Understanding Multiple Linear Regression: Exploring its Core Assumptions

Authored by

Mohammed looti

November 1, 2025

Multiple Linear Regression (MLR) is a powerful **statistical method** used to model the relationship between several independent variables, known as **predictor variables**, and a single continuous dependent variable, often called the **response variable**. It is essential in fields ranging from economics to engineering for predictive modeling and understanding variable influence.

However, the validity and reliability of the MLR results depend critically on meeting several underlying statistical prerequisites. Before conducting any analysis, researchers must ensure five core assumptions are satisfied. Violating one or more of these assumptions can lead to unreliable conclusions, inflated standard errors, and incorrect hypothesis testing.

The five critical assumptions for Multiple Linear Regression are:

Linear relationship: A linear relationship must exist between each predictor variable and the response variable.

No Multicollinearity: Predictor variables must not be highly correlated with each other.

Independence: All observations and their respective residuals must be independent of one another.

Homoscedasticity: The variance of the residuals must be constant across all levels of the predictor variables.

Multivariate Normality: The residuals of the model must be normally distributed.

This article provides a detailed explanation of each assumption, outlining the diagnostic steps required to verify them and offering practical remedies when violations occur.

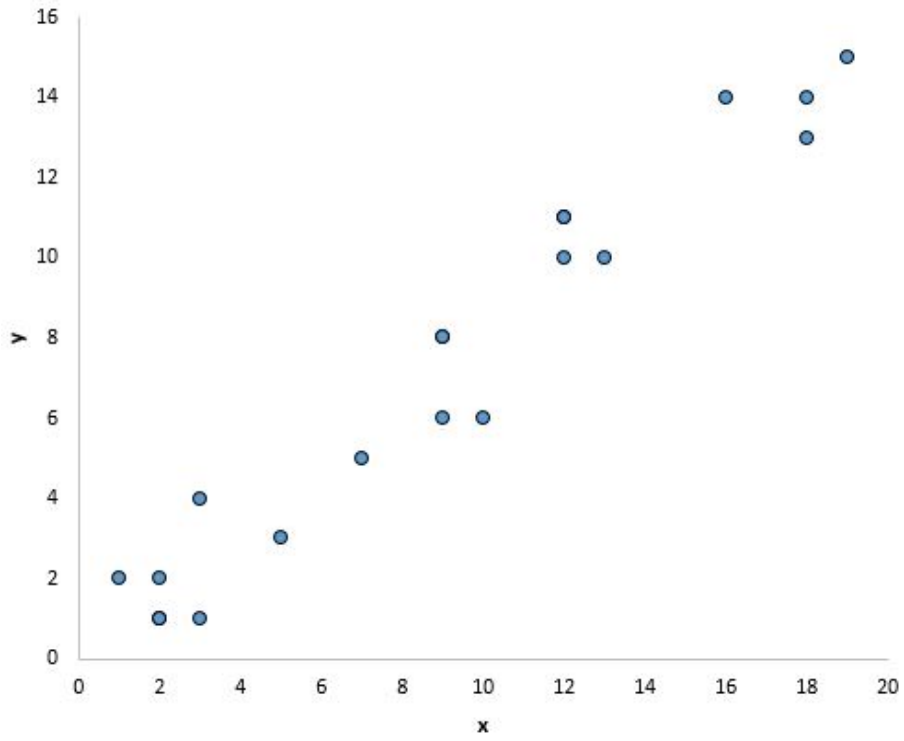
Assumption 1: Linear Relationship

The foundation of Multiple Linear Regression is the premise that the relationship between the independent variables (predictors) and the dependent variable (response) can be accurately modeled using a straight line. If the true relationship is non-linear--for instance, parabolic or exponential--the MLR model will inaccurately estimate the coefficients, leading to poor predictive power.

How to Determine if this Assumption is Met

The most straightforward method for assessing linearity is through visual inspection. By generating a **scatter plot** for each individual predictor variable against the response variable, analysts can visually confirm whether the data points generally align along a straight diagonal path. This graphical representation is highly intuitive and quickly reveals potential non-linear patterns.

If the points in the scatter plot exhibit a clear linear trend, meaning they roughly follow a straight line, the assumption is likely satisfied for that specific variable. Consider the example below, where the relationship between variable X (predictor) and variable Y (response) appears strongly linear:



What to Do if this Assumption is Violated

When a lack of linearity is detected for one or more predictor variables, several corrective actions can be taken to improve the model:

Apply a Nonlinear Transformation: Often, applying a mathematical transformation--such as taking the **logarithm** (log) or the square root of the predictor variable--can linearize a curved relationship. This transformation often makes the variable suitable for inclusion in the linear model.

Introduce Polynomial Terms: If the plot of x versus y exhibits a clear parabolic shape, it might make sense to include a squared term, such as X2, as an additional predictor variable in the model. This allows the linear framework to capture the non-linear curvature.

Drop the Predictor Variable: In the most extreme case, if no linear relationship exists between a specific predictor variable and the response variable, and transformations fail to yield improvement, the predictor variable may not be useful and should be excluded from the model.

Assumption 2: No Multicollinearity

Multiple Linear Regression assumes that none of the predictor variables are highly correlated with each other. When one or more predictor variables are highly correlated, the regression model suffers from **Multicollinearity**.

Severe collinearity causes the coefficient estimates in the model to become highly unstable and unreliable. This difficulty in isolating the unique effect of each variable makes it nearly impossible to interpret the true impact of individual predictors, as slight changes in the data can lead to large swings in the estimated coefficients.

How to Determine if this Assumption is Met

The easiest and most rigorous way to determine if this assumption is met is to calculate the Variance Inflation Factor (VIF) value for each predictor variable. The VIF quantifies how much the variance of an estimated regression coefficient is inflated due to collinearity with other predictors.

VIF values start at 1 and have no upper limit. As a general rule of thumb, VIF values greater than 5* indicate potential multicollinearity, though some researchers use a more conservative threshold of 10, depending on the research context.

The following tutorials show how to calculate VIF in various statistical software:

(Placeholder for VIF tutorial link 1)
(Placeholder for VIF tutorial link 2)
(Placeholder for VIF tutorial link 3)

*The appropriate VIF threshold often depends on the field of study and data characteristics.

What to Do if this Assumption is Violated

If one or more of the predictor variables has a VIF value greater than the established threshold, the following options are available:

Remove High VIF Variables: The simplest way to resolve this issue is to remove the predictor variable(s) with the highest VIF values. Since these variables are largely redundant, their removal stabilizes the model parameters.

Use Alternative Methods: Alternatively, if retaining every predictor variable is mandatory, analysts can employ different statistical techniques such as **Ridge Regression**, **Partial Least Squares Regression**, or **Principal Component Regression**, which are specifically designed to handle predictor variables that are highly correlated.

Assumption 3: Independence of Observations

Multiple Linear Regression requires that each observation in the dataset is independent. This extends directly to the model's residuals: there should be no systematic relationship between the residual of one observation and the residual of the next. Violation of this assumption, known as **autocorrelation** or serial correlation, is most common in time-series data.

When observations are dependent, the standard errors calculated by the model are incorrect. This can severely distort hypothesis testing, potentially leading researchers to erroneously conclude that predictor variables are statistically significant.

How to Determine if this Assumption is Met

The simplest and most common method to determine if this assumption is met is to perform a Durbin-Watson test. This formal statistical test assesses whether the residuals exhibit autocorrelation. A Durbin-Watson statistic close to 2 indicates that the observations (and thus the residuals) are independent.

What to Do if this Assumption is Violated

Depending on the nature of the dependence, several options are available to correct this violation:

For **positive serial correlation**, consider augmenting the model by adding lagged values of the dependent variable and/or independent variables as new predictors.

For **negative serial correlation**, verify that none of your variables have been *overdifferenced*. Overdifferencing can artificially introduce negative correlation patterns.

For **seasonal correlation** (often seen in quarterly or monthly data), consider adding **seasonal dummy variables** to the model to account for recurring temporal patterns.

