

XI. CHART TRENDLINE REGRESSION.....	2
DESCRIPTION AND LIMITATIONS	2
SOME TYPICAL CHARTS THAT ARISE, GIVEN THE EQUATIONS IN TABLE 11-1	3
WHERE THE FITTED EQUATION IS SHOWN.....	6
CHOICE OF THE ORIGIN OF THE FITTED EQUATION	6
TRENDLINE INTERNAL EQUATION-DATA VALIDITY CHECKS	6
FITTING THE LINEAR MODEL	7
FITTING THE EXPONENTIAL MODEL	7
FITTING POLYNOMIAL MODEL	8
FITTING A POWER FUNCTION MODEL.....	8
TRENDLINE PERFORMANCE.....	8
SETTING UP THE CHART FOR A TRENDLINE	9
REMOVING A TRENDLINE FROM A CHART.....	10
TRENDLINE FAULTS AND ERRORS.....	10
CHART AND EQUATION SELECTION FAULTS.....	10
ACCURACY OF THE DISPLAYED EQUATION COEFFICIENTS.....	11
THE TEST FOR COEFFICIENT ACCURACIES.....	11
THE ANALYSIS OF TEST RESULTS	11
TRENDLINE EQUATION PARAMETER VALUE INSTABILITIES.....	12
WAMPLER SET 1	12
WAMPLER 2.....	14
WAMPLER 3.....	15
WAMPLER 4.....	16
WAMPLER 5.....	19
TRENDLINE EXPONENTIAL FUNCTION	22
EXCEL 2003-EXCEL 2007 EQUATION DISPLAY FAULT	25
CONCLUSIONS.....	27
THE BOTTOM LINE.....	27

XI. CHART TRENDLINE REGRESSION

DESCRIPTION AND LIMITATIONS

The Excel Chart module has an operation called Trendline that will fit chart-plotted variables to an equation. Microsoft defines it as, “Trendlines are graphical representations of trends in data that you can use to analyze problems of prediction. Such analysis is also called **regression analysis**. By using regression analysis, you can extend a trendline in a chart beyond the actual data to predict future values.”

The user selects the type:

- Linear Trendlines
- Logarithmic Trendlines
- Polynomial Trendlines
- Power Trendlines
- Exponential Trendlines
- Moving Average Trendlines

Then through a series of menu selections, the equation fitted to the data and its “curve” are shown on the chart. The curve can also be projected beyond the range of the data.

Trendline is a very attractive regression tool, but it misleads the user into using wrong regressions. The logarithmic and exponential transformations of data to obtain a linear set for exponential and power types, is a wrong approach, see Hesse (2006). This is a distinct error on the part of Microsoft that has persisted since the very first Excel version. The returned transformed coefficient values are basically in error, since the statistical objective of minimizing the sum-of-the-true-residuals-squared is not minimized. Also for “through-the-origin power and exponential models”, the origin defined by Microsoft may not be the user’s intended origin.

Microsoft in the specific ”Help” files talks about the use of the normal equations for solutions. The normal equations are basically used only when ONE or TWO parameters are to be found given only X and Y vectors.

$$m = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2}$$

$$b = \bar{y} - m\bar{x}$$

For the linear zero intercept model, only m is calculated and reported.

Table 11-1 lists the specific functions that can be fitted:

Table 11-1: Normal Equation Input Data

Equation	The Y variable	The X variable	Location of Origin
Linear	Y	X	Y=0 and X=0.
Logarithmic	Y	Ln (X)	Y=0 and X=1
Power	Ln (Y)	Ln (X)	Y=1 and X=1
Exponential	Ln (Y)	X	Y=1 and X=0

“Ln()” refers to the natural log (to the base e).

The Excel Help file shows these normal equations. Shown are the forms for only one X variable.

SOME TYPICAL CHARTS THAT ARISE, GIVEN THE EQUATIONS IN TABLE 11-1

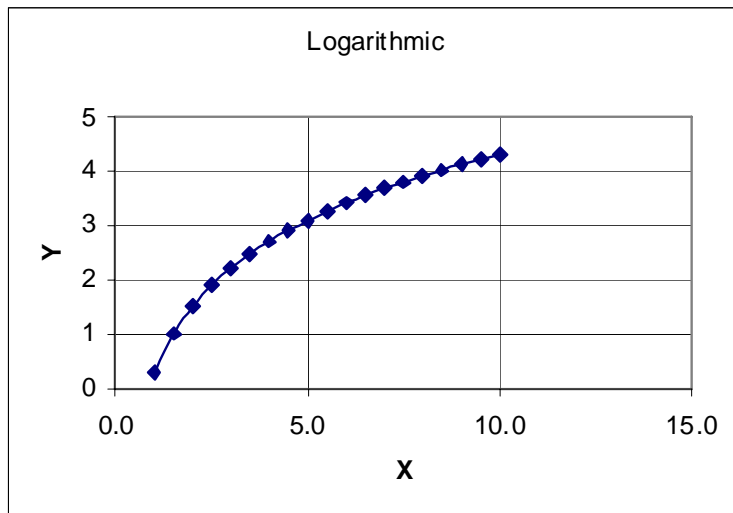
The following are essentially illustrative of the general behavior of Y values from the different equations given in table 11-1. The shapes with negative exponents are different from the shapes with positive exponents.¹ Also included is information about the fitted trendline. When the word “exact” is used, the Excel calculated value has an LRE value with reference to the true value greater than 14,

LINEAR

This is of course a straight line on the chart.

LOGARITHMIC

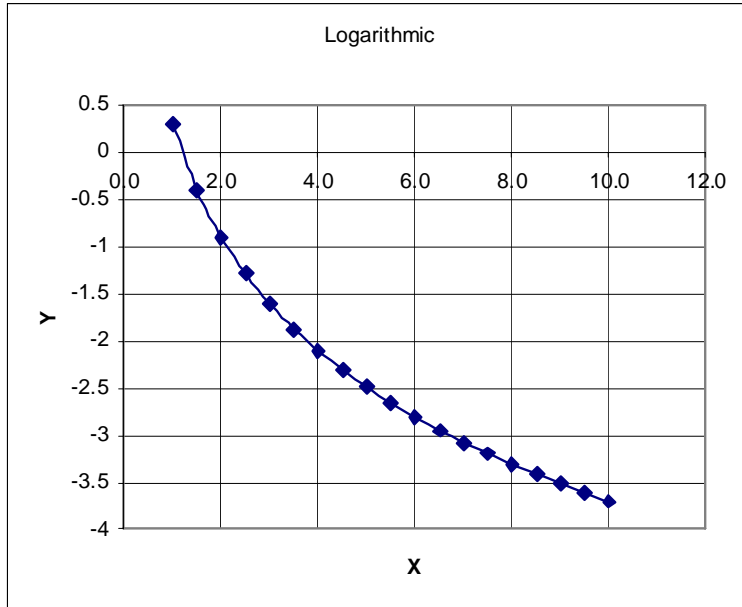
Figure 11-1: A Logarithmic Curve, $Y = 0.3 + 4 * \text{Log}(x)$



¹ For these charts the plotted y values are exact. The Trendline on the chart would be an exact copy of the shown equation generating the points.

Here the logarithm is to the base 10. The Excel computations are to the natural log. The returned Trendline is $Y = 0.3 + 1.737177927613 * \ln(x)$. The coefficients are exact.

Figure 11-2: A Logarithmic Curve, $Y = 0.3 - 4 * \text{Log}(x)$

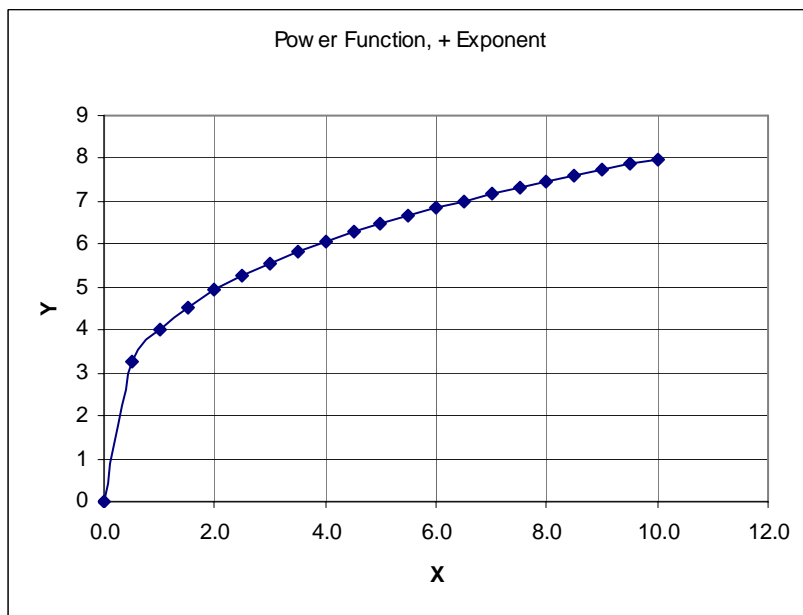


Here again the logarithm is to the base 10. The Excel computations are to the natural log. The returned Trendline is $Y = 0.3 - 1.737177927613 * \ln(x)$. The coefficients are exact.

The logarithmic curve has a distinct shape. The left hand region does not level out, but remains curved upwards or downwards.

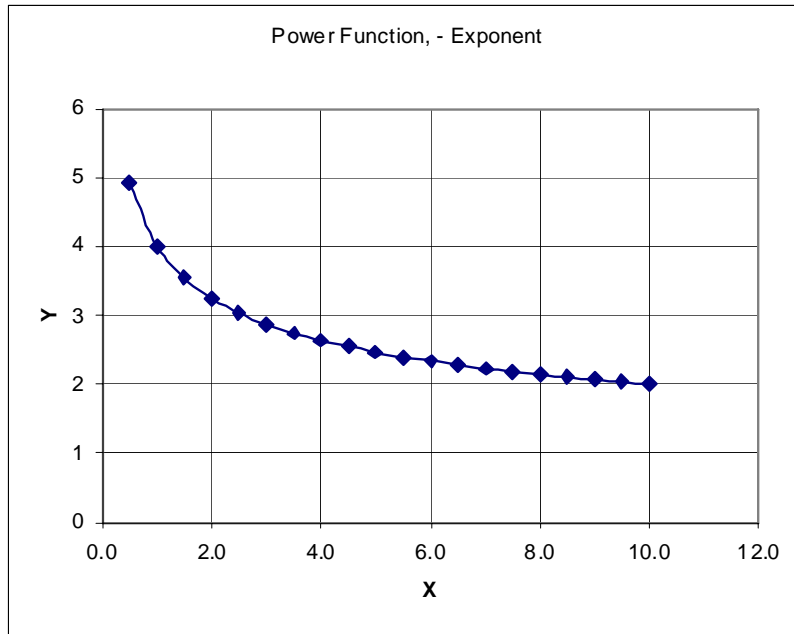
POWER

Figure 11-3: A Power Function Curve, $Y = 4x^{0.3}$



In this case, the data set includes the point 0,0. From a function standpoint this is a valid point, but from the transform standpoint, the 0,0 is not valid. The Power Function option (in the menu) is blocked from being used. When 0,0 is removed, the Power Option can be selected, and the Trendline equation is exactly as shown above in the title.

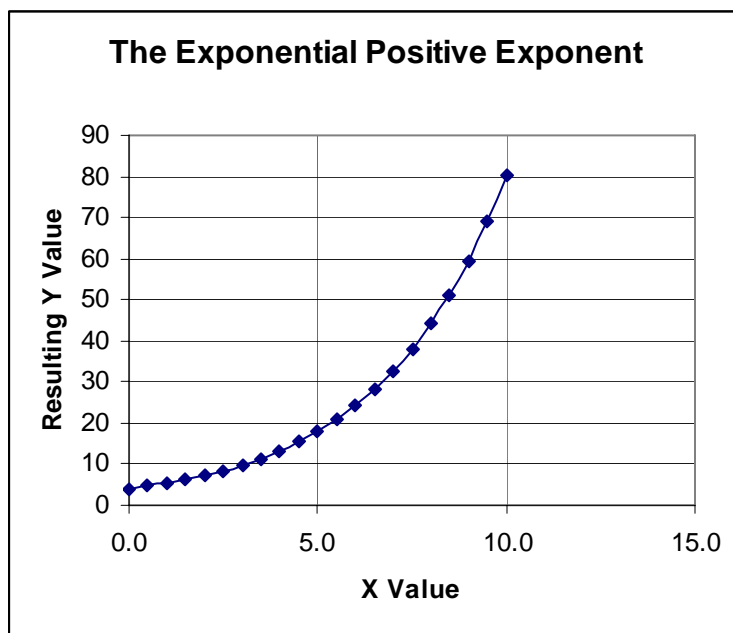
Figure 11-4: A Power Function Curve, $Y = 4x^{-0.3}$



Here again, the trendline coefficients are exact.

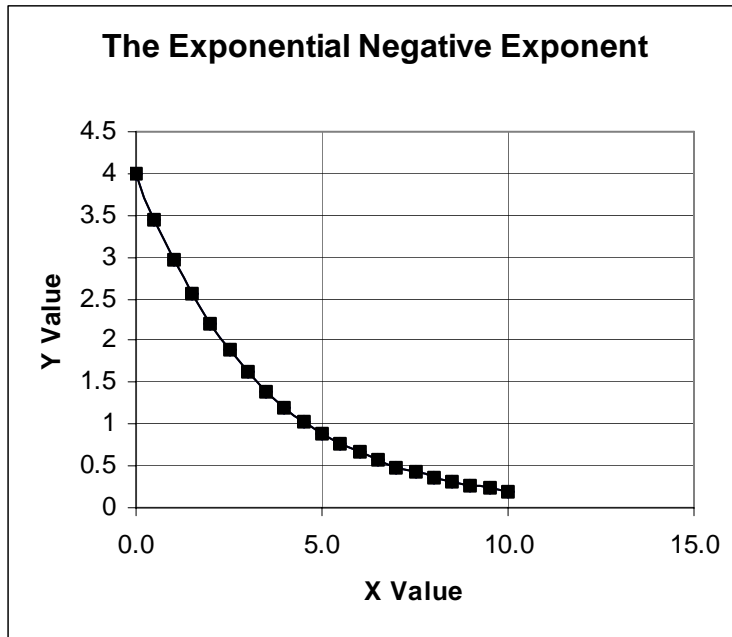
EXPONENTIAL

Figure 11-5: An Exponential Curve, $Y = 4e^{+0.3x}$



The trendline coefficients here are exact.

Figure 11-6: An Exponential Curve, $Y = 4e^{-0.3x}$



The trendline coefficients here are exact.

One should note that there is a certain similarity among the above curves, in the manner and shape of the curves. Consequently the only way to select one from the above would be to look at the theoretical aspects and to look at the variance of the data from the fitted curve.

WHERE THE FITTED EQUATION IS SHOWN

On a chart (see figure 11-7 below), the fitted equation line is shown together with equation parameter values (when the equation display option is selected) above and to the left (or right) of the curve. The number formats are fixed. The equation envelope² can be moved within the chart to where the coefficient values can be read. Once the equation envelope is selected, the shown coefficient values can be expanded, using the same method that numbers in cells are expanded.

CHOICE OF THE ORIGIN OF THE FITTED EQUATION

If the Trendline equation must pass through the origin, this option must be set in the Trendline setup menu. It is a selection from “The Intercept Option” input.

TRENDLINE INTERNAL EQUATION-DATA VALIDITY CHECKS

Trendline limits the choices of equations, based on the selected source data. If the data set includes zero-zero, negative values, a zero X value, etc, the equations that would not be valid for the given data set, do not appear in the Trendline equation selection menu.

² Move the cursor pointer over the shown equation. Right click the mouse, and the equation envelope box will appear.

FITTING THE LINEAR MODEL

The linear equation can be directly solved using any one of many matrix solution methods. The term “normal equations” is either of the form shown above, or a matrix solution method using Gaussian Elimination.

The chart module is separate from the main Excel function/subroutine set. It does not use the Excel LINEST function that is installed in the Excel version being used. Microsoft does not give any information in the KBA files or in Help about the Chart solution methods actually used.

FITTING THE EXPONENTIAL MODEL

There are actually two forms of the exponential that are commonly in use:

1. $y = A * \text{EXP}(B * x)$
2. $y = A * (1 - \text{EXP}(B * x))$

where A and B are fitted constants. The resulting curve has basically four shapes, choice of forms 1 or 2 and the two signs of B.

