

What are Residuals in Statistics?

Authored by
stats writer

December 17, 2025

RECOMMENDED CITATION

stats writer (2025). *What are Residuals in Statistics?*. PSYCHOLOGICAL SCALES.
Retrieved from <https://scales.arabpsychology.com/?p=107659>

In the field of statistics, **residuals** are fundamental metrics used to evaluate the performance of a predictive model. Formally, a residual represents the vertical distance--the error--between an actual, **observed value** in a dataset and the **predicted value** generated by the statistical model, typically within the context of regression analysis.

These discrepancies are crucial because they quantify the accuracy of the model's estimations for individual data points. By analyzing the characteristics and patterns within these residuals, statisticians can gain insights into the model's suitability, detect potential flaws, or identify significant outliers that might skew the overall analysis.

A **residual** is fundamentally the difference between an observed value and a predicted value in statistical modeling.

Understanding the Residual Calculation

The calculation of a residual is straightforward and defined by the simple arithmetic difference between the data's true value and the model's estimate. The formula below is universally applied in statistical modeling:

Residual = Observed value - Predicted value

To contextualize this, consider the primary objective of linear regression: to define and quantify the relationship between one or more independent (predictor) variables and a single dependent, or response variable. This process involves fitting a straight line through the data points, which is mathematically optimized to minimize the sum of squared errors--hence its name, the **least squares regression line**.

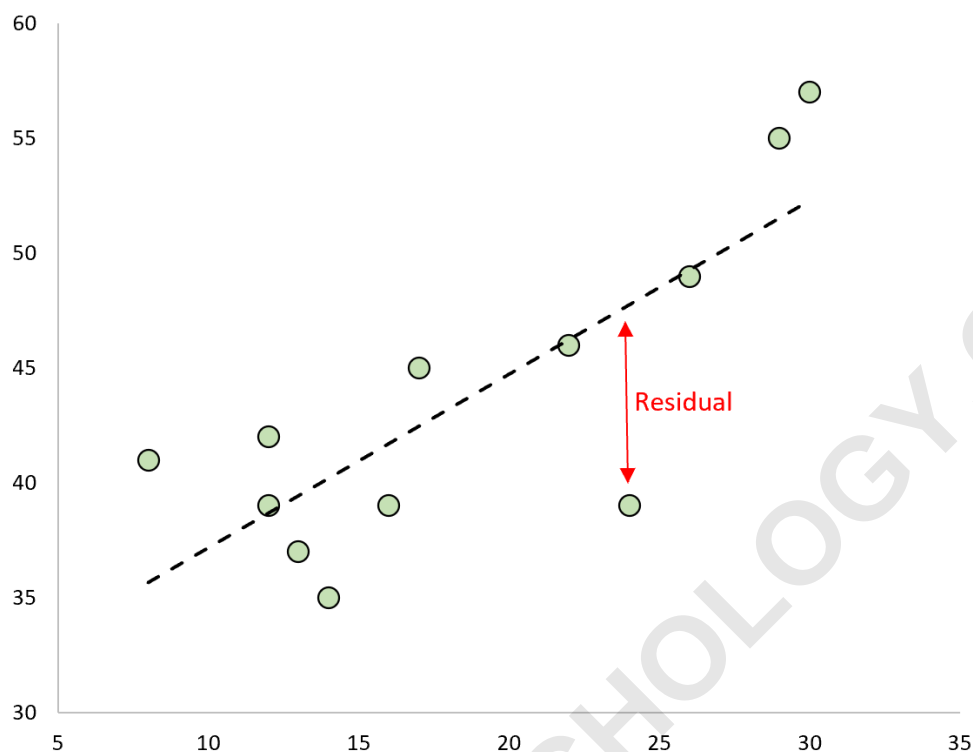
This optimal line generates a prediction for every single data observation in the sample. However, it is exceedingly rare in real-world data for the prediction generated by the regression line to perfectly align with the observed value. The inevitable gap between these two values is precisely what the residual measures, providing a vital tool for error analysis.