Assumption 4: Homoscedasticity

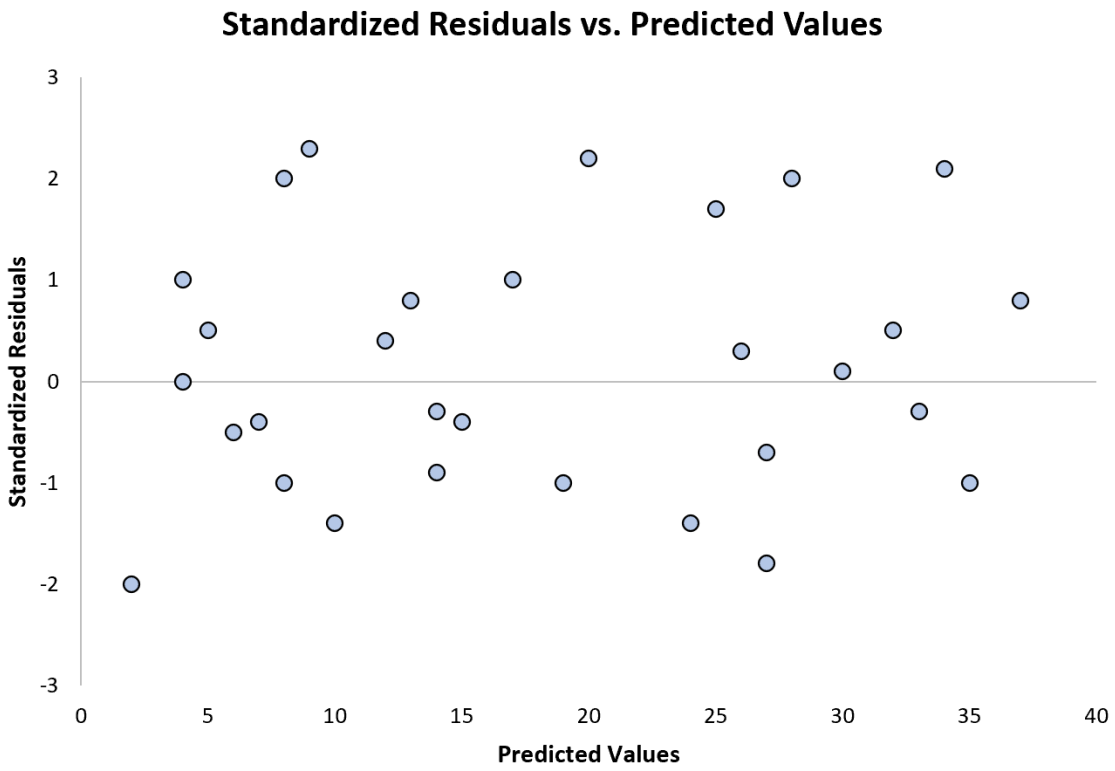
This assumption mandates that the residuals have constant variance across all predicted values in the linear model. When this condition is not met, the residuals exhibit non-constant variance, a phenomenon known as **Heteroscedasticity**.

When heteroscedasticity is present, the model's results become unreliable for inference. The primary issue is that heteroscedasticity increases the variance of the regression coefficient estimates, yet the model fails to detect this increase. This misestimation of variance makes it much more likely for a researcher to incorrectly declare a term in the model to be statistically significant.

How to Determine if this Assumption is Met

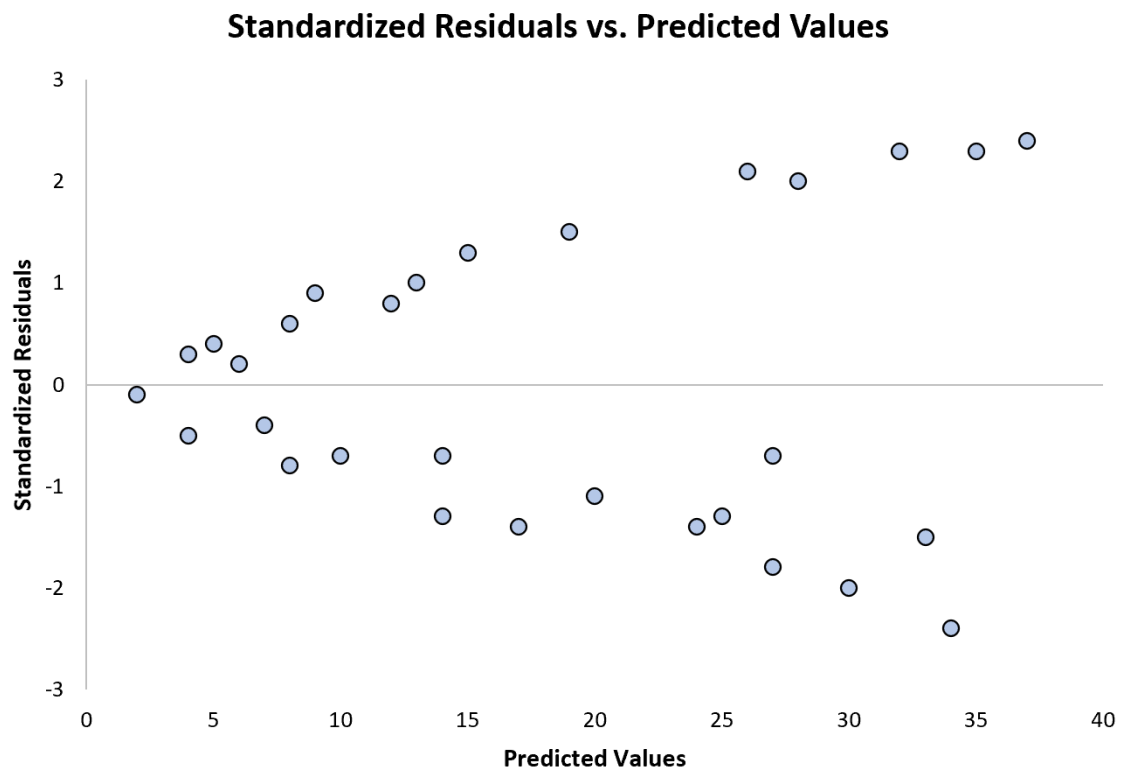
The most effective way to check for Homoscedasticity is to create a **plot of standardized residuals versus predicted values**.

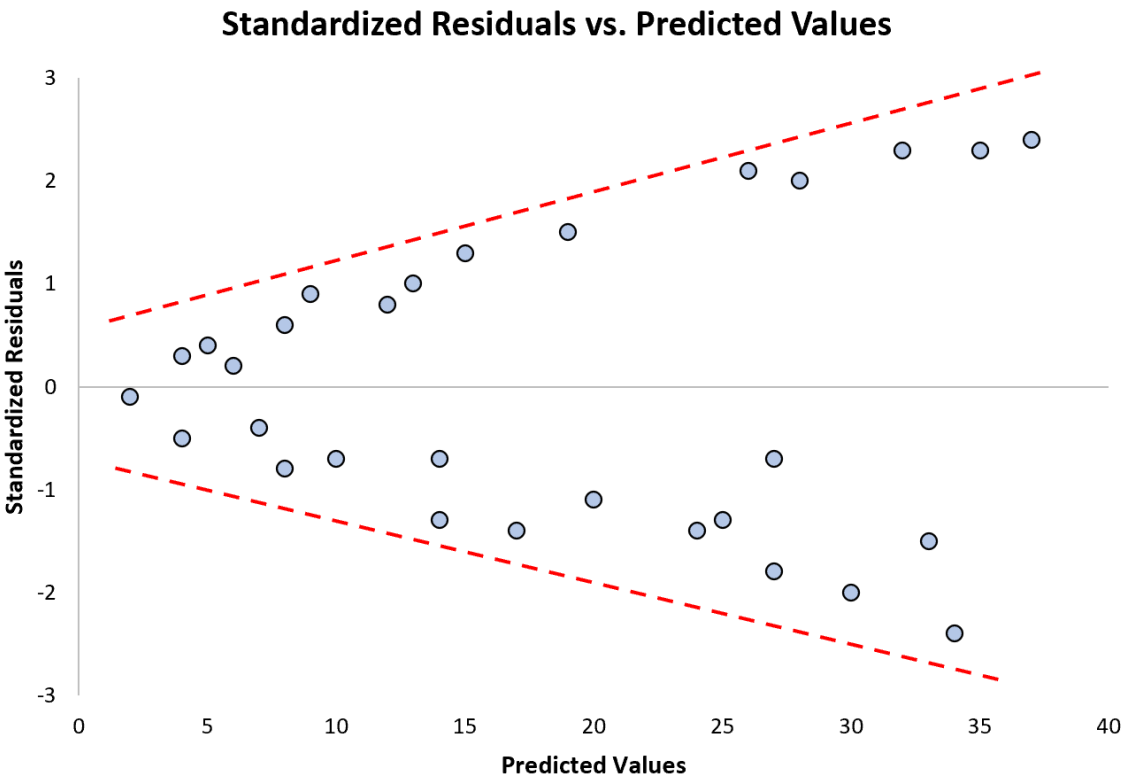
Once the regression model is fitted, a scatter plot is generated showing the predicted values for the response variable on the x-axis and the standardized residuals of the model on the y-axis. If the points are randomly scattered about zero with no clear pattern (as shown below), the assumption is met:



Conversely, if the standardized residuals exhibit a pattern--for instance, becoming wider or narrower as the predicted values increase--then Heteroscedasticity is present.

The following plot shows an example where the standardized residuals spread out as the predicted values increase, forming the classic "cone" shape indicative of heteroscedasticity:


Standardized Residuals vs. Predicted Values
Standardized Residuals
Predicted Values
3
2
1
0
-1
-2
-3
0
5
10
15
20
25
30
35
40


Standardized Residuals vs. Predicted Values
Standardized Residuals
Predicted Values
3
2
1
0
-1
-2
-3
0
5
10
15
20
25
30
35
40

What to Do if this Assumption is Violated

There are three common ways to address heteroscedasticity when it is detected:

Transform the Response Variable: The most frequent remedy is to apply a nonlinear transformation (such as the log, square root, or cube root) to all values of the response variable. This often successfully stabilizes the variance.

Redefine the Response Variable: Sometimes, the variability can be reduced by redefining the response variable as a *rate* rather than a raw value. For instance, instead of predicting the raw number of flower shops, one might predict the number of flower shops **per capita**, which inherently controls for overall population size and typically reduces variability in larger data points.

Use Weighted Regression: Another robust solution is **Weighted Least Squares Regression** (WLS). WLS assigns weights to each data point inversely proportional to the variance of its fitted value. This technique gives small weights to data points with higher variability, thereby minimizing the influence of high-variance observations and effectively eliminating the problem of heteroscedasticity.

Related:

Assumption 5: Multivariate Normality of Residuals

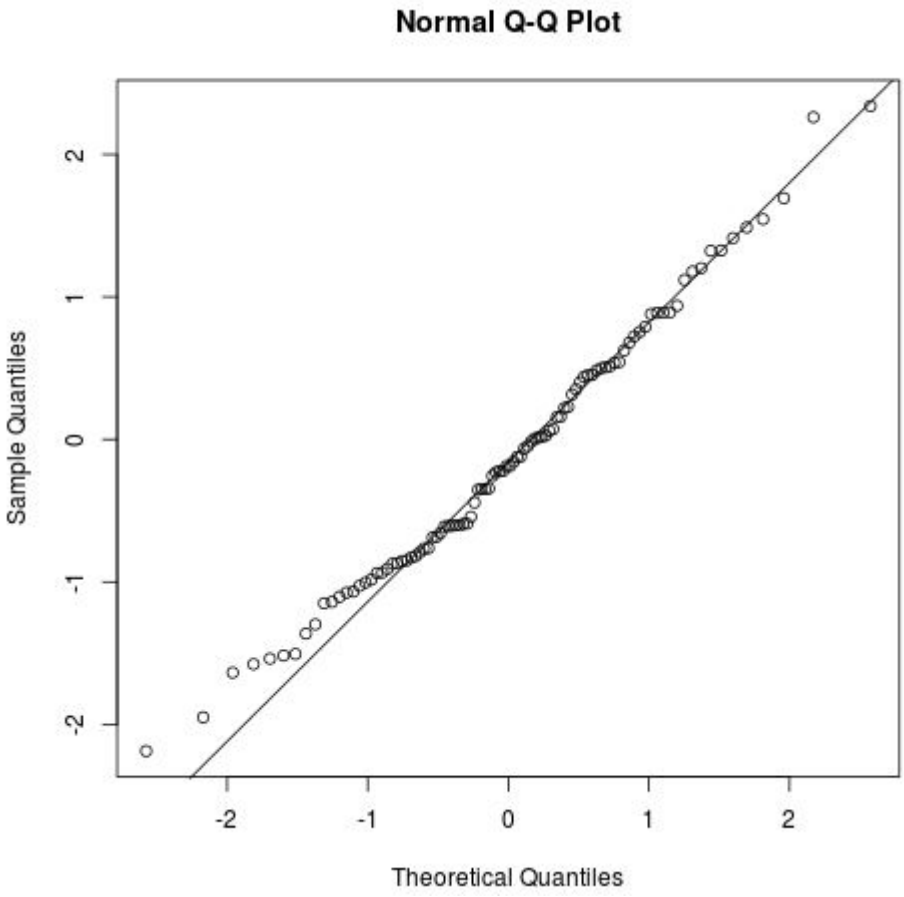
The final assumption for robust statistical inference is that the model's residuals are normally distributed. While Multiple Linear Regression is relatively robust to minor deviations from normality, particularly with large sample sizes, severe violations can undermine the confidence intervals and significance tests.

How to Determine if this Assumption is Met

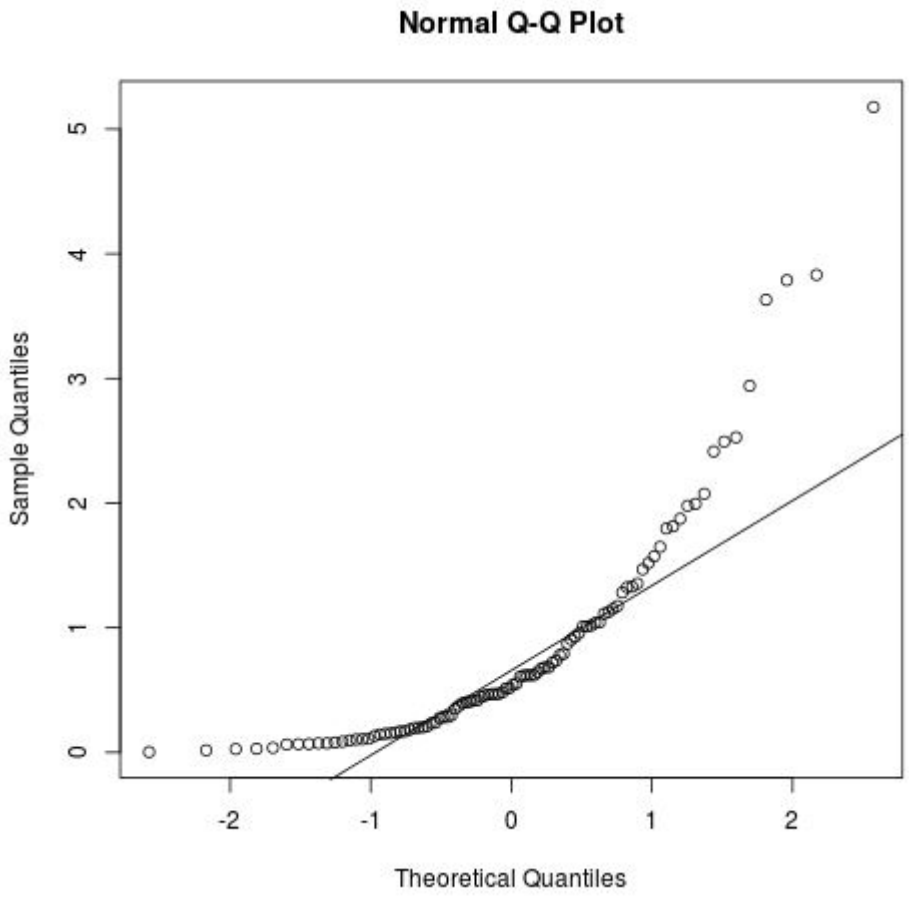
There are two primary methods for verifying the normality assumption:

Visual Check using a Q-Q Plot: A quantile-quantile plot compares the observed residuals against the theoretical quantiles of a normal distribution. If the residuals follow a normal distribution, the points on the plot will closely form a straight diagonal line.

The following Q-Q plot illustrates residuals that adhere to the normality assumption:



Conversely, the plot below shows residuals that clearly depart from the straight diagonal line, indicating non-normality:



Formal Statistical Tests: Normality can also be checked using formal tests such as **Shapiro-Wilk**, **Kolmogorov-Smirnov**, **Jarque-Barre**, or **D'Agostino-Pearson**. However, analysts should exercise caution with these tests, as they are highly sensitive to large sample sizes--often concluding that the residuals are non-normal even when the practical impact is minimal. For this reason, graphical methods like the Q-Q plot are often preferred.

What to Do if this Assumption is Violated

If the normality assumption is violated, these steps should be considered:

Verify Outliers: First, confirm that the non-normality is not simply due to the presence of a few **extreme outliers** in the data set. Addressing or removing problematic outliers can often resolve the issue.

Apply Transformation: Next, apply a nonlinear transformation (such as taking the square root, the log, or the cube root) to the **response variable**. Transforming the dependent variable often successfully causes the model's residuals to become more normally distributed.

Additional Resources and Next Steps

The following tutorials provide additional information about multiple linear regression and its assumptions:

The following tutorials provide step-by-step examples of how to perform multiple linear regression using different statistical software: