

VII. RELATIONSHIPS BETWEEN X-Y TYPES OF DATA SETS

GENERAL CONCEPTS

Given that we have an X-Y type of dataset, in general there are 4 ways to get information about the relationships between the X and Y values in Excel and to make some predictions about how the X and Y values relate. They are

1. Correlation (Section 8)
2. Regression, linear (Section 9)
3. Regression, non-linear (Section 10)
4. Trend (trendlines) (Section 10)
5. Forecast (Section 11)

There is also another general class where there is a complex relationship between variables that cannot be expressed in the form of an equation. This is the general “what-if” type of problem.

6. What-If problems. (In section 12)

In the X-Y data set, complex variables (i.e. $X = \text{real} + \text{imaginary terms}$) cannot be considered, since the Excel functions and VB routines needed for solutions cannot handle the Excel complex number, since it is a string variable, not a number variable. Also X as “text” (e.g. words, symbols, phrases, etc) cannot be input to the standard Excel numerical functions and routines. Essentially the X-Y data set must exist as a floating point double precision set or be able to be converted to a double precision floating point number set.

If X or Y represent non-numerical variables, then we have to recognize that Excel (as-is) is not able to obtain a solution. Some form of a solution can be obtained, if the non-numerical values can be ranked, and the rank (as a number) is used as a measure of the variable.

In a general sense, there are also these conditions inherent in Excel

- a. X in a sense is a vector, and Y is a single value.
- b. X is a matrix and Y is a vector.
- c. Both X and Y are vectors.

If X and/or Y are expressed as matrices (e.g. more dimensions than a simple vector), there are size limitations in Excel.

SOME GENERAL CONSIDERATIONS

These six types in a sense overlap, leading to application problems and confusion on what these are intended to do. This confusion shows up on all the Excel based websites and the many sites that deal with questions and “peer” responses. Many of those Excel reference, how-to and self-help books do not distinguish between these six ways.

CORRELATION

A correlation cannot be extrapolated to give trends or to make predictions. Correlations can be tested for statistical properties. Correlations occur as worksheet functions.

REGRESSION

A regression only provides numerical values (coefficients of a stated model) and statistical information about the fit of the model to the data. If the equation is extended beyond the range of the data, the extensions may be considered as “predictions” rather than “forecasts”. From the results of the regression, one can also construct upper and lower confidence bands about the regression equation that can be extended beyond the range of the data. These confidence bands allow us to estimate the uncertainty of any prediction. Excel does not provide functions to calculate these confidence intervals.

REGRESSION-LINEAR

The core is the array function LINEST. It handles multivariate data, and does polynomial regression by setting the powers of each data X value as additional multivariate X values. The Excel 2003 and 2007 versions have corrected major problems with prior Excel versions.

REGRESSION –NONLINEAR

Excel does a form of non-linear regression using LINEST. It does a linear regression on the logarithms of the variables, and therefore is not a true non-linear regression. The resulting transformed equation parameter values are not equal to parameter values from a true non-linear regression.

The Excel Solver routine can be used to directly solve non-linear regression problems. However it has some major faults and problems with the accuracy of results and the stability of the solution.

TREND

This is a function that is only available in the CHART module. It is called “trendline”. It combines regression tools with an undefined block of logic that changes the regression. As a result the output equation on the chart and the corresponding line is not a true regression on the data. It does not output directly to worksheet cells, but is shown as a line and equation directly in the chart. The equation parameter values are shown in a very limited form. An R-squared value (as a measure of fit) can be selected and shown in addition to the curve and equation.

A very common problem is that Excel “users” assume that getting a “trendline” from a data set is the same as a “regression” on the data set. This is not always the case. Another common problem is that Excel “users” assume that a “trend” is the same as a “forecast”.

FORECAST

Forecasting is basically associated with time series. It is a means of predicting values of variables important in decision processes from historic values of selected variables, Typical applications involve, Inventory Management, Production Planning, Financial Planning, Staff Scheduling, Facilities Planning and Process Control.

The Excel array function FORECAST does a LINEST on the input data and outputs a range of Y values for an input range of X values. It is really a “prediction”.

The Analysis Toolpac Exponential Smoothing tool does a simple forecast. It predicts a value based on the prior period data and prior forecasts. The new forecast is adjusted for the error in prior forecasts. The tool uses the smoothing constant α , the magnitude of which determines how strongly forecasts respond to errors in the prior forecast.

Excel does not provide better forecasting tools.

THE APPROACH IN THIS ANALYSIS OF EXCEL

Section 8 covers the area of correlations and covariances.

Section 9 covers the area of linear regression*.

Section 10 covers the area of non-linear regression

Section 11 covers the area of trendlines

Section 12 covers the area of forecasts

Section 13 covers the what-if solutions

Excel does provide solutions in all these areas, but they may be limited in what they can handle and whether they output correct solutions.

Reports of operational problems, faults, errors and specific computing problems relating to each of these sections. are included in each section. These come from Emails, messages, articles relating to Excel, and tests performed by the author.