

NOTE V: REGRESSION NORMAL PROBABILITY PLOT

LINEAR REGRESSION OUTPUTS:

In the Data Analysis Regression routine, an input box appears. At the bottom there are boxes to check if specific outputs are requested. They are:

- Residuals
- Standardized Residuals
- Residual Plots
- Line Fit Plots
- Normal Probability Plots

Residuals: Gives three columns of data below the regression output information block. The first column is observation and is just a sequence number referring the sequence of input data values. The second column is the predicted Y value, and the third column is the difference between the actual data Y value and the calculated Y value, based on the regression coefficients listed.

Standardized Residuals: Gives a column of values, which come from the division of the residual by the standard deviation of the residuals. Use it to look for outliers.

Residual Plots: Gives numerous plots of the residual versus each of the X variables. If only one X is input, there is only one chart. The chart can be changed just like all other Excel charts. Use it to detect patterns, indicated deviations from the regression equation.

Line Fit Plots: Generates a chart with a plot of predicted Y values versus observed Y values. Use it to determine if the regression is a suitable fit over the entire range of the data.

Normal Probability Plots:

Two new columns of values are generated in the same space below the regression output. The first column is a calculated value from:

$$100 * ((\text{sequence number}/n) - (1/2*n)),$$

where sequence number is the numbers from 1 to n. The second column is the sorted observed Y values from lowest to highest. These columns bear no relationship to the residual columns described above, although they occupy the same row space. The chart is just a scatter plot of these two columns. It bears no relationship to normal distribution characteristics.

The intent was to generate a form of a Meier-Kaplan plot of the Y data to determine if the distribution of Y values is normal. If it is normal, then correlation coefficients may be more important than regression coefficients.

The more common application today of the term “Normal Probability Plot”, is to plot the residuals versus a z value (or cumulative normal percentile) derived from the normal probability distribution for the ranking location of the residual. This gives a visual look to determine if the residuals are actually normally distributed. The assumption of error (as residuals) being normally distributed is an important assumption. The probability values on the coefficients and coefficient confidence intervals are based on this assumption. If the plot is roughly a straight line, then the assumption is valid.