**Correlation and Regression**

We use Correlation and Regression to analyze the direction and strength of the mathematical relationship between variables. For simplicity, we will just use two variables, which we will call “X” and “Y”. “X” will be the independent variable, and “Y” will be the dependent variable. Again for simplicity we will explore a linear relationship. We will use data from a chemistry experiment to test whether the amount of a copper compound generated in a particular experiment is related to the temperature at which the experiment was run. Because we are free to select various values for the temperature, it is the “independent” variable, and thus will be depicted on the “X” axis. The amount of copper is the “dependent” variable, and it will be on the “Y” axis.

|  |  |
| --- | --- |
| TEMP | COPPER |
| 10 | 17 |
| 20 | 21 |
| 30 | 25 |
| 40 | 28 |
| 50 | 33 |
| 60 | 40 |
| 70 | 49 |

Start with a Scatter-plot of the data (Graph > Scatterplot).



Is there an apparent pattern here?

Go to Stat > Regression > Fitted line plot



Note the value of R-Sq. This is the amount of variation in copper that is explained by variation in temperature, and it is a very high value indicating a strong relationship between the two variables. The upward slope of the line indicates that the relationship is positive ( a downward slope would indicate an inverse relationship). R-Sq is called the” Coefficient of Determination”. It varies from 0 to 1. The square root of R-Sq is called the “Correlation Coefficient” and it varies from -1 (strong inverse relation) to 1 (strong positive relation). The formula at the top of the panel is the mathematical relationship that was our original goal. For example the amount of copper to be expected at a temperature of 45 is 10.14 plus 0.5071 times 45, or 32.96 units.

Note that we did not require either of the variables to have any particular distribution. However, for the analysis to be valid the “residuals”, described below must be normally distributed. Go to Stat > Regression > Fitted Line Plot > Storage > Residuals. Minitab will calculate the residuals, and store the data in the first available column. Go to Stats > Basic Stats > Normality test., Select the column containing the residuals, then observe the P-value. If it is greater than 0.05 the residuals are normally distributed.

The straight line is called a “Least Squares Line” meaning that the sum of the squared values of the vertical distances from the points to this line, called the “residuals”, is less than the corresponding value to any other line…we could have drawn an infinite number of straight lines through the array of data-points!

It is also true, but not obvious that the line must pass through the point X-Bar, Y-Bar.

Finally, when using the formula to calculate “Y” values, we should not extrapolate outside the bounds of our data.