Regression Analysis Tutorial

In a simple linear regression, it is assumed that only one variable, the independent variable, determines the dependent variable. These variables may also be referred to as the explanatory variable (x) and the response variable (y). Performing a linear regression means to determine which straight line best cancels out the deviations above and below the line. These deviations or distances with respect to the regression line are called residuals.

In theory, a straight line can always be found that cancels out the deviations. However, caution must be used as a straight line may not always be the "best fit".

Whether a TI graphing calculator or Microsoft Excel is used, one method of determining the regression type is by trial and error. Simply try the logarithmic, exponential, or polynomial, to determine which gives an R-squared value closest to one. A value of R^2 less than 0.90 corresponds to a poor fit of data to the regression type.

An alternative to using trial and error is to plot a graph of Residuals vs Independent Variable. A residual plot has the residuals on the y-axis and the independent variable on the x-axis. Residual plots magnify the residuals which allows for easier identification of patterns.

When creating residual plots, either a TI graphing calculator or Excel can be used. When using Excel, check if Data Analysis is found on the bottom of the Tools menu. If Data Analysis is not displayed, use the Add-Ins option also found under the Tools menu to load the Analysis ToolPak. Using the TI calculator is a bit different and the steps can be found in the TI-XX Graphing Calculator Guidebook.

The following analysis applies to both the TI graphing calculator and Excel.

Important points to consider when analyzing residual plots:

- When the residuals are equally distributed above and below the zero-line (Y = 0), the regression line is a good model for the data.

  When the residual plot generates a curved path, another regression type is needed. Other regression types include quadratic, cubic, exponential, etc.

- An increase or decrease in the spread of residuals as the independent variable increases is important to note. As the values for the dependent variable increase for larger values of the independent variable, the y values become less accurate.

- Data associated with large residuals called outliers can either be ignored or examined more closely for errors.

- Data associated with points that are extreme in the x-direction should be examined closely. While these points may not have high residuals, they may change the position (the slope) of the regression line. When these points, called influential points, are not surrounded by other data points, they can "pull" the regression line towards themselves and change the slope.

- Just looking at the residual plot, these influential points may be overlooked. It would be beneficial to perform a linear regression on the original data, leaving out the influential points.

The graphs found in Data Set 1 and Data Set 2 were made with Excel and a TI-83 Plus graphing calculator. The data was entered using an Excel spreadsheet and the data and graphs were copied from Excel into Word. To create the TI graphs, the data was copied into TI Interactive, stored in a list, and copied to the TI via TI Connectivity Cable. After generating the graphs on the TI, the TI Screen Capture feature found in TI Connect was used to upload the graphs to the PC.

Data Set 1 is a graph of Pressure vs Temperature for the vapor pressure of water at various temperatures. A linear regression for this set of data is clearly incorrect but it shows that, in theory, a straight line can always be found that cancels out the deviations.

Data Set 2 is a graph of Temperature vs Time which models Newton's Law of Cooling. In this example, the incorrectness of a linear regression is more subtle. If one examines the Temperature Residual Plot, it becomes apparent that a straight line is not the "best fit".