MORRISVILLE
EST. 1908
STATE UNIVERSITY OF NEW YORK

School of Agriculture, Business, and Technology

Common Excel Functions for Correlation and Regression

Developed By

Dr. Walid H. Shayya

A. CORREL: Returns the correlation coefficient which determines the relationship between two properties.

Syntax: CORREL(Array1,Array2)
Array1 is a cell range of values while Array2 is a second cell range of values. The arguments must be numbers, array, or references that contain numbers. If an array or reference argument contains text, logical values, or empty cells, those values are ignored; however, cells with the value zero are included. CORREL returns the "#N/A" error value when Array1 and Array2 have a different number of data points.

Example: CORREL({2,4,3,5,6},{17,15,16,13,12}) equals -0.99124

B. FORECAST: Calculates a future value by using existing values. The known values are existing x-values and y-values, and the new value is predicted by using linear regression.

Syntax: FORECAST(x,Known_y's,Known_x's)
X is the data point for which you want to predict a value while Known_y's is the dependent array (or range) of data and Known_x's is the independent array (or range) of data. If x is nonnumeric, FORECAST returns the "#VALUE!" error value. FORECAST returns the #N/A error value when known_y's and/or known_x's are empty or contain a different number of data points.

Example: FORECAST(20,{2,4,3,5,6},{17,15,16,13,12}) equals 0.67442

C. INTERCEPT: Calculates the point at which a line will intersect the y-axis by using existing x-values and y-values. The intercept point is based on a best-fit regression line plotted through the known x-values and known y-values.

Syntax: INTERCEPT(Known_y's,Known_x's)
Known_y's is the dependent set of observations (or data) while Known_x's is the independent set of observations or data. The arguments should be either numbers, arrays, or references that contain numbers. If an array (or reference argument) contains text or empty cells, those values are ignored; however, cells with the value zero are included. Again, INTERCEPT returns the "#N/A" error value when Known_y's and Known_x's contain a different number of data points (or contain no data points)

Example: INTERCEPT({2,4,3,5,6},{17,15,16,13,12}) equals 15.03488

D. LINEST: Calculates the statistics for a line by using the "least squares" method to calculate a straight line that best fits your data. LINEST returns an array of values that describes the line. For this reason, it must be entered as an array formula (use the “Help” option in Excel for more information about array formulas).

The equation for the line is:

$y = mx + b$ (for simple regression)
$y = m_1x_1 + m_2x_2 + ... + b$ (for multiple regression)

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. The array that LINEST returns is $\{m_n,m_n\text{-}1,...,m_1,b\}$. 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 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 (refer to the following table for the right order)
$\{m_n,m_{n-1},...,m_1,b;se_n,se_{n-1},...,se_1,seb;r2,sey;F,df;ssreg,ssresid\}$

	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	d_f				
5	ss_{reg}	ss_{resid}				

In addition to the standard regression statistics that include the regression coefficients and the intercept, the coefficient of determination is of interest and results from setting Stats as True. This parameter is represented by r2 which compares estimated and actual y-values. r2 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.

If Stats is FALSE or omitted, LINEST returns only the m-coefficients and the constant b

Example: LINEST({2,4,3,5,6},{17,15,16,13,12}) equals {-0.755813953,15.03488372}, with the slope = m = -0.755813953 and y-intercept = b = 15.03488372. Other examples will be discussed in the class.

E. RSQ: Returns the square of the correlation coefficient through data points in known_y's and known_x's. The r-squared value can be interpreted as the proportion of the variance in y attributable to the variance in x.

Syntax: RSQ(Known_y's,Known_x's)
Known_y's is an array (or range) of data points while Known_x's is an array (or range) of data points. The arguments must be either numbers, arrays, or references that contain numbers. If an array or reference argument contains text, or empty cells, those values are ignored; however, cells with the value zero are included. If Known_y's and Known_x's are empty or have a different number of data points, RSQ returns the "#N/A" error value.

Example: RSQ({2,4,3,5,6},{17,15,16,13,12}) equals 0.98255814

F. SLOPE: Calculates the slope of the linear regression line through data points in known_y's and known_x's. The slope is the vertical distance divided by the horizontal distance between any two points on the line, which is the rate of change along the regression line.

Syntax: SLOPE(Known_y's,Known_x's)
Known_y's is an array (or cell range) of numeric dependent data points while Known_x's is the set of independent data points. The arguments must be either numbers, arrays, or references that contain numbers. If an array or reference argument contains text, logical values, or empty cells, those values are ignored; however, cells with the value zero are included. If Known_y's and Known_x's are empty or have a different number of data points, SLOPE returns the "#N/A" error value.

Example: SLOPE({2,4,3,5,6},{17,15,16,13,12}) equals -0.755813953

G. TREND: Returns values along a linear trend after it fits a straight line (using the method of least squares) through the arrays known_y's and known_x's. It returns the y-values along that line for the array of new_x's that you specify.

Syntax: TREND(Known_y's,Known_x's,New_x's,Const)
Known_y's is the set of y-values you already know in the relationship y = mx + b while Known_x's is an optional set of x-values that you may already know in the same relationship. 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. 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. New_x's are new x-values for which you want TREND to return corresponding y-values. New_x's must include a column (or row) for each independent variable, just as Known_x's does. So, if Known_y's is in a single column, Known_x's and New_x's must have the same number of columns. If Known_y's is in a single row, Known_x's and New_x's must have the same number of rows. If New_x's is omitted, it is assumed to be the same as Known_x'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 so that y = mx.

TREND can also be used for polynomial curve fitting by regressing against the same variable raised to different powers. For example, suppose column A contains y-values and column B contains x-values. You can enter x^2 in column C, x^3 in column D, and so on, and then regress columns B through D against column A. Formulas that return arrays must be entered as array formulas.

When entering an array constant for an argument such as Known_x's, use commas to separate values in the same row and semicolons to separate rows.

Example: TREND({2,4,3,5,6},{17,15,16,13,12},{17,15,16,13,12}) equals {2.186046512, 3.697674419,2.941860465,5.209302326,5.965116279}