

Perform Linear Regression in Google Sheets

Authored by
Mohammed looti

November 7, 2025

RECOMMENDED CITATION

Mohammed looti (2025). *Perform Linear Regression in Google Sheets*. PSYCHOLOGICAL STATISTICS. Retrieved from <https://statistics.arabpsychology.com/?p=12076>

Linear regression is a cornerstone of statistical analysis, employed universally to model and quantify the linear relationship existing between variables. Fundamentally, this technique helps analysts determine precisely how changes in one or more **explanatory variables** (predictors) influence a single **response variable** (the outcome).

The specific form of regression utilized depends entirely on the complexity of the model and the number of predictors involved. We rely on **simple linear regression** when our analysis focuses on the impact of only a single explanatory variable. Conversely, **multiple linear regression** is the appropriate method when the outcome is assumed to be influenced by two or more independent explanatory variables.

Fortunately, executing both simple and multiple linear regression analyses is remarkably straightforward within Google Sheets, thanks to the powerful, integrated **LINEST() function**. This array function is specifically designed to calculate the statistics necessary for a straight line that best fits the data, utilizing the **least squares method**. Understanding its structure is the first step toward advanced data analysis in the spreadsheet environment.

The general syntax for the **LINEST() function** is defined as follows:

LINEST(known_data_y, known_data_x, calculate_b, verbose)

Each argument is crucial for accurately defining the scope and desired output of the regression model:

known_data_y: This is the required array or range containing all known values of the **response variable** (the Y values). This is the variable we are attempting to predict.

known_data_x: This is the required array or range containing the known values of the **explanatory variables** (the X values). For performing multiple regression, this range must include multiple adjacent columns corresponding to each predictor.

calculate_b: A Boolean value (TRUE or FALSE) that dictates whether the regression equation should calculate the **y-intercept** (the constant 'b'). The default setting is **TRUE**, which is standard for most linear models.

verbose: A Boolean value (TRUE or FALSE) that controls the amount of detail returned by the function. Setting this to **TRUE** instructs the function to return a full suite of detailed regression statistics, such as R-squared, standard errors, and F-statistics, beyond just the core **slope** and intercept. While the default is FALSE, setting it to TRUE is essential for comprehensive model diagnosis, as demonstrated in the examples below.

The following practical demonstrations illustrate the effective deployment of the **LINEST() function**, moving from a simple model to a more complex multivariate analysis.

Executing Simple Linear Regression in Google Sheets

To begin our practical application, let us analyze a common educational scenario: determining the correlation between the amount of time a student dedicates to studying and their final exam score. This scenario is ideal for **simple linear regression** because it involves only a single predictor variable.

In setting up this specific model, we formally designate the *hours studied* as our single explanatory variable (X) and the resulting *exam score* as the response variable (Y). Our primary objective is to derive a statistically estimated linear equation that allows us to predict a student's score based solely on their study time.

The screenshot below showcases the application of simple linear regression to a hypothetical dataset covering 20 students. Note the formula entered in cell D2; it is configured to extract the full range of comprehensive regression statistics by setting the final argument to TRUE:

=LINEST(B2:B21, A2:A21, TRUE, TRUE)

	A	B	C	D	E
1	hours	score			
2	1	76		5.250257467	67.16168898
3	2	78		0.7578474311	2.663270396
4	2	85		0.7272528139	5.280516451
5	4	88		47.99518133	18
6	2	72		1338.290628	501.9093718
7	1	69			
8	5	94			
9	4	94			
10	2	88			
11	4	92			
12	4	90			
13	3	75			
14	6	96			
15	5	90			
16	3	82			
17	4	85			
18	6	99			
19	2	83			
20	1	62			
21	2	76			
22					
23					
24					
25					

Because [LINEST\(\)](#) is an array function, the result automatically spills across a range of cells, producing a matrix of various statistical measures. These metrics are vital for assessing the model's overall fit and interpreting the relationship found in the data. The following annotated screenshot clarifies the meaning and position of the key elements within this statistical output:

	A	B	C	D	E
1	hours	score		slope	intercept
2	1	76	coefficient for slope, coefficient for intercept	5.250257467	67.16168898
3	2	78	Standard error for slope, standard error for intercept	0.7578474311	2.663270396
4	2	85	coefficient of determination, standard error	0.7272528139	5.280516451
5	4	88	F stat w/ degrees of freedom	47.99518133	18
6	2	72	Regression SS, Residual SS	1338.290628	501.9093718
7	1	69			
8	5	94			
9	4	94			
10	2	88			
11	4	92			
12	4	90			
13	3	75			
14	6	96			
15	5	90			
16	3	82			
17	4	85			
18	6	99			
19	2	83			
20	1	62			
21	2	76			
22					
23					

Interpreting the Results of Simple Regression Analysis

Drawing meaningful conclusions from any linear regression analysis requires a clear understanding of the key metrics returned by the LINEST output array. Here, we focus on interpreting the most relevant statistics derived from the simple regression model above:

R Square (Coefficient of Determination): The computed value is **0.72725**. This critical metric, officially known as the [coefficient of determination](#), quantifies the proportion of the total variance in the response variable (Exam Score) that can be reliably predicted or explained by the explanatory variable (Hours Studied). In this specific example, approximately 72.73% of the observed differences in student exam scores are accounted for by the differences in the number of hours they studied.

Coefficients (Slope and Intercept): These values, located in the first row of the output, provide the essential components--the slope and the y-intercept--required to construct the estimated regression equation. Based on the data analysis, the model is formulated as:

$$\text{Exam score} = 67.16 + 5.2503 * (\text{hours})$$

The coefficient for hours (5.2503) is interpreted as the estimated change in the exam score for every one-unit increase in the predictor. Specifically, for each additional hour studied, the exam score is expected to increase by **5.2503** points, holding all other factors constant. The intercept (67.16) represents the predicted exam score for a hypothetical student who dedicates zero hours to studying (where the line crosses the Y-axis).

This estimated equation is a powerful tool for making predictions. For instance, using this model, a student who studies for exactly three hours is expected to achieve a predicted score of **82.91**, calculated as follows:

$$\text{Exam score} = 67.16 + 5.2503 * (3) = 82.91$$

Implementing Multiple Linear Regression in Google Sheets

While simple regression is informative, real-world outcomes are typically influenced by complex interactions among several variables. To reflect this complexity, we expand our model to perform [multiple linear regression](#). We will now investigate whether the number of hours spent studying *and* the number of preparatory exams taken both contribute significantly to the student's final exam score.

In this expanded framework, we utilize *hours studied* and *prep exams taken* as our two distinct explanatory variables, while *exam score* remains the single response variable. This model allows us to assess the independent contribution of each predictor.

The methodology within Google Sheets remains largely consistent with the simple regression approach, but a critical adjustment is made to the second argument of the [LINEST\(\) function](#). The range specifying the known X values must now encompass both columns of predictors (Hours Studied and Prep Exams). The following screenshot illustrates the data setup and the modified formula entered in cell E2:

=LINEST(C2:C21, A2:B21, TRUE, TRUE)

fx							
	A	B	C	D	E	F	G
1	hours	prep_exams	score				
2	1	1	76		-0.6016868046	5.555748295	67.67352554
3	2	3	78		0.9143850312	0.8991916992	2.815802228
4	2	3	85		0.734027217	5.365703261	#N/A
5	4	5	88		23.45815717	17	#N/A
6	2	2	72		1350.756885	489.4431152	#N/A
7	1	2	69				
8	5	1	94				
9	4	1	94				
10	2	0	88				
11	4	3	92				
12	4	4	90				
13	3	3	75				
14	6	2	96				
15	5	4	90				
16	3	4	82				
17	4	4	85				
18	6	5	99				
19	2	1	83				
20	1	0	62				
21	2	1	76				
22							
23							
24							
25							

Analyzing the Output of Multiple Regression

The output array generated by the multiple regression analysis offers a significantly richer set of statistical information. This allows for a detailed evaluation of the combined and individual influence exerted by the two explanatory factors (study hours and prep exams) on the student scores.

R Square: The calculated value is **0.734**. This indicates that 73.4% of the total variability observed in the exam scores is collectively explained and accounted for by the combination of the number of hours studied and the number of prep exams taken. While only slightly higher than the simple regression model, the increase suggests that the second variable adds minor explanatory power.

Standard error of the Estimate: This metric is calculated as **5.3657**. The standard error measures the average magnitude of the residual (the difference between the observed data point and the predicted point on the regression plane). In practical terms, this means that the actual student scores typically deviate from the scores predicted by our model by an average of 5.3657 units.

Estimated Regression Equation: The [coefficients](#) from the output model are presented in the same order as the explanatory variables were listed in the X range, followed by the intercept. This configuration allows us to construct the final prediction equation:

$$\text{Exam score} = 67.67 + 5.56 * (\text{hours}) - 0.60 * (\text{prep exams})$$

The coefficients reveal interesting insights: every additional hour studied increases the predicted score by 5.56 points, while surprisingly, every prep exam taken decreases the predicted score by 0.60 points (though the statistical significance of this negative coefficient would need further testing). We can use this equation to predict the expected score for a student who studies for three hours and takes one prep exam, resulting in a score of **83.75**:

$$\text{Exam score} = 67.67 + 5.56 * (3) - 0.60 * (1) = 83.75$$

Next Steps for Robust Data Analysis

Mastering the [LINEST\(\) function](#) provides a solid foundation for performing linear regression directly in Google Sheets. However, a complete regression analysis requires more than just calculating coefficients and R-squared values. It is essential to perform model diagnostics to ensure the fundamental assumptions of linear regression are met.

Advanced statistical visualization, such as analyzing residual plots, is crucial for detecting non-linearity, heteroscedasticity, and outliers within your dataset. The following tutorial provides guidance on this essential diagnostic step:

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