

Learning Linear Regression: A Step-by-Step Guide to Deriving the Equation from Data

Authored by
Mohammed looti

November 16, 2025

RECOMMENDED CITATION

Mohammed looti (2025). *Learning Linear Regression: A Step-by-Step Guide to Deriving the Equation from Data*. PSYCHOLOGICAL STATISTICS. Retrieved from <https://statistics.arabpsychology.com/?p=2796>

In analytical disciplines ranging from scientific research to financial modeling, the ability to quantify the relationship between different factors is paramount for informed decision-making. One of the most essential statistical techniques employed for this purpose is [linear regression](#). This robust method allows researchers and analysts to derive a mathematical formula that accurately models the linear association between an [independent variable](#) (X) and a [dependent variable](#) (Y), based on a collected set of empirical [data points](#). Understanding the mechanics of this derivation is fundamental to statistical literacy.

The core objective of this process is to determine the precise [regression equation](#), conventionally expressed in the format $\hat{y} = b_0 + b_1x$. In this formula, $\hat{y}$ signifies the predicted outcome of the dependent variable, x is the input from the independent variable, b_0 represents the [y-intercept](#) (the predicted value of Y when X is zero), and b_1 denotes the [slope](#) of the fitted line. Once established, this equation serves as a powerful instrument for both [forecasting](#) and gaining a deep understanding of how changes in X quantitatively influence Y. While software automates this process today, a manual derivation provides unparalleled insight into the underlying mathematical principles.

The Foundational Principle of Least Squares

Fundamentally, [linear regression](#) functions as a statistical modeling technique designed to find the straight line that best summarizes the trend observed in a scatter plot of paired data. The model operates under the key assumption that the relationship between the explanatory variable (X) and the outcome variable (Y) can be approximated linearly. For instance, in economics, analysts might use this to predict consumer spending (Y) based on disposable income (X), or in public health, researchers might predict disease prevalence (Y) based on vaccination rates (X). The power of linear regression lies in its ability to condense complex data relationships into a simple, interpretable mathematical formula.

To ensure the resulting line is indeed the "best fit" possible, the methodology relies on the renowned [least squares](#) principle. This sophisticated statistical technique works by minimizing the sum of the squared vertical distances between every actual [data point](#) and the regression line itself. These vertical distances, representing the errors in prediction, are formally known as [residuals](#). By minimizing these squared errors, the least squares method guarantees that the calculated regression line is the statistically most representative linear model for the observed data distribution, providing the most accurate summary of the trend.

Although modern computational tools handle these calculations instantaneously, mastering the manual derivation of the slope and intercept is crucial for genuine statistical comprehension. The step-by-step calculations we are about to perform are the direct computational realization of the least squares formulas. This manual process offers profound, practical insight into precisely how

the coefficients (the slope and intercept) are derived and what they mathematically represent in the context of minimizing prediction error across the dataset.

Preparing and Organizing the Raw Data

To successfully derive a linear regression equation, the first necessary step is the meticulous organization and processing of our raw dataset. Our starting point is invariably a table of observations, where each row contains a paired value for the [independent variable](#) (X) and the corresponding [dependent variable](#) (Y). This structured, tabular format is the standard prerequisite for almost all quantitative statistical analyses.

We will utilize the following practical dataset, which consists of ten pairs of (X, Y) observations. Note that the total number of observations, denoted as **n**, is 10 for this specific analysis. Our ultimate objective is to systematically transform this initial collection of raw numbers into a robust, predictive mathematical model capable of describing the relationship between X and Y.

x	y
4	6
5	8
5	5
6	9
8	15
9	13
10	15
12	16
12	19
14	22

The manual calculation process requires careful augmentation of this basic table and precise aggregation of the resulting values. This ensures that every intermediate calculation accurately feeds into the complex formulas that define the parameters of the optimized regression line, guaranteeing computational accuracy throughout the derivation.

Step 1: Augmenting the Data Table for Calculation

The first imperative computational step in manually calculating the linear regression coefficients is

expanding our simple X-Y data table. The formulas used to determine the [slope](#) ($b?$) and the [y-intercept](#) ($b?$) are based on intermediate products and squares derived directly from the original X and Y [data points](#). Therefore, we must prepare these prerequisite values before moving to the aggregation stage.

We need to create three new columns, calculating three distinct metrics for every single observation in the dataset:

X * Y: The cross-product of the independent variable value (x) and the dependent variable value (y). This term captures the covariance component.

X²: The square of the independent variable (x) value. This is critical for calculating the variance of X.

Y²: The square of the dependent variable (y) value. Although not strictly needed for the slope and intercept, it is essential for calculating the correlation coefficient (R) and the coefficient of determination (R-squared).

These computations represent the essential groundwork, preparing the data for the final [summation](#) steps required by the core regression formulas. The augmented table below clearly demonstrates how these necessary calculations are applied to our dataset, providing a comprehensive basis for the subsequent statistical analysis:

x	y	x*y	x ²	y ²
4	6	24	16	36
5	8	40	25	64
5	5	25	25	25
6	9	54	36	81
8	15	120	64	225
9	13	117	81	169
10	15	150	100	225
12	16	192	144	256
12	19	228	144	361
14	22	308	196	484

By meticulously executing these preliminary squaring and multiplication steps, we ensure that the foundation for accurately determining the parameters of our linear model is robust and free from

computational errors.

Step 2: Aggregating Key Statistics via Summation

Once the table has been augmented with the products ($X * Y$) and squares (X^2 and Y^2), the second critical step involves calculating the sum of each relevant column. This [summation](#) process provides the essential aggregate statistics necessary for direct insertion into the formulas used to calculate the slope ($b?$) and intercept ($b?$) of the [linear regression](#) equation. The Greek letter Σ (Sigma) is the universal notation used in statistics to signify the sum of a series of values.

We are required to compute the following five crucial aggregate sums:

ΣX : The total sum of all independent variable (X) observations.

ΣY : The total sum of all dependent variable (Y) observations.

$\Sigma X * Y$: The total sum of all ($X * Y$) products, capturing the relationship between the variables.

ΣX^2 : The total sum of all (X^2) squared values.

ΣY^2 : The total sum of all (Y^2) squared values.

These five consolidated sums, alongside n (the total number of observations, which is 10 in our example), encapsulate all the numerical evidence needed to solve the regression equations derived from the [least squares](#) principle. The following visualization confirms the calculation of these totals, found at the base of their corresponding columns, which will serve as the inputs for the final coefficient formulas:

	x	y	x*y	x²	y²
	4	6	24	16	36
	5	8	40	25	64
	5	5	25	25	25
	6	9	54	36	81
	8	15	120	64	225
	9	13	117	81	169
	10	15	150	100	225
	12	16	192	144	256
	12	19	228	144	361
	14	22	308	196	484
Σ	85	128	1258	831	1926

With these comprehensive aggregate statistics now prepared--specifically $\Sigma X = 85$, $\Sigma Y = 128$, $\Sigma XY = 1258$, and $\Sigma X^2 = 831$ --we can proceed directly to the final stages of calculating the specific coefficients that rigorously define our predictive model.

Step 3: Calculation of the Y-Intercept (b_0)

The [y-intercept](#), represented by the coefficient b_0 , is a defining element of the linear regression model. Conceptually, b_0 estimates the expected value of the [dependent variable](#) (Y) precisely at the point where the independent variable (X) is zero. Graphically, this is the crucial point where the regression line intersects the vertical axis. It serves as the baseline value for the prediction.

The formula utilized for calculating b_0 is derived directly from the constraints of the [least squares](#) minimization process, leveraging the aggregate sums determined in Step 2. The formula is structured to account for the overall scaling and positioning of the data:

$$b_0 = ((\Sigma Y)(\Sigma X^2) - (\Sigma X)(\Sigma XY)) / (n(\Sigma X^2) - (\Sigma X)^2)$$

By substituting our calculated values-- $n=10$, $\Sigma X=85$, $\Sigma Y=128$, $\Sigma XY=1258$, and $\Sigma X^2=831$ --into the formula, we perform the necessary arithmetic to solve for the intercept. Note that n represents the total count of paired observations in the dataset, which is 10:

$$b_0 = ((128)(831) - (85)(1258)) / (10(831) - (85)^2)$$

$$b_0 = (106368 - 106930) / (8310 - 7225)$$

$$b_0 = -562 / 1085$$

$$b_0 \approx -0.518$$

This result, $b_0 \approx -0.518$, provides the starting point for our linear prediction model, establishing the intercept of the best-fit line.

Step 4: Determining the Slope (b1) and Final Equation

The second fundamental coefficient required is the [slope](#), denoted as b_1 . This coefficient is perhaps the most important, as it dictates the steepness and direction of the regression line, quantifying the expected change in the dependent variable (Y) for every single unit increase in the independent variable (X). A positive slope signifies a direct relationship, meaning Y increases as X increases, while a negative slope indicates an inverse relationship.

The formula for calculating b_1 is also derived from the mathematical constraints of the [least squares](#) technique. Crucially, the denominator for the slope calculation is identical to the denominator used for the intercept calculation, representing the variance component of X:

$$b_1 = (n(\sum XY) - (\sum X)(\sum Y)) / (n(\sum X^2) - (\sum X)^2)$$

Using the identical sums and 'n' value utilized for the intercept calculation, we determine the slope coefficient:

$$b_1 = (10(1258) - (85)(128)) / (10(831) - (85)^2)$$

$$b_1 = (12580 - 10880) / (8310 - 7225)$$

$$b_1 = 1700 / 1085$$

$$b_1 \approx 1.5668$$

With both the intercept ($b_0 \approx -0.518$) and the slope ($b_1 \approx 1.5668$) successfully calculated, we now possess all the required elements to formulate the complete [regression equation](#). We substitute these specific values into the general form $Y = b_0 + b_1x$. Our final, specific linear regression equation is:

$$Y = -0.518 + 1.5668x$$

This equation is now the definitive tool derived from the data, ready to be used for making informed [predictions](#). For example, predicting Y when X equals 10 yields: $Y = -0.518 + 1.5668(10) = 15.15$.

Interpretation and Model Verification

The derived [regression equation](#), $Y = -0.518 + 1.5668x$, offers profound quantitative insights into

the modeled relationship. Understanding the context of each coefficient is essential for accurate analysis. The [y-intercept](#) ($b_0 = -0.518$) suggests that if the independent variable X were zero, the predicted value for Y would be -0.518 . However, analysts must always exercise caution: if $X=0$ is far outside the practical or observed range of the original data, the intercept may simply be a mathematical necessity rather than a value with genuine real-world significance.

More critically, the [slope](#) ($b_1 = 1.5668$) indicates a strong, positive, linear relationship between X and Y . Specifically, the model predicts that for every incremental unit increase in X , Y is expected to increase by approximately 1.57 units. This quantitative measure of association is the primary goal of applying the [least squares](#) method, establishing how closely and positively the two variables move together based on the observed data.

While this equation represents the mathematically best linear fit, statistical rigor dictates that manual calculations should always be cross-verified using reliable statistical software or a dedicated computational tool. This verification step is crucial for building confidence in the accuracy of the hand-calculated coefficients. To confirm our derived equation, we input the exact set of original [data points](#) into a professional statistical application. The output generated by the software should align precisely, or very closely due to controlled rounding, with our calculated intercept and slope values.

The image below displays the results obtained from a standard statistical application processing the same input dataset:

Predictor values:

4, 5, 5, 6, 8, 9, 10, 12, 12, 14

Response values:

6, 8, 5, 9, 15, 13, 15, 16, 19, 22

CALCULATE

Linear Regression Equation:

$$\hat{y} = -0.5180 + (1.5668) \cdot x$$

As confirmed by the screenshot, the linear regression equation generated by the software matches our meticulous manual calculations, confirming the validity and accuracy of our step-by-step derivation process. A complete analysis would also require assessing the model's goodness-of-fit (R-squared) and checking for influential outliers to ensure the model is reliable for generalized inference.

For those seeking to deepen their knowledge of [linear regression](#), correlation analysis, and other essential statistical methodologies, the following resources are recommended for further reading and advanced exploration: