How to Find Residuals in Regression: A Step-by-Step Guide

Authored by

stats writer

December 30, 2025

Understanding how to calculate residuals is fundamental to mastering regression analysis. A residual represents the discrepancy between the observed data point and the value predicted by the regression model. Specifically, it is calculated as the difference between the actual observed value of the dependent variable (y) and its corresponding predicted value (?).

These residuals are crucial for assessing the fit and validity of a statistical model. While the calculation is simple (actual value minus the predicted value), the analysis of residuals provides profound insight into whether the model is appropriate and if its underlying statistical assumptions are met. Furthermore, in standard Ordinary Least Squares (OLS) regression, the sum of these residuals is used as a foundational element in evaluating the overall fit and reliability of the model parameters.

The Role of Regression in Statistical Modeling

Regression analysis is a powerful statistical framework utilized to investigate and model the linear relationship between two or more variables. This method allows analysts to determine how changes in one variable correlate with changes in another, enabling accurate prediction and inference about real-world phenomena.

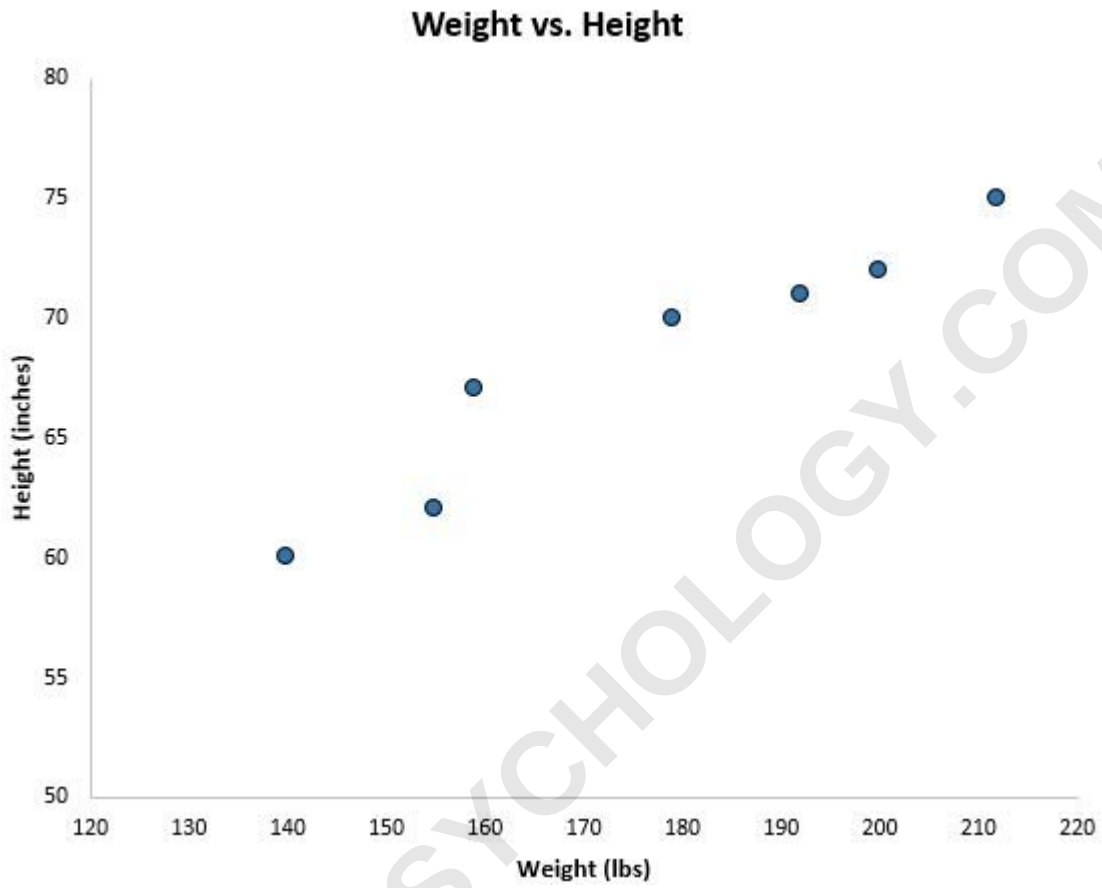
In the context of simple linear regression, we examine two main variables: the independent variable, often denoted as **x**, and the dependent variable, denoted as **y**. The variable **x** is formally known as the **predictor variable**, as its value is used to predict the outcome. Conversely, the variable **y** is known as the **response variable**, as its value is presumed to depend on **x**.

To illustrate this concept, consider a common example exploring the relationship between physical attributes. Suppose we collect a dataset containing the weight and height measurements for a small sample of seven individuals:

Weight (lbs)	Height (inches)
140	60
155	62
159	67
179	70
192	71
200	72
212	75

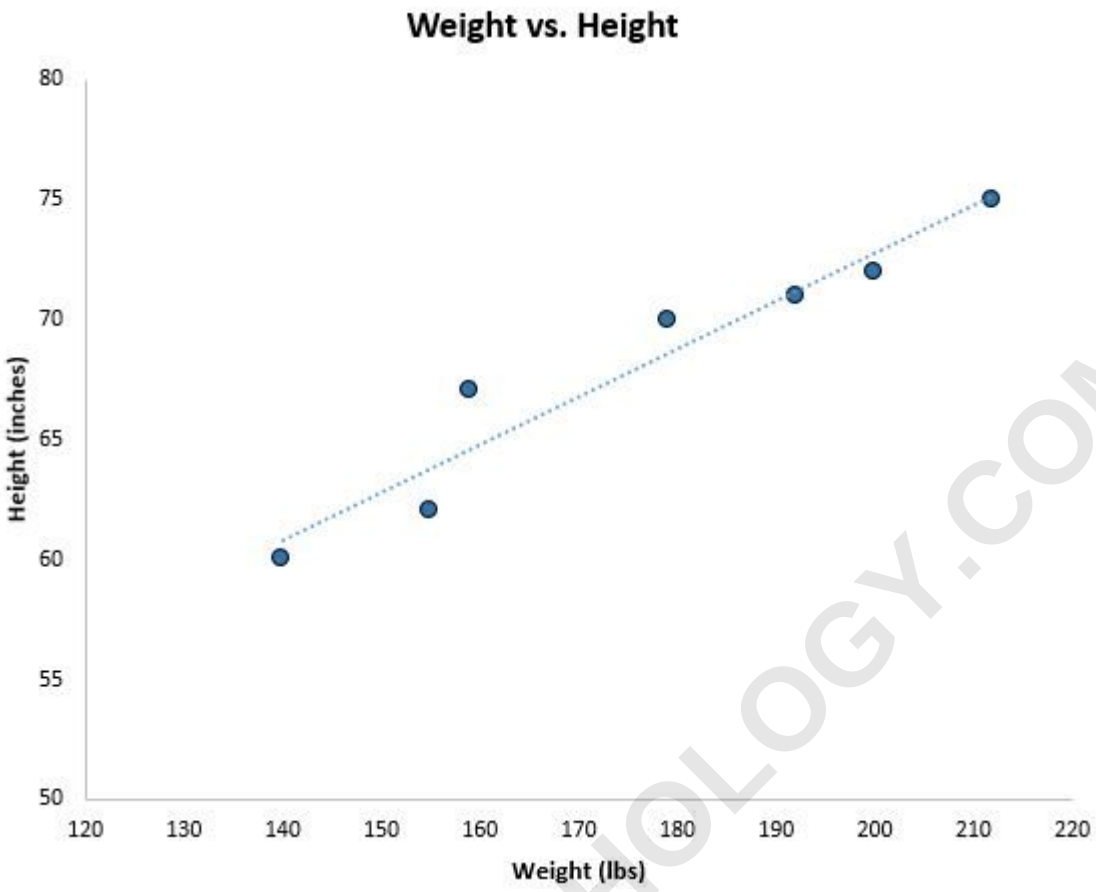
In this study, we designate *weight* as the **predictor variable** and *height* as the **response variable**.

The initial step in any regression study is often visualization. By plotting these paired observations on a scatterplot, with weight on the horizontal (x) axis and height on the vertical (y) axis, we can visually inspect the relationship.



The resulting scatterplot clearly demonstrates a positive correlation: as weight increases, height generally tends to increase as well. While the plot provides qualitative insight, to rigorously *quantify* the strength and nature of this linear relationship, we must apply linear regression techniques.

The primary goal of linear regression is to identify the line of best fit that minimizes the overall distance between the line and all data points. This line models the expected linear trend within the data, effectively summarizing the relationship:



Understanding the Regression Equation

The mathematical representation of the line of best fit derived from simple linear regression is formalized by the equation:

? = b0 + b1x

In this standard equation, ? (pronounced 'y-hat') signifies the **predicted value** of the response variable. The term **b?** is the y-intercept, and the term **b?** is the regression coefficient (or slope), which quantifies the expected change in y for every one-unit change in x. The variable **x** is the specific value of the predictor variable used for prediction.

For our specific dataset relating weight to height, the statistical analysis determined the following fitted equation for the regression line:

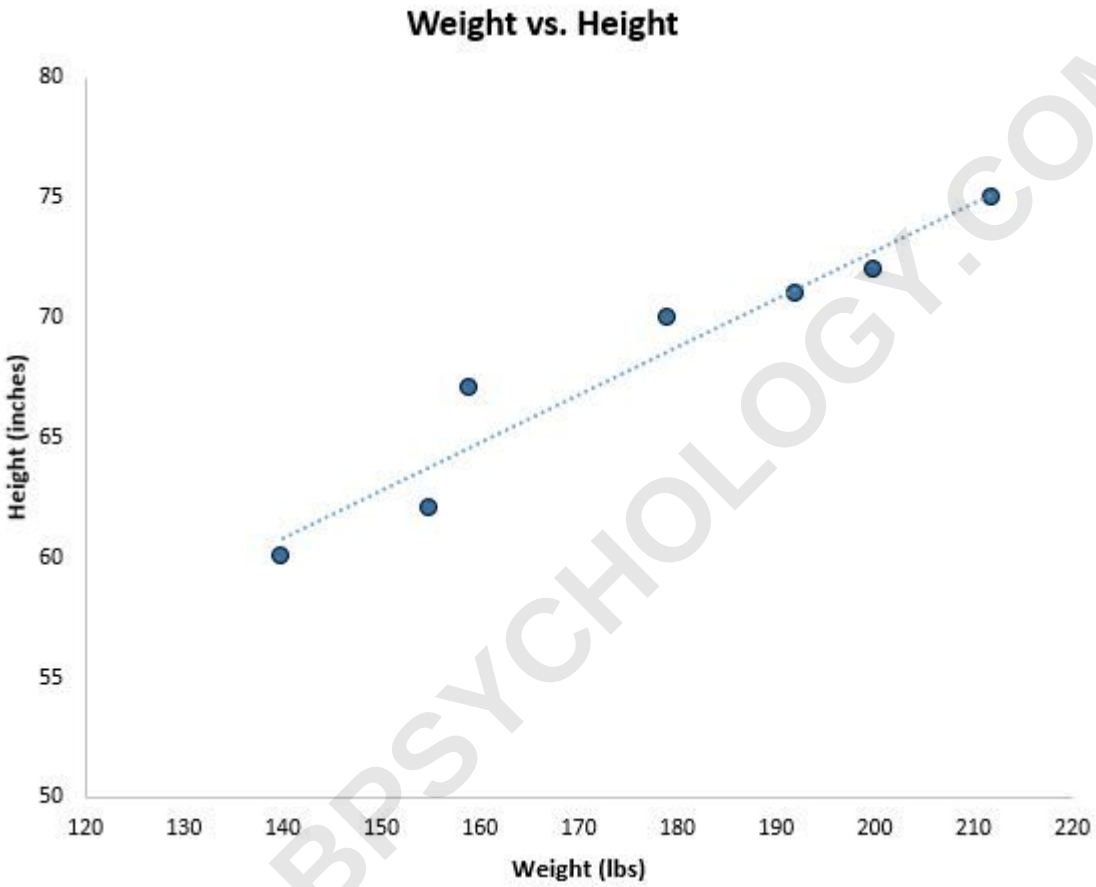
height = 32.783 + 0.2001*(weight)

This equation now serves as our core tool for generating predictions (?). Because real-world data rarely aligns perfectly with a straight line, the predicted value will often differ slightly from the actual

observed value (y). This difference is exactly what we define as the residual.

The Concept and Calculation of Residuals

As illustrated in the scatterplot below, the observed data points do not perfectly align with the line of best fit:



The vertical distance between an observed data point and the regression line is defined as the residual (e). Mathematically, the residual for any given observation is calculated using the simple formula:

e = y - ?

Where **y** is the actual observed value (the data point) and **?** is the predicted value (the point on the line). We must calculate a residual for every single data point in our analysis to accurately assess the model's performance across the entire range of the dependent variable.

Example 1: Calculating the Residual for the First Observation

To demonstrate the calculation, let us focus on the first observation in our dataset. We recall the original data table, which defines the observed values:

Weight (lbs)	Height (inches)
140	60
155	62
159	67
179	70
192	71
200	72
212	75

The first individual has an observed weight (x) of **140** lbs. and an observed height (y) of **60** inches. We use the established line of best fit equation to calculate the predicted height (?) by substituting the observed weight (x = 140) into the formula:

height = 32.783 + 0.2001*(weight)

Performing the calculation yields the predicted height for this individual:

height = 32.783 + 0.2001*(140)

height = 60.797 inches

The residual is the difference between the actual observed height (60 inches) and the predicted height (60.797 inches). Calculation: 60 - 60.797 = **-0.797**. A negative residual means the regression line predicted a value slightly higher than the actual observation, indicating the data point lies below the fitted line.

Example 2: Calculating the Residual for the Second Observation

The exact methodology is repeated to calculate the residual for the second individual in the dataset:

Weight (lbs)	Height (inches)
140	60
155	62
159	67
179	70
192	71
200	72
212	75

The second individual has an observed weight (x) of **155** lbs. and an observed height (y) of **62** inches. We substitute this weight into the line of best fit equation:

height = 32.783 + 0.2001*(weight)

The predicted height (?) for this individual is:

height = 32.783 + 0.2001*(155)

height = 63.7985 inches

To find the residual, we subtract the predicted height (63.7985 inches) from the actual observed height (62 inches). The resulting residual is 62 - 63.7985 = **-1.7985**. This larger negative value suggests that the model significantly overestimated this individual's height, resulting in a greater distance from the trend line.

The Property of Residuals in Ordinary Least Squares

By systematically applying the residual calculation (e = y - ?) to every data point, we generate a comprehensive view of the model's errors across the entire dataset:

Weight (lbs)	Height (inches)	Predicted Height	Residual
140	60	60.797	-0.797
155	62	63.7985	-1.7985
159	67	64.5989	2.4011
179	70	68.6009	1.3991
192	71	71.2022	-0.2022
200	72	72.803	-0.803
212	75	75.2042	-0.2042

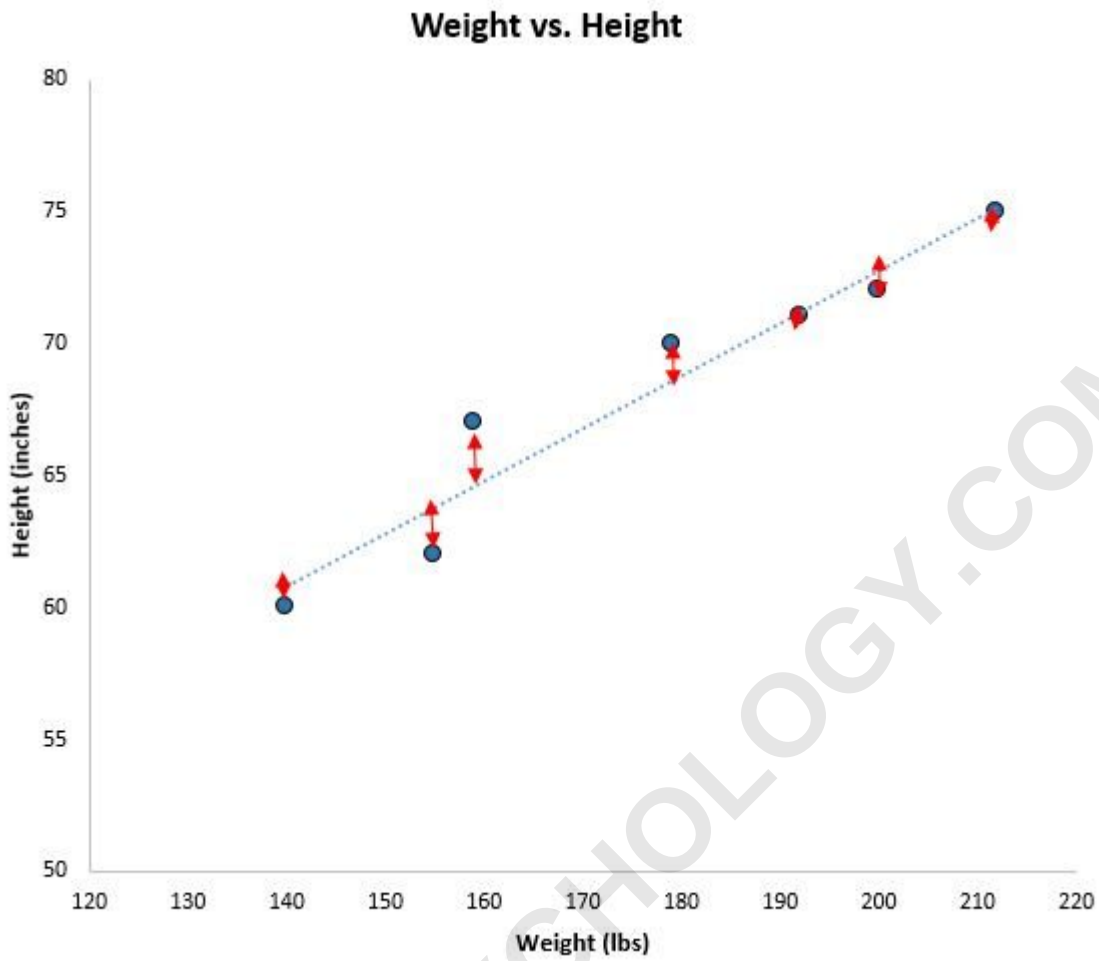
Note that the calculated residuals include both positive and negative values. A **positive residual** means the actual observation lies **above** the regression line, indicating the model underestimated the outcome. Conversely, a **negative residual** means the observation lies **below** the regression line, indicating the model overestimated the outcome.

An essential characteristic of the Ordinary Least Squares (OLS) method used in standard linear regression is that the sum of all the residuals will always equal zero. This occurs because OLS defines the regression line as the one that minimizes the total squared residuals, which perfectly balances the positive and negative errors across the dataset.

Visual Representation of Residuals

To gain a clearer, intuitive understanding, it is helpful to visualize the residuals directly on the scatterplot. Recall that the residual is the vertical distance separating an observed data point from the corresponding predicted point on the line of best fit.

The following illustration emphasizes these vertical distances (the residuals) as lines extending from each data point to the regression line. This visualization makes it clear how the magnitude and sign of the residual relate to the position of the data point relative to the model.



From this chart, we can see that the residuals vary significantly in size. Larger residuals correspond to data points that are poorly predicted by the model, while smaller residuals indicate a close fit. The visual representation confirms the existence of both positive (above the line) and negative (below the line) residuals, maintaining the zero-sum balance inherent in the OLS calculation.

Utilizing Residuals for Model Diagnostics: The Residual Plot

The fundamental purpose of calculating and analyzing residuals extends beyond simple error measurement; it serves as a critical diagnostic tool for evaluating the overall quality of the regression analysis model. By examining the patterns within these errors, statisticians can determine if the underlying assumptions of linear regression have been violated.

The magnitude of the residual is directly proportional to the model's predictive error for that point. **Larger residuals** suggest that the regression line is a poor fit for that specific data cluster, meaning the observed values are far from the predicted trend. Conversely, **smaller residuals** indicate an excellent fit, where the observed data points align closely with the regression line.

A specialized visualization tool used for this assessment is the residual plot. This plot graphs the predicted values (?) on the x-axis against the corresponding residual values (e) on the y-axis. Analyzing the pattern in a residual plot is crucial for assessing model appropriateness, particularly for checking for signs of non-linearity or heteroscedasticity (non-constant variance) of residuals.

For those interested in practical application, detailed tutorials are available on how to create a residual plot for a simple linear regression model using standard software like Excel.