

LINEST

 Hide All

Calculates the statistics for a line by using the "least squares" method to calculate a straight line that best fits your data, and then returns an array that describes the line. You can also combine LINEST with other functions to calculate the statistics for other types of models that are linear in the unknown parameters, including polynomial, logarithmic, exponential, and power series. Because this function returns an array of values, it must be entered as an array formula.

The equation for the line is:

$$y = mx + b \text{ or}$$

$$y = m_1x_1 + m_2x_2 + \dots + b \text{ (if there are multiple ranges of x-values)}$$

where the dependent y-value is a function of the independent x-values. The m-values are coefficients corresponding to each x-value, and b is a constant value. Note that y, x, and m can be vectors. The array that LINEST returns is {m_n,m_{n-1},...,m₁,b}. LINEST can also return additional regression statistics.

Syntax

LINEST(known_y's,known_x's,const,stats)

Known_y's is the set of y-values you already know in the relationship $y = mx + b$.

- If the array known_y's is in a single column, then each column of known_x's is interpreted as a separate variable.
- If the array known_y's is in a single row, then each row of known_x's is interpreted as a separate variable.

Known_x's is an optional set of x-values that you may already know in the relationship $y = mx + b$.

- The array known_x's can include one or more sets of variables. If only one variable is used, known_y's and known_x's can be ranges of any shape, as long as they have equal dimensions. If more than one variable is used, known_y's must be a vector (that is, a range with a height of one row or a width of one column).
- If known_x's is omitted, it is assumed to be the array {1,2,3,...} that is the same size as known_y's.

Const is a logical value specifying whether to force the constant b to equal 0.

- If const is TRUE or omitted, b is calculated normally.
- If const is FALSE, b is set equal to 0 and the m-values are adjusted to fit $y = mx$.

Stats is a logical value specifying whether to return additional regression statistics.

- If stats is TRUE, LINEST returns the additional regression statistics, so the returned array is {mn,mn-1,...,m1,b;sen,sen-1,...,se1,seb;r2,sey;F,df;ssreg,ssresid}.
- If stats is FALSE or omitted, LINEST returns only the m-coefficients and the constant b.

The additional regression statistics are as follows.

Statistic	Description
se1,se2,...,sen	The standard error values for the coefficients m1,m2,...,mn.
seb	The standard error value for the constant b (seb = #N/A when const is FALSE).
r2	The coefficient of determination. Compares estimated and actual y-values, and ranges in value from 0 to 1. If it is 1, there is a perfect correlation in the sample — there is no difference between the estimated y-value and the actual y-value. At the other extreme, if the coefficient of determination is 0, the regression equation is not helpful in predicting a y-value. For information about how r2 is calculated, see "Remarks" later in this topic.
sey	The standard error for the y estimate.
F	The F statistic, or the F-observed value. Use the F statistic to determine whether the observed relationship between the dependent and independent variables occurs by chance.
df	The degrees of freedom. Use the degrees of freedom to help you find F-critical values in a statistical table. Compare the values you find in the table to the F statistic returned by LINEST to determine a confidence level for the model. For information about how df is calculated, see "Remarks" later in this topic. Example 4 below shows use of F and df.
ssreg	The regression sum of squares.
ssresid	The residual sum of squares. For information about how ssreg and ssresid are calculated, see "Remarks" later in this topic.

The following illustration shows the order in which the additional regression statistics are returned.

	A	B	C	D	E	F
1	m _n	m _{n-1}	...	m ₂	m ₁	b
2	se _n	se _{n-1}	...	se ₂	se ₁	seb
3	r ₂	se _y				
4	F	df				
5	ssreg	ssresid				

Remarks

- You can describe any straight line with the slope and the y-intercept:

Slope (m):

To find the slope of a line, often written as m, take two points on the line, (x1,y1) and (x2,y2); the slope is equal to $(y2 - y1)/(x2 - x1)$.

Y-intercept (b):

The y-intercept of a line, often written as b, is the value of y at the point where the line crosses the y-axis.

The equation of a straight line is $y = mx + b$. Once you know the values of m and b , you can calculate any point on the line by plugging the y - or x -value into that equation. You can also use the TREND function.

- When you have only one independent x -variable, you can obtain the slope and y -intercept values directly by using the following formulas:

Slope:

`=INDEX(LINEST(known_y's,known_x's),1)`

Y -intercept:

`=INDEX(LINEST(known_y's,known_x's),2)`

- The accuracy of the line calculated by LINEST depends on the degree of scatter in your data. The more linear the data, the more accurate the LINEST model. LINEST uses the method of least squares for determining the best fit for the data. When you have only one independent x -variable, the calculations for m and b are based on the following formulas:

$$m = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2}$$

$$b = \bar{y} - m\bar{x}$$

where \bar{x} and \bar{y} are sample means, i.e., $\bar{x} = \text{AVERAGE}(\text{known } x\text{'s})$ and $\bar{y} = \text{AVERAGE}(\text{known } y\text{'s})$.

- The line- and curve-fitting functions LINEST and LOGEST can calculate the best straight line or exponential curve that fits your data. However, you have to decide which of the two results best fits your data. You can calculate TREND(known_y's,known_x's) for a straight line, or GROWTH(known_y's, known_x's) for an exponential curve. These functions, without the new_x's argument, return an array of y -values predicted along that line or curve at your actual data points. You can then compare the predicted values with the actual values. You may want to chart them both for a visual comparison.
- In regression analysis, Microsoft Excel calculates for each point the squared difference between the y -value estimated for that point and its actual y -value. The sum of these squared differences is called the residual sum of squares, ssresid. Microsoft Excel then calculates the total sum of squares, sstotal. When const = TRUE, or omitted, the total sum of squares is the sum of the squared differences between the actual y -values and the average of the y -values. When const = FALSE, the total sum of squares is the sum of the squares of the actual y -values (without subtracting the average y -value from each individual y -value). Then regression sum of squares, ssreg, can be found from: $\text{ssreg} = \text{sstotal} - \text{ssresid}$. The smaller the residual sum of squares is, compared with the total sum of squares, the larger the value of the coefficient of determination, r^2 , which is an indicator of how well the equation resulting from the regression analysis explains the relationship among the variables. r^2 equals $\text{ssreg}/\text{sstotal}$.
- In some cases, one or more of the X columns (assume that Y 's and X 's are in columns) may have no additional predictive value in the presence of the other X columns. In other words, eliminating one or more X columns might lead to predicted Y values that are equally accurate. In that case these redundant X columns should be omitted from the regression model. This phenomenon is called "collinearity" because any redundant X column can be expressed as a sum of multiples of the non-redundant X columns. LINEST checks for collinearity and removes any

redundant X columns from the regression model when it identifies them. Removed X columns can be recognized in LINEST output as having 0 coefficients as well as 0 se's. If one or more columns are removed as redundant, then df is affected because df depends on the number of X columns actually used for predictive purposes. For details on the computation of df, see Example 4 below. If df is changed because redundant X columns are removed, values of sey and F are also affected. Collinearity should be relatively rare in practice. However, one case where it is more likely to arise is when some X columns contain only 0's and 1's as indicators of whether a subject in an experiment is or is not a member of a particular group. If const = TRUE or omitted, LINEST effectively inserts an additional X column of all 1's to model the intercept. If you have a column with a 1 for each subject if male, or 0 if not, and you also have a column with a 1 for each subject if female, or 0 if not, this latter column is redundant because entries in it can be obtained from subtracting the entry in the "male indicator" column from the entry in the additional column of all 1's added by LINEST.

- df is calculated as follows when no X columns are removed from the model due to collinearity: if there are k columns of known_x's and const = TRUE or omitted, then $df = n - k - 1$. If const = FALSE, then $df = n - k$. In both cases, each X column removed due to collinearity increases df by 1.
- Formulas that return arrays must be entered as array formulas.
- When entering an array constant such as known_x's as an argument, use commas to separate values in the same row and semicolons to separate rows. Separator characters may be different depending on your locale setting in **Regional and Language Options** in **Control Panel**.
- Note that the y-values predicted by the regression equation may not be valid if they are outside the range of the y-values you used to determine the equation.
- The underlying algorithm used in the LINEST function is different than the underlying algorithm used in the SLOPE and INTERCEPT functions. The difference between these algorithms can lead to different results when data is undetermined and collinear. For example, if the data points of the known_y's argument are 0 and the data points of the known_x's argument are 1:
 - LINEST returns a value of 0. The LINEST algorithm is designed to return reasonable results for collinear data, and in this case at least one answer can be found.
 - SLOPE and INTERCEPT return a #DIV/0! error. The SLOPE and INTERCEPT algorithm is designed to look for one and only one answer, and in this case there can be more than one answer.
- In addition to using LOGEST to calculate statistics for other regression types, you can use LINEST to calculate a range of other regression types by entering functions of the x and y variables as the x and y series for LINEST. For example, the following formula:

=LINEST(yvalues, xvalues^COLUMN(\$A:\$C))

works when you have a single column of y-values and a single column of x-values to calculate the cubic (polynomial of order 3) approximation of the form:

$$y = m1*x + m2*x^2 + m3*x^3 + b$$

You can adjust this formula to calculate other types of regression, but in some cases it requires the adjustment of

the output values and other statistics.

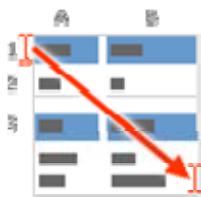
Example 1 Slope and Y-Intercept

The example may be easier to understand if you copy it to a blank worksheet.

 How to copy an example

1. Create a blank workbook or worksheet.
2. Select the example in the Help topic.

NOTE Do not select the row or column headers.



Selecting an example from Help

3. Press CTRL+C.
4. In the worksheet, select cell A1, and press CTRL+V.
5. To switch between viewing the results and viewing the formulas that return the results, press CTRL+` (grave accent), or on the **Formulas** tab, in the **Formula Auditing** group, click the **Show Formulas** button.

	A	B
1	Known y	Known x
2	1	0
3	9	4
4	5	2
5	7	3
	Formula	Formula
	=LINEST(A2:A5,B2:B5,,FALSE)	

NOTE The formula in the example must be entered as an array formula. After copying the example to a blank worksheet, select the range A7:B7 starting with the formula cell. Press F2, and then press CTRL+SHIFT+ENTER. If the formula is not entered as an array formula, the single result is 2.

When entered as an array, the slope (2) and the y-intercept (1) are returned.

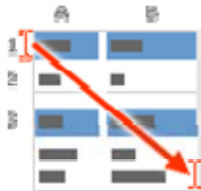
Example 2 Simple Linear Regression

The example may be easier to understand if you copy it to a blank worksheet.

How to copy an example

1. Create a blank workbook or worksheet.
2. Select the example in the Help topic.

NOTE Do not select the row or column headers.



Selecting an example from Help

3. Press CTRL+C.
4. In the worksheet, select cell A1, and press CTRL+V.
5. To switch between viewing the results and viewing the formulas that return the results, press CTRL+` (grave accent), or on the **Formulas** tab, in the **Formula Auditing** group, click the **Show Formulas** button.

	A	B
1	Month	Sales
2	1	3100
3	2	4500
4	3	4400
5	4	5400
6	5	7500
7	6	8100
	Formula	Description (Result)
	=SUM(LINEST(B2:B7, A2:A7)*{9,1}) Estimate sales for the ninth month (11000)	

In general, $SUM(\{m,b\}*x,1\}$ equals $mx + b$, the estimated y-value for a given x-value. You can also use the TREND function.

Example 3 Multiple Linear Regression

Suppose a commercial developer is considering purchasing a group of small office buildings in an established business district.

The developer can use multiple linear regression analysis to estimate the value of an office building in a given area based on the following variables.

Variable	Refers to the
y	Assessed value of the office building
x1	Floor space in square feet
x2	Number of offices
x3	Number of entrances
x4	Age of the office building in years

This example assumes that a straight-line relationship exists between each independent variable (x1, x2, x3, and x4) and the dependent variable (y), the value of office buildings in the area.

The developer randomly chooses a sample of 11 office buildings from a possible 1,500 office buildings and obtains the following data. "Half an entrance" means an entrance for deliveries only.

The example may be easier to understand if you copy it to a blank worksheet.

How to copy an example

1. Create a blank workbook or worksheet.
2. Select the example in the Help topic.

NOTE Do not select the row or column headers.



Selecting an example from Help

3. Press CTRL+C.
4. In the worksheet, select cell A1, and press CTRL+V.
5. To switch between viewing the results and viewing the formulas that return the results, press CTRL+` (grave accent), or on the **Formulas** tab, in the **Formula Auditing** group, click the **Show Formulas** button.

	A	B	C	D	E
1	Floor space (x1)	Offices (x2)	Entrances (x3)	Age (x4)	Assessed value (y)

2	2310	2	2	20	142,000
3	2333	2	2	12	144,000
4	2356	3	1.5	33	151,000
5	2379	3	2	43	150,000
6	2402	2	3	53	139,000
7	2425	4	2	23	169,000
8	2448	2	1.5	99	126,000
9	2471	2	2	34	142,900
10	2494	3	3	23	163,000
11	2517	4	4	55	169,000
12	2540	2	3	22	149,000

Formula

=LINEST(E2:E12,A2:D12,TRUE,TRUE)

NOTE The formula in the example must be entered as an array formula. After copying the example to a blank worksheet, select the range A14:E18 starting with the formula cell. Press F2, and then press CTRL+SHIFT+ENTER. If the formula is not entered as an array formula, the single result is -234.2371645.

When entered as an array, the following regression statistics are returned. Use this key to identify the statistic you want.

	A	B	C	D	E	F
1	m_n	m_{n-1}	...	m_2	m_1	b
2	se_n	se_{n-1}	...	se_2	se_1	se_b
3	r^2	se_y				
4	F	df				
5	ss_{reg}	ss_{resid}				

The multiple regression equation, $y = m_1*x_1 + m_2*x_2 + m_3*x_3 + m_4*x_4 + b$, can now be obtained using the values from row 14:

$$y = 27.64*x_1 + 12,530*x_2 + 2,553*x_3 - 234.24*x_4 + 52,318$$

The developer can now estimate the assessed value of an office building in the same area that has 2,500 square feet, three offices, and two entrances and is 25 years old, by using the following equation:

$$y = 27.64*2500 + 12530*3 + 2553*2 - 234.24*25 + 52318 = \$158,261$$

Or you can copy the following table to cell A21 of the example workbook.

Floor space (x1)	Offices (x2)	Entrances (x3)	Age (x4)	Assessed value (y)
2500	3	2	25	=D14*A22 + C14*B22 + B14*C22 + A14*D22 + E14

You can also use the TREND function to calculate this value.

Example 4 Using the F and r2 Statistics

In the previous example, the coefficient of determination, or r^2 , is 0.99675 (see cell A17 in the output for LINEST), which would indicate a strong relationship between the independent variables and the sale price. You can use the F statistic to determine whether these results, with such a high r^2 value, occurred by chance.

Assume for the moment that in fact there is no relationship among the variables, but that you have drawn a rare sample of 11 office buildings that causes the statistical analysis to demonstrate a strong relationship. The term "Alpha" is used for the probability of erroneously concluding that there is a relationship.

F and df in LINEST output can be used to assess the likelihood of a higher F value occurring by chance. F can be compared with critical values in published F-distribution tables or Excel's FDIST can be used to calculate the probability of a larger F value occurring by chance. The appropriate F distribution has v_1 and v_2 degrees of freedom. If n is the number of data points and $\text{const} = \text{TRUE}$ or omitted, then $v_1 = n - \text{df} - 1$ and $v_2 = \text{df}$. (If $\text{const} = \text{FALSE}$, then $v_1 = n - \text{df}$ and $v_2 = \text{df}$.) Excel's FDIST(F, v_1 , v_2) will return the probability of a higher F value occurring by chance. In Example 4, $\text{df} = 6$ (cell B18) and $F = 459.753674$ (cell A18).

Assuming an Alpha value of 0.05, $v_1 = 11 - 6 - 1 = 4$ and $v_2 = 6$, the critical level of F is 4.53. Since $F = 459.753674$ is much higher than 4.53, it is extremely unlikely that an F value this high occurred by chance. (With Alpha = 0.05, the hypothesis that there is no relationship between known_y's and known_x's is to be rejected when F exceeds the critical level, 4.53.) Using Excel's FDIST you can obtain the probability that an F value this high occurred by chance. FDIST(459.753674, 4, 6) = 1.37E-7, an extremely small probability. You can conclude, either by finding the critical level of F in a table or by using Excel's FDIST, that the regression equation is useful in predicting the assessed value of office buildings in this area. Remember that it is critical to use correct values of v_1 and v_2 computed in the previous paragraph.

Example 5 Calculating the t-Statistics

Another hypothesis test will determine whether each slope coefficient is useful in estimating the assessed value of an office building in example 3. For example, to test the age coefficient for statistical significance, divide -234.24 (age slope coefficient) by 13.268 (the estimated standard error of age coefficients in cell A15). The following is the t-observed value:

$$t = m_4 \div \text{se}_4 = -234.24 \div 13.268 = -17.7$$

If the absolute value of t is sufficiently high, it can be concluded that the slope coefficient is useful in estimating the assessed value of an office building in Example 3. The table below shows the absolute values of the 4 t-observed values.

If you consult a table in a statistics manual, you will find that t-critical, two tailed, with 6 degrees of freedom and Alpha = 0.05 is 2.447. This critical value can also be found using Excel's TINV function. TINV(0.05,6) = 2.447. Because the absolute value of t , 17.7, is greater than 2.447, age is an important variable when estimating the assessed value of an office building. Each of the other independent variables can be tested for statistical significance in a similar manner. The following are the t-observed values for each of the independent variables.

Variable	t-observed value
Floor space	5.1

Number of offices	31.3
Number of entrances	4.8
Age	17.7

These values all have an absolute value greater than 2.447; therefore, all the variables used in the regression equation are useful in predicting the assessed value of office buildings in this area.

See Also

- [GROWTH](#)
- [LOGEST](#)
- [Statistical functions](#)
- [TREND](#)