Visualizing Residuals: The Vertical Error

When we visualize a regression model, the meaning of the residual becomes geometrically clear. Imagine a scatterplot where the data points (the observed values) are plotted, and the fitted **least squares regression line** is superimposed. The residual for any given data point is the **vertical distance** between that observation and the regression line.

If the residual is small, the model performed well for that particular observation; if the residual is large, the model significantly missed the mark. This visual representation allows analysts to quickly

grasp how tightly the model adheres to the underlying data structure. The image below illustrates this concept, showing the distinct vertical segments that represent the residuals for various data points relative to the line of best fit.

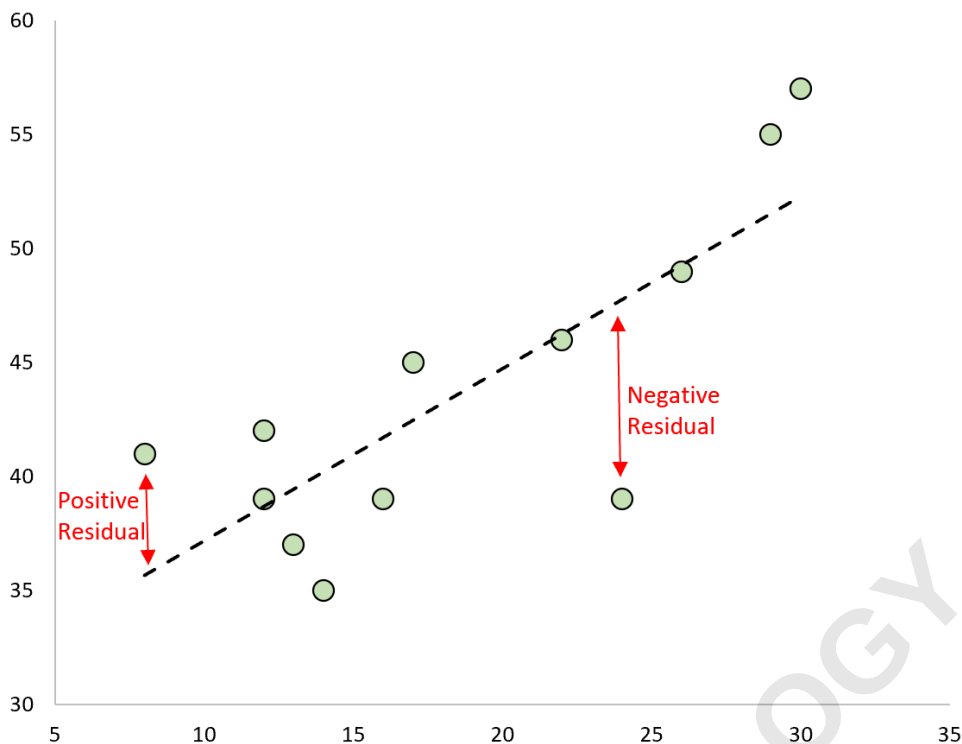


Positive vs. Negative Residuals

Residuals can be categorized based on their sign, providing immediate information about whether the model overestimated or underestimated the observed value. Understanding this directional difference is key to interpreting model errors.

A data point is said to have a **positive residual** when its actual, observed value is higher than the value predicted by the regression line. In visual terms, the point lies vertically above the fitted line, indicating that the model systematically underestimated the outcome for that specific instance.

Conversely, an observation registers a **negative residual** when its observed value falls below the predicted value. This means the model overestimated the outcome, and the data point is located vertically beneath the regression line. Across a well-fitted model, observations will typically be balanced, resulting in a mixture of positive and negative residuals, as shown in the visualization below.



A fundamental property of the least squares method is that the sum of all individual residuals within the model dataset must aggregate to exactly **zero**. This cancellation effect is essential for ensuring the line of best fit passes through the mean of the data cloud.

Step-by-Step Example of Calculating Residuals

To solidify the theoretical definition, let us walk through a practical example of how residuals are derived from a simple dataset. Suppose we begin with twelve total observations, each pairing an X value (predictor) with a corresponding Y value (observed response):

X	Y
8	41
12	42
12	39
13	37
14	35
16	39
17	45
22	46
24	39
26	49
29	55
30	57

When this data is processed using common statistical software (such as R, Excel, or Python), the algorithm determines the optimal line of best fit. For this particular dataset, the resulting equation for the least squares regression line is:

$$y = 29.63 + 0.7553x$$

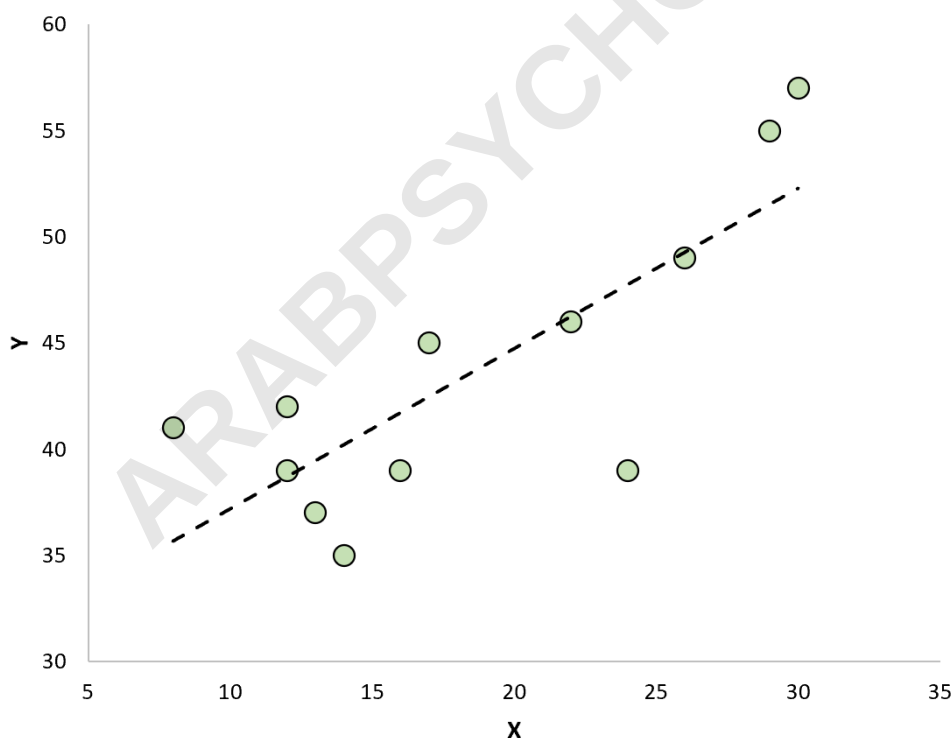
Using this formula, we can predict the Y value for any given X. For instance, considering the first observation where $X=8$, the predicted value is calculated as $Y_{\text{predicted}} = 29.63 + 0.7553(8) = 35.67$. The observed Y value for this point was 41. We then calculate the residual:

$$\text{Residual} = \text{Observed value} - \text{Predicted value} = 41 - 35.67 = \mathbf{5.33}$$

This positive residual indicates that the model underestimated the outcome by 5.33 units. By applying this simple calculation across all twelve observations, we generate a complete set of residuals, highlighting the specific error for every data point, as summarized in the table below:

X	Y	Predicted Value	Residual
8	41	35.67	5.33
12	42	38.69	3.31
12	39	38.69	0.31
13	37	39.45	-2.45
14	35	40.20	-5.20
16	39	41.71	-2.71
17	45	42.47	2.53
22	46	46.25	-0.25
24	39	47.76	-8.76
26	49	49.27	-0.27
29	55	51.53	3.47
30	57	52.29	4.71

When these results are plotted on a scatterplot alongside the fitted regression line, the distribution of residuals becomes visually apparent. Points corresponding to positive residuals lie above the line, while those with negative residuals fall below. This visualization is essential for qualitative assessment of the model fit.



Key Statistical Properties of Residuals

Residuals possess distinct mathematical properties that are direct consequences of using the least squares method to fit the regression line. These properties are critical for statistical theory and for validating the assumptions required for reliable inference.

The following list summarizes the inherent characteristics of residuals derived from a simple linear regression model:

One-to-One Correspondence: Every single observation used in fitting the model results in a unique residual. If the training dataset contains 100 observations, the model will inherently generate 100 predicted values, resulting in an equivalent count of 100 calculated residuals.

Sum is Zero: A defining characteristic of the least squares method is that the algebraic sum of all residuals in the sample must equal **zero**. This guarantees that the line of best fit is centrally positioned among the data points.

Mean is Zero: Directly stemming from the property above, if the sum of the residuals is zero, it follows logically that the average (mean) value of the residuals must also be zero.

Practical Application 1: Assessing Overall Model Fit (RSS)

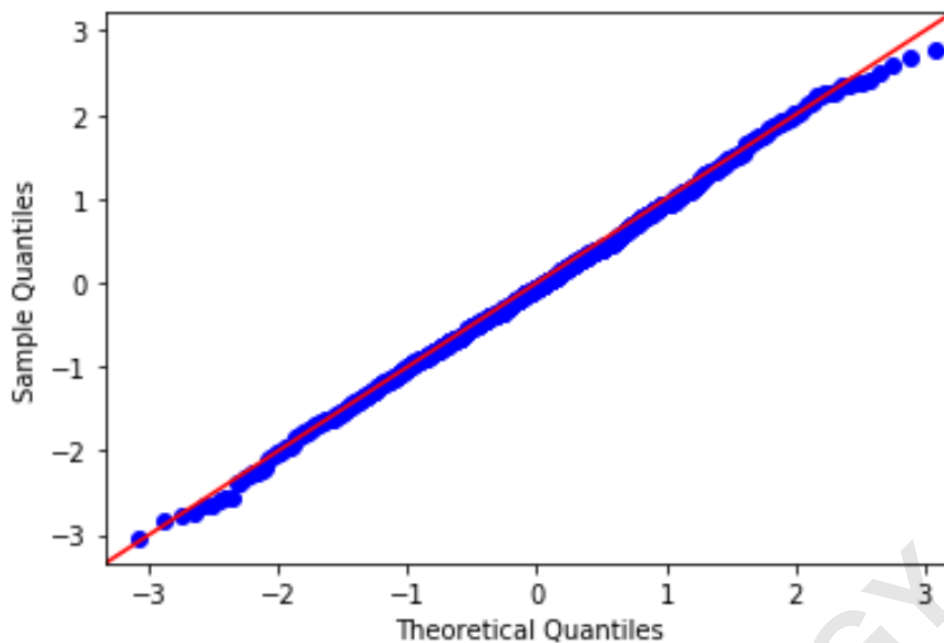
In real-world statistical modeling, residuals serve several vital functions beyond simply calculating prediction errors. One of their most frequent uses is to quantitatively assess the overall goodness-of-fit of a regression model. This is primarily achieved through the calculation of the **Residual Sum of Squares (RSS)**.

RSS is derived by squaring each individual residual and then summing these squared values together. By squaring the errors, we eliminate the issue of positive and negative residuals canceling each other out, thereby obtaining a true measure of the total unexplained variation in the model. A fundamental principle in regression is that the lower the RSS, the tighter the fit between the predicted values and the observed data, signifying a statistically superior model.

Practical Application 2: Checking the Normality Assumption

A crucial requirement for valid statistical inference in linear regression is that the random errors--represented by the residuals--must be normally distributed. If this assumption is severely violated, the confidence intervals and hypothesis tests derived from the model may be unreliable.

To visually inspect this requirement, statisticians employ a specialized graph called a **Q-Q plot** (Quantile-Quantile plot). This plot compares the distribution of the calculated residuals against the expected quantiles of a theoretical normal distribution. If the residuals adhere to the assumption of normality, the points plotted on the Q-Q plot should congregate closely along a straight, diagonal reference line. Significant deviations suggest that the underlying distribution of errors is non-normal.



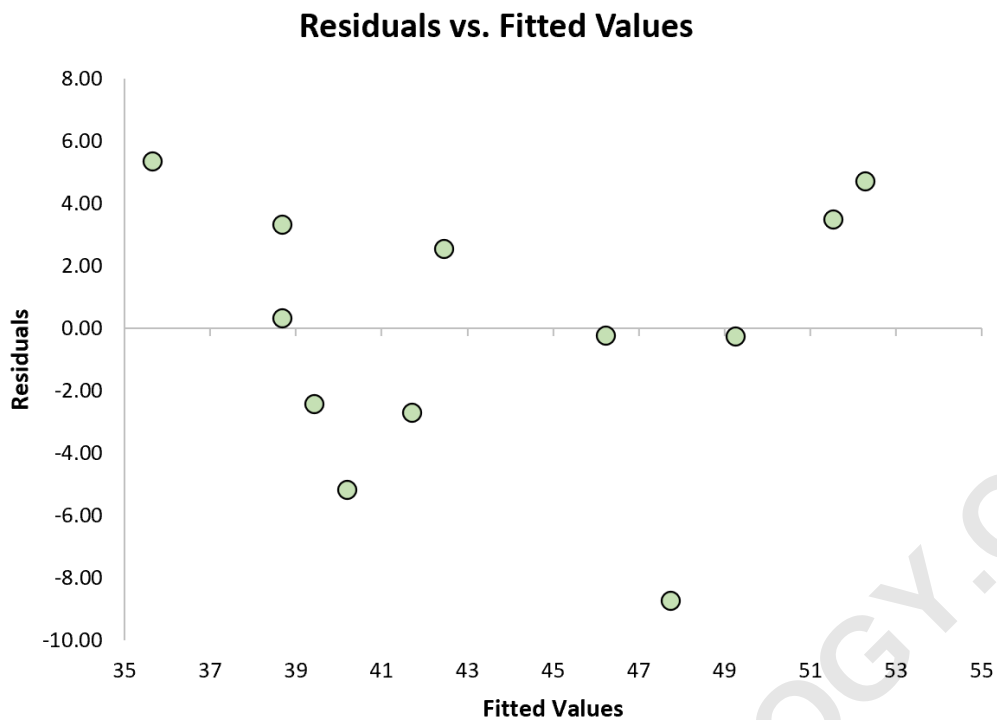
Example of a Q-Q

plot for checking normality

Practical Application 3: Checking for Homoscedasticity

Another core tenet of classical regression analysis is the assumption of **homoscedasticity**. This assumption dictates that the variance (spread) of the residuals must remain constant across all levels of the predictor variable (X). In simpler terms, the error of prediction should be consistent regardless of whether we are predicting low or high values of the response variable.

If the residual variance changes systematically across the range of X values--for example, if the errors become larger for higher predicted values--the model suffers from heteroscedasticity, which can lead to biased standard errors. To evaluate this crucial assumption, a **residual plot** is created, showing the residuals plotted against the fitted (predicted) values.



Example of residual

vs. fitted values plot

A plot that adheres to homoscedasticity will show residuals scattered randomly and uniformly around the zero horizontal line, resembling a horizontal band or cloud without any noticeable pattern, such as a cone shape (which would indicate heteroscedasticity).

Related Statistical Concepts

Explore further resources related to regression analysis and model validation:

[Introduction to Simple Linear Regression](#)

[Introduction to Multiple Linear Regression](#)

[The Four Assumptions of Linear Regression](#)

[How to Create a Residual Plot in Excel](#)