

Learning Multiple Linear Regression: A Step-by-Step Guide

Authored by
Mohammed looti

November 6, 2025

RECOMMENDED CITATION

Mohammed looti (2025). *Learning Multiple Linear Regression: A Step-by-Step Guide*. PSYCHOLOGICAL STATISTICS. Retrieved from <https://statistics.arabpsychology.com/?p=11749>

[Multiple linear regression](#) is a cornerstone statistical technique used across various disciplines--from economics to engineering--to model and quantify the complex relationship between multiple inputs and a single output. This robust method enables researchers to assess how two or more [predictor variables](#) collectively influence a single [response variable](#).

While sophisticated statistical software packages efficiently automate these demanding numerical operations, achieving a deep, intuitive understanding of the model requires knowing the underlying mechanics. This comprehensive guide breaks down the manual, step-by-step process for calculating the parameters of a multiple linear regression model, providing invaluable insight into how the model estimates its **coefficients**.

Introduction and Theoretical Foundation

Unlike simple linear regression, which restricts the analysis to just one predictor, multiple linear regression extends this framework by allowing for the simultaneous evaluation of several independent variables ($X_1, X_2, \dots, X_k$). This capability is crucial when analyzing real-world systems where outcomes are rarely determined by a single factor, but rather by the interplay of several factors working in tandem. The model seeks to fit a **hyperplane** (the regression plane) through the data points in k -dimensional space.

The fundamental objective of this calculation is to identify the optimal set of [regression coefficients](#) (b_0, b_1, b_2 , etc.) that adheres to the principle of the [least squares method](#). This method ensures that the sum of the squared residuals--the vertical distances between the observed data points and the estimated regression plane--is **minimized**, thus generating the mathematically "best-fit" linear model for the observed data structure.

Setting Up the Manual Calculation Example

To clearly demonstrate the manual computational steps, we will employ a small, manageable dataset. Our example dataset consists of eight observations ($n=8$). We have one response variable, Y , and two [predictor variables](#), X_1 and X_2 . This restricted number of variables makes the calculation feasible without the aid of computational tools, allowing us to focus purely on the algebraic procedure.

The structure of our raw observational data, including the values for the response variable and the two predictors, is presented below. The initial step requires calculating the sum (Σ) of each column, which forms the basis for all subsequent variance and covariance calculations.

y	X ₁	X ₂
140	60	22
155	62	25
159	67	24
179	70	20
192	71	15
200	72	14
212	75	14
215	78	11

Step 1: Calculating Raw Sums of Squares and Products

The initial essential step toward solving the system of simultaneous equations involves transforming the raw data into a set of necessary intermediate sums. This requires expanding the original dataset significantly. We must calculate the squares of the predictor variables (X_1^2 , X_2^2), the products of each predictor with the response variable (X_1Y , X_2Y), and the **interaction product** of the two predictors (X_1X_2).

These extended variables allow us to compute the raw totals (Σ) for every column. These raw sums are the foundational components for calculating the covariance and variance terms necessary for correcting the data in the next step. Ensuring **absolute accuracy** in this raw data transformation is paramount, as mistakes here will propagate through the entire regression solution.

The augmented table, which includes all calculated products and squares, along with the totals (Σ) for each respective column, is shown here for reference:

	y	X ₁	X ₂		X ₁ ²	X ₂ ²	X ₁ y	X ₂ y	X ₁ X ₂
	140	60	22		3600	484	8400	3080	1320
	155	62	25		3844	625	9610	3875	1550
	159	67	24		4489	576	10653	3816	1608
	179	70	20		4900	400	12530	3580	1400
	192	71	15		5041	225	13632	2880	1065
	200	72	14		5184	196	14400	2800	1008
	212	75	14		5625	196	15900	2968	1050
	215	78	11		6084	121	16770	2365	858
Mean	181.5	69.375	18.125	Sum	38767	2823	101895	25364	9859
Sum	1452	555	145						

Step 2: Correcting the Sums of Squares (SS) and Products (SP)

To accurately estimate the partial slopes, which measure the effect of one variable independent of the others, we must calculate the corrected [Sum of Squares](#) (SS) and Sum of Products (SP). These corrected sums effectively "center" the data by measuring the variation and covariation around the respective means, thereby removing the influence of the overall magnitude of the data points. In regression notation, these corrected sums are typically represented by lowercase variables (e.g., Σx_{12}).

The general formula for correcting a raw sum is: $\text{Corrected Sum} = \Sigma X^2 - (\Sigma X)^2 / n$. By subtracting the correction factor (which accounts for the mean), we isolate the true **variability** within the dataset. We require five crucial corrected sums derived from the totals in Step 1 to solve the coefficient formulas.

The calculations below show how the corrected sums are derived. These specific numerical results will serve as the inputs for the complex formulas used to determine the partial [regression coefficients](#) (b_1 and b_2).

$$\Sigma x_{12} = \Sigma X_{12} - (\Sigma X_1)^2 / n = 38,767 - (555)^2 / 8 = \mathbf{263.875}$$

$$\Sigma x_{22} = \Sigma X_{22} - (\Sigma X_2)^2 / n = 2,823 - (145)^2 / 8 = \mathbf{194.875}$$

$$\Sigma x_{1y} = \Sigma X_{1y} - (\Sigma X_1 \Sigma y) / n = 101,895 - (555 * 1,452) / 8 = \mathbf{1,162.5}$$

$$\Sigma x_{2y} = \Sigma X_{2y} - (\Sigma X_2 \Sigma y) / n = 25,364 - (145 * 1,452) / 8 = \mathbf{-953.5}$$

$$\Sigma x_{1x_2} = \Sigma X_{1X_2} - (\Sigma X_1 \Sigma X_2) / n = 9,859 - (555 * 145) / 8 = \mathbf{-200.375}$$

The resulting table summarizing these five critical corrected sums is provided below, serving as the definitive input matrix for the subsequent coefficient calculations:

	y	X ₁	X ₂		X ₁ ²	X ₂ ²	X ₁ y	X ₂ y	X ₁ X ₂
	140	60	22		3600	484	8400	3080	1320
	155	62	25		3844	625	9610	3875	1550
	159	67	24		4489	576	10653	3816	1608
	179	70	20		4900	400	12530	3580	1400
	192	71	15		5041	225	13632	2880	1065
	200	72	14		5184	196	14400	2800	1008
	212	75	14		5625	196	15900	2968	1050
	215	78	11		6084	121	16770	2365	858
Mean	181.5	69.375	18.125	Sum	38767	2823	101895	25364	9859
Sum	1452	555	145						
Reg Sums					263.875	194.875	1162.5	-953.5	-200.375

Steps 3 & 4: Solving for the Partial Slopes and Intercept

The partial slopes, b_1 and b_2 , are the core parameters of the model. They quantify the expected change in the [response variable](#) (Y) associated with a one-unit change in that specific predictor, while rigorously holding the effects of all other predictors constant (a statistical concept known as **ceteris paribus**). These values are typically found by solving a system of normal equations, which, when performed manually, translates into complex formulas derived from [matrix algebra](#).

We first calculate the coefficient for X_1 (b_1). This formula utilizes a ratio of specific corrected sums, ensuring the partial effect is isolated:

$$b_1 = \frac{S_{YX_1} - \frac{S_{YX_2} S_{X_1X_2}}{S_{X_2X_2}}}{S_{X_1X_1} - \frac{S_{X_1X_2}^2}{S_{X_2X_2}}}$$

Substituting our calculated values from Step 2 into the equation yields the numerical result: **$b_1 = 3.148$**

Next, we determine the coefficient for X_2 (b_2). The denominator remains the same for all slopes, as it represents the overall variance structure of the predictors:

$$b_2 = \frac{S_{YX_2} - \frac{S_{YX_1} S_{X_1X_2}}{S_{X_1X_1}}}{S_{X_2X_2} - \frac{S_{X_1X_2}^2}{S_{X_1X_1}}}$$

Substituting our derived values: **$b_2 = -1.656$**

Finally, we calculate the intercept (b_0). The intercept ensures that the regression plane passes precisely through the centroid (the means) of the dataset ($\bar{X}_1$, $\bar{X}_2$, $\bar{Y}$). This process is vital as it guarantees that the sum of the residuals across the entire model equals zero.

The general formula for b_0 is: $b_0 = \bar{Y} - b_1\bar{X}_1 - b_2\bar{X}_2$. Given the means $\bar{Y} = 181.5$, $\bar{X}_1 = 69.375$, and $\bar{X}_2 = 18.125$, we compute the final intercept value: **$b_0 = 181.5 - 3.148(69.375) - (-1.656)(18.125) = -6.867$**

Step 5: Constructing and Interpreting the Final Model Equation

With all three [regression coefficients](#) (b_0 , b_1 , and b_2) successfully calculated, we can now construct the complete [estimated linear regression equation](#). This equation provides the mathematical framework for predicting the response variable based on new values of the predictors.

The general form of the multiple linear regression model is expressed as: $\hat{Y} = b_0 + b_1x_1 + b_2x_2$. Substituting our derived values into this structure yields the final estimated equation for this specific dataset: **$\hat{Y} = -6.867 + 3.148x_1 - 1.656x_2$** .

Interpreting the individual components of this model is critical for drawing meaningful conclusions

from the statistical analysis:

The Intercept ($b_0 = -6.867$): This value estimates the mean value for Y when both predictor variables, X1 and X2, are set to zero. If zero falls outside the observed data range, the intercept may represent an abstract mathematical starting point rather than a physically meaningful prediction.

The Partial Slope for X1 ($b_1 = 3.148$): This **positive coefficient** signifies that, holding X2 constant, a one-unit increase in X1 is associated with an average increase of 3.148 units in Y.

The Partial Slope for X2 ($b_2 = -1.656$): This **negative coefficient** indicates an inverse relationship, suggesting that, controlling for X1, a one-unit increase in X2 is associated with an average decrease of 1.656 units in Y.

Further Statistical Resources

To deepen your understanding of regression techniques, statistical methodology, and advanced modeling concepts, the following resources are recommended:

[An Introduction to Multiple Linear Regression](#)

[How to Perform Simple Linear Regression by Hand](#)