

NOTE L: AUTOCORRELATION

McCullough (2000) concluded that Excel's CORREL function gives wrong values when it is used to do a lag-1 autocorrelation on univariate data. He said, "...it is also obvious that Excel uses a deficient algorithm for computing the sample correlation coefficient". This was in regard to the low LRE values CORREL outputs on the StRD univariate data sets.

The StRD reference sheets on the univariate data sets state that the univariate correlation being used is "Autocorrelation Coefficient (lag 1) r(1):" The StRD formula for the autocorrelation is:

$$\text{Tau} = \frac{\sum_{i=1}^n \{(Y_i - Y_m) \times (Y_{i-1} - Y_m)\}}{\{\sum_{i=1}^n (Y_i - Y_m)^2\}}$$

Where: Y_m is the average of the 1 to n data set.

McCullough refers to it as "the first-order autocorrelation coefficient".

He used the Excel CORREL function, which is defined as:

$$\text{CORREL}(X,Y) = \text{COVAR}(X,Y) / (\text{STDEV}(X) * \text{STDEV}(Y))$$

Where:

$$\text{COVAR} = \frac{\sum_{i=1}^n \{(X_i - X_m) \times (Y_i - Y_m)\}}{n}$$

STDEV = Excel's standard deviation function

These are two different formulas. The equation for the Autocorrelation Coefficient uses a common mean and a common standard deviation. The CORREL function uses separate means and separate standard deviations. They become numerically different because of the 1-lag, which leads to different means and different standard deviations.

The algorithm shown in note Z was used in testing new algorithms to the Longley data base.

CORRELATION (EXCEL 2000)

The covariance function (COVAR) is robust against number of significant figures. As shown in the Function Reference, The input range of both the X and Y values is internally centered about the means of the two variables. The divisor of the summed product of the differences of each variable from its mean is n, rather than n-2. Most statistics books take the covariance computation as being divided by n, the number of values. Rather than the number of degrees of freedom.

The correlation coefficient function (CORREL) is not robust against a large number of significant figures. According to the Function Reference, CORREL uses COVAR and STDEV functions, COVAR for the numerator and STDEV on both variables to obtain (as a product) the denominator. STDEV is not robust and is the main contributor to inaccuracies. Also the value is biased to $n/(n-1)$ when X and Y are identical.

The Pearson product moment (PEARSON) is essentially CORREL without the $n/(n-1)$ bias. It however is not robust, and easily produces inaccurate values as the number of significant figures increases. If used, all data should be first centered.