best Linear Unbiased Estimator (BLUE), meaning our parameter estimates are the most efficient and reliable possible among linear unbiased estimators.

In practical terms, **homoscedasticity** ensures that the predictive power and precision of the model remain consistent, regardless of whether the predictor variable is low or high. If we were predicting house prices, for instance, this assumption means the uncertainty (variance of the errors) in the predicted price should be the same for small starter homes as it is for large mansions. Failing to meet this assumption introduces significant complications into the analysis.

While OLS remains unbiased even when this assumption is violated, the standard errors calculated for the coefficients become unreliable. This unreliability leads to incorrect t-statistics and p-values, potentially causing us to draw erroneous conclusions about the significance of our predictor variables. Therefore, ensuring or addressing **homoscedasticity** is a crucial step in building a trustworthy regression model.

The Problem of Heteroscedasticity

When the assumption of equal variance is violated, we encounter the problem of **heteroscedasticity**. This occurs when the variance of the **residuals** changes systematically across the range of the independent variable. Typically, this presents itself as a pattern where the spread of the residuals increases (or decreases) as the fitted values increase, often visualized as a "cone" shape in residual plots. This unequal spread indicates that the error associated with predicting the outcome is dependent on the magnitude of the predictor.

The primary consequence of **heteroscedasticity** is that the OLS estimators lose their property of being BLUE. They are still unbiased, but they are no longer the most efficient. Observations associated with a larger variance (larger errors) contain less information and should logically have less influence on the parameter estimation than observations with smaller variance. Since OLS treats all observations equally, it does not account for this disparity, leading to inefficient and imprecise estimates of the standard errors.

If left unaddressed, severe heteroscedasticity can render the hypothesis testing procedures, such as t-tests and F-tests, completely invalid. When the standard errors are underestimated, we might falsely conclude that a predictor is statistically significant (Type I error). Conversely, if standard errors are overestimated, we might overlook genuinely important predictors (Type II error).

Therefore, identifying and mitigating this issue is paramount for accurate statistical inference in **linear regression**.

Introducing Weighted Least Squares (WLS) Regression

One robust solution for addressing known or suspected heteroscedasticity is to employ **Weighted Least Squares (WLS) regression**. Unlike OLS, WLS does not assume equal variance across all observations. Instead, it systematically assigns a weight to each observation, effectively allowing the analyst to model the non-constant error variance structure. The core philosophy of WLS is to give more influence (or weight) to the observations that are measured with higher precision, i.e., those with smaller error variance.

In the WLS framework, the weights are defined inversely proportional to the variance of the **residuals**. Specifically, observations that have a small error variance are assigned a larger weight, as they are considered more reliable and informative. Conversely, observations exhibiting a large error variance receive a smaller weight, reducing their impact on the overall fit of the regression line. By implementing this weighting scheme, **WLS regression** returns to the desirable property of BLUE, providing more efficient and reliable coefficient estimates than standard OLS when heteroscedasticity is present.

The successful application of **WLS regression** hinges entirely on correctly defining these weights. While sometimes the variance structure is known theoretically, such as when data points are averages of different numbers of underlying measurements, often the variance structure must be estimated empirically from the data itself. This tutorial provides a step-by-step example of how to perform **weighted least squares regression** in R using an empirical estimation approach.

Step 1: Preparing the Sample Data in R

To illustrate the methodology of WLS regression, we will begin by creating a simple dataset. This dataset simulates a typical scenario in educational statistics, where we examine the relationship between the number of hours a student spends studying and their corresponding exam score. We will intentionally use a small sample size to keep the example manageable, though real-world applications usually involve much larger datasets.

The following R code snippet initializes a data frame named `df`. It contains two variables: `hours`, representing the time spent studying (our predictor), and `score`, representing the resulting exam score (our response variable). The data captures results for 16 hypothetical students:

```
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))
```

Examining the raw data, we can observe that for small values of `hours` (e.g., 1 hour), the scores show a high degree of variability (48 vs. 78). This initial observation hints that the variance in scores might be higher when study time is low, suggesting that heteroscedasticity may be a factor we need to account for, reinforcing the need for formal testing.

Step 2: Fitting the Initial OLS Model

Before implementing WLS, it is standard practice to fit the conventional **Ordinary Least Squares** model. This initial model serves two main purposes: first, to establish a baseline performance metric, and second, to generate the residuals necessary for diagnosing the assumption violations. We utilize the built-in `lm()` function in R to perform this standard **linear regression** analysis, using `hours` as the predictor variable and `score` as the response variable.

The following command fits the model and then displays the summary output, which includes the estimated coefficients, standard errors, and overall model fit statistics like R-squared and the Residual Standard Error (RSE):

```
#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

From the output of this standard OLS model, we see that the coefficient for `hours` is 5.500, suggesting that for every additional hour studied, the predicted exam score increases by 5.5 points. This coefficient is highly statistically significant. The model explains approximately 63% of the variance in scores (Multiple R-squared: 0.6296), and the Residual Standard Error (RSE) is 9.224. This RSE represents the typical deviation between the observed scores and the scores predicted by the model. If **homoscedasticity** holds, this model is adequate; otherwise, these results are inefficient.

Step 3: Detecting Heteroscedasticity Visually and Formally

The crucial next step is to test the assumption of constant variance. This is typically done through a combination of visual inspection and formal statistical testing. Visual inspection provides an intuitive understanding of the residual patterns, while formal tests provide objective statistical evidence.

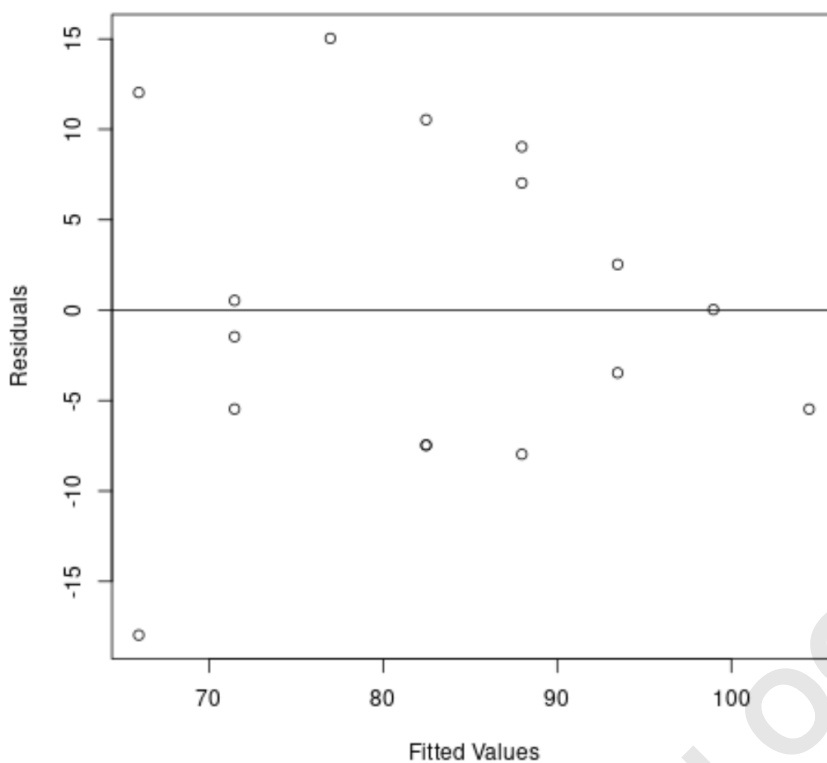
First, we create a residual vs. fitted values plot, which is the standard graphical tool for detecting heteroscedasticity. If the assumption of constant variance were met, the points in this plot should exhibit a random scatter around the zero line, forming a roughly horizontal band of equal width.

#create residual vs. fitted plot

```
plot(fitted(model), resid(model), xlab='Fitted Values', ylab='Residuals')
```

#add a horizontal line at 0

```
abline(0,0)
```



Upon visual inspection of the resulting plot, a distinct pattern is evident: the spread of the residuals, which represents the model error, increases as the fitted values increase. The residuals fan out into a clear "cone" shape. This graphical evidence strongly suggests that the variance is not constant, thus confirming the presence of **heteroscedasticity**.

To formally confirm this finding, we utilize the **Breusch-Pagan test**. This test assesses whether the variance of the errors is related to the values of the independent variables. We first need to load the `lmtest` package in R to access the `bptest()` function.

```
#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
```

The hypothesis structure for the **Breusch-Pagan test** is as follows:

Null Hypothesis (H0): Homoscedasticity is present (the error variance is constant).

Alternative Hypothesis (HA): Heteroscedasticity is present (the error variance is non-constant).

Since the resulting p-value (**0.0466**) is marginally less than the conventional significance level of 0.05, we must reject the null hypothesis (H0). Both the visual evidence and the formal **Breusch-Pagan test** confirm that heteroscedasticity is a statistically significant problem in our initial OLS model, necessitating the use of a robust alternative method like **Weighted Least Squares regression**.

Step 4: Defining and Applying Weights for WLS

To correct for the varying error structure, we must define the weights (w_i). In empirical estimation, a common heuristic approach is to model the variance of the errors as a function of the independent variable or the fitted values. Since the variance (or standard deviation) of the errors appears to increase with the fitted score in our example, we will model the weights based on the fitted values obtained from the initial OLS model.

The optimal weight is defined as the reciprocal of the variance of the error terms: $w_i = 1 / \text{Var}(\epsilon_i)$. When $\text{Var}(\epsilon_i)$ is unknown, a two-step procedure is often used: first, model the absolute residuals (a proxy for the standard deviation of the error) as a function of the fitted values, and second, use the inverse of the square of these predicted standard deviations as the weights. The inverse squaring is necessary because the weights must be inversely proportional to the variance, not the standard deviation.

The R code below implements this two-step estimation. We fit a temporary **linear regression** where the absolute values of the OLS model residuals are regressed against the OLS fitted values. We then define the weights (w_t) as the reciprocal of the squared fitted values from this temporary model. This method ensures that observations associated with smaller predicted errors receive substantially higher weights.

#define weights to use (Inverse of the squared predicted standard deviations)

```
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)
```

Step 5: Interpreting the WLS Model Results

Once the weights are calculated and applied, we fit the final WLS model using the `lm()` function, simply specifying the newly created `wt` vector in the `weights` argument. This results in the WLS model, which minimizes the weighted sum of squared errors, giving less penalty to errors

associated with observations having high inherent variance.

#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

Reviewing the summary output, we observe several key differences compared to the initial OLS model. Firstly, the coefficient estimate for `hours` has changed from 5.500 to 4.7091. This shift is significant; because the **WLS model** down-weighted the observations with high variance, the overall slope relating hours to score has become slightly flatter. This adjustment provides a more accurate and efficient estimate of the underlying linear relationship when accounting for the non-constant error structure.

Furthermore, the standard error associated with the `hours` coefficient has decreased substantially, moving from 1.127 in the OLS model to 0.8709 in the WLS model. This reduction in the standard error confirms the theoretical advantage of WLS: by correctly accounting for heteroscedasticity, we achieve more precise estimates, leading to a higher t-value and a lower p-value. This increase in precision is the primary goal of using WLS.

Comparing WLS and OLS Performance Metrics

A direct comparison of the overall model fit statistics between the OLS and WLS models clearly illustrates the benefits of employing the weighted approach when dealing with non-constant variance. We should focus specifically on the Residual Standard Error (RSE) and the R-squared

value, as these metrics quantify the goodness of fit.

Residual Standard Error (RSE): The original simple linear regression (OLS) model had an RSE of **9.224**. The **Weighted Least Squares** model dramatically reduced this value to **1.199**. This massive reduction indicates that, after applying the weights, the predicted scores produced by the WLS model are significantly closer to the actual, observed scores. The WLS framework provides a much tighter fit to the data once the heteroscedastic nature of the errors is factored in.

R-squared: The OLS model achieved a Multiple R-squared of **0.6296**. The WLS model, despite having a slightly flatter slope, resulted in an improved R-squared value of **0.6762**. This improvement demonstrates that the WLS model is capable of explaining a larger proportion of the total variance in the exam scores, offering superior explanatory power compared to the standard OLS approach.

These statistical metrics collectively confirm that the **WLS regression** model provides a substantially better, more efficient, and more reliable fit to the data than the initial OLS model. By appropriately down-weighting observations with high error variability and up-weighting those with low error variability, WLS ensures that the model parameters are estimated with optimal precision, fulfilling the requirement for BLUE estimators in the presence of heteroscedasticity.

Conclusion: When to Use WLS

This tutorial demonstrated that while linear regression is a powerful tool, its reliability depends heavily on satisfying underlying assumptions, particularly **homoscedasticity**. When diagnostics, such as the residual plot or the **Breusch-Pagan test**, indicate the presence of significant heteroscedasticity, standard OLS estimates become inefficient, and their associated standard errors become untrustworthy.

The application of **WLS regression** provides a viable and robust alternative. By meticulously calculating weights that reflect the inverse variance of the error terms--even when that variance must be empirically estimated--we restore the efficiency of the estimators. The resulting WLS model yields coefficient estimates that are less sensitive to high-variance outliers and provides more accurate statistical inference, as demonstrated by the improved RSE and R-squared metrics in our study example.

It is essential to note that WLS should be considered whenever the inherent precision of observations differs systematically. If the pattern of variance is complex or unknown, alternative methods like using Heteroscedasticity-Consistent Covariance Matrix Estimators (such as Huber-White standard errors) might be preferred, as they adjust the standard errors without modifying the coefficient estimates. However, if the functional form of the heteroscedasticity can be reasonably modeled, WLS remains the preferred method for achieving the most efficient parameter estimates possible.