The x data value signs have to be considered. Generally the data sets would have x values all positive or all negative. If there is a mix, and the data appears to be generally symmetrical about the Y axis, then the exponential as shown above may not be applicable. The y values all have to be positive or all negative since the EXP function does not change signs, The case defines the sign of A.

The data points at (or near) the Y axis at X=0 need to be considered since $\text{EXP}(0) = 1$.

If the Y values at X=0 appear to be zero then form 2 would be more appropriate than form 1. If however the values at X=0 appear to be positive, then form 1 would be more appropriate, where the value of B represents the point where the function crosses the Y axis.

If the Y values appear to increase as X increases, then form 1 may be appropriate with the fitted B value as a positive number. If the Y values tend to level out as X increases, then form 2 (with B negative) would be more appropriate. Form 1 cannot be fitted to data that “levels out” as X increases.

Therefore both forms 1 and 2 should be considered as valid representations of the exponential function.

Microsoft only thought of form 1, because it is the only form that allows a simple transformation of x variables to where the linear regression module in Chart can be used. The transformation forces all x and y values to be positive. This transformation also results in wrong coefficient values, and the failures of Excel to pass the NIST tests. The errors in returned coefficient values when the log transformation is used, is covered in Section 10 on non-linear equation fitting.

When the exponential equation is of form 2, Trendlines is of no help. One has to go to non-linear equation fitting to get correct A and B values.

FITTING POLYNOMIAL MODEL

A heavy use of Trendlines is to fit polynomials. Polynomials are in essence a multi-variant X equation, where the powers of X are treated as the different X variables. The normal equations shown above from Help cannot be used here.

The polynomial model is basically solved using a linear multivariate solution. The normal equations do not work well in this application, since the strong dependence (i.e. co-linearity) of the successive, x^2 , x^3 , x^4 , etc terms may result in computational errors. See section 9.

FITTING A POWER FUNCTION MODEL

The power function is a non-linear equation. Microsoft transforms the X and Y values by taking the logs of the input data and then solving the resulting transformed values by the “normal equations” such as that described above. The transformation requires all X and Y values to be positive. The transformation of true X and Y values to the logs, basically results in incorrect A and B values when a linear solver is applied to the transformed data. The error comes from the fact that the variance between the data Y values and the computed Y values is not truly minimized.

TRENDLINE PERFORMANCE

Trendline is erratic, in that the values of the fitted parameters will not be consistent, day-to-day or week-to-to week for the identical input data set. The set of values you get today, may not be the set of values you get tomorrow. The shown equations with coefficient values in the chart plots may not show this variation, because Excel here only shows the first few digits of the fitted parameter value. This problem is shown and discussed below.

There will always be a numerical difference between the Trendline equation parameter values, and the corresponding Excel 2003 LINEST parameter values. LINEST 2003 is stable and uses a good reduction algorithm. Trendline is unstable, and the algorithm is unknown.

Trendline is a limited tool in that only a very limited amount of information is shown about the fitted equation on the chart itself. The main criteria is a visual one between the data points and the trendline curve. The only statistical measure that can be shown is an “R-SQUARED” value.

Trendline however is intended to project trends, so it may change internally from doing a regression to doing a projection in the form of a straight line. If only the projections beyond the data set are of interest, then the straight line projection may be acceptable. However any statistics about the variability of the projections still depends on a linear regression basis and on assumptions about the variances.

DO NOT USE TRENDLINE FOR A VALID REGRESSION OR A VALID PROJECTION ON DATA. THE EQUATION GIVEN IS MICROSOFT'S VERSION OF A TREND EQUATION, NOT NECESSARILY A CORRECT REGRESSION EQUATION. TRENDLINE MAY MIS-INTERPRET YOUR INTENTIONS AND ALSO GIVE WRONG PARAMETER VALUES.

SETTING UP THE CHART FOR A TRENDLINE

When the data has been selected and plotted, the Trendline option will appear in the chart (main) menu. Select the menu item “Add Trendline”.

Excel allows you to put in trendlines on columnar data (use of a XY Line Chart). However each column has a uniformly spaced x value number count assigned to it, independent of any labels indicating the true X value. See Peltier (2008) for an expanded discussion of this problem and a fix.

A new menu appears allowing selection of the type of curve, choice of smooth or straight line segments and the desired order. Linear, logarithmic, polynomial, power, exponential or moving average plots can be selected. A box appears (“Based on Series:”) that lists each of the plots by the series name set when the data for the plot was given a name and the worksheet X and Y data ranges set.

Select the type of function:

Linear	Smooth Lines	Select Order
Logarithmic	Smooth Lines	Select Order ³
Polynomial	Smooth Lines	Select Order
Power	Smooth Lines	
Exponential	Smooth Lines	Select Period
Moving Average		Select Period

Select the specific chart data set:

Select “Based on Series :” (the chart data set name)

Set or select the options:

Trendline Name	(Automatic or Custom)
Forecast,	Forward (# of units) or Backward (# of units)
Set Intercept	(select)
Display equation on chart	(select)
Display R Squared Value On Chart	(select)

The trendline curve and the data curve then appears on the chart (See Figure 11-7 below). The parameter (coefficient) numbers for the fitted equation are given in a fixed format form and arranged above or to the right of the equation line.

The formats can be changed, by first selecting the equation (a box surrounding the equation appears). Then go to the pop-up menu, select **format** and the standard cell format list appears. Select the form and number of digits to display. Then move the box

³ This may be misleading. What is done is convert the data to natural logs and then do a linear regression on the logs. Then the resulting coefficients are transformed back to correspond to the original data. The Help charts only show order 1 forms.

to the right and exit the box. The process here is also described below under the “THE DISPLAY” heading.

Trendline evaluates the selected “type” with respect to the data. Consequently Trendline may block certain combinations. For example, a data set with Y fluctuating may limit the polynomial to a linear trend fit.

Because coefficient values are not accurately shown (or computed), the shown equation with the initially shown coefficient values will calculate Y values that may not even be close to the data. Peltier (Peltier 2008) shows in an example how far the trend equation line can be from the data. Hargreaves (Hargreaves 2008) also complains about how far the computed equation is from the input data set. It’s not easy to say it here, but the fault is squarely due to the user who assumes 1 digit coefficient values are correct coefficient values.

The choice of whether the equation goes through the origin or not is determined by the user, checking the appropriate box in the set up process described above. If log transforms are involved, the shown line-through-the-origin may be oddly misplaced from the data.

The internal logic that shifts the regression equation to a lower linear form is not known or disclosed by Microsoft.

One of the issues that come up in arguments about TREND is how to interpret the terms “projection”, “regression”, “fit” or “trend”, since they do not mean the same to everybody.

REMOVING A TRENDLINE FROM A CHART

When charts become cluttered with plots and trendlines there is a problem.

Deleting trendline equations from a chart shuts Excel down. Trendline generates both the equation and the equation plot. First shown is the equation, then the plot as an overlay. Deleting the equation creates a lost connection and Excel shuts down. (UK Forums-3492345)

To remove a trendline:

- Click Trendline

- Select Layout (tab)

- Click Trendline

- From menu, select “none”

TRENDLINE FAULTS AND ERRORS

CHART AND EQUATION SELECTION FAULTS

Peltier () in his many replies to questions, points out that a major error made by users is to select the wrong chart option. The correct option is of course is to select the XY Scatter Plot (shown in the selection menu as a bunch of points in a miniature XY first quadrant figure).

When users select the XY line plot (two broken lines crossing in a miniature XY first quadrant figure), they will get incorrect equation coefficients with Trendline. This is because Excel interprets each X value, not as selected from the worksheet, but as a sequential bar-chart with a height equal to the given Y value. The result is a “new, unexpected” X value, based on sequential intervals (1, 2, 3, etc.). Trendline fits the input Y values to an “X” value equal to an interpretation of the actual X values based on a sequential interval value. Consequently the resulting Trendline equation (when plotted) will not even be close to the data points.

ACCURACY OF THE DISPLAYED EQUATION COEFFICIENTS
THE TEST FOR COEFFICIENT ACCURACIES

The basic test used here is the NIST Wampler set.

This test has 5 data sets to evaluate the ability of the software to properly fit a polynomial to X-Y data. The equation to be fitted is:

$$Y = B0 + B1 * X + B2 * X^2 + B3 * X^3 + B4 * X^4 + B5 * X^5$$

In 4 cases, the B values are equal to 1.0 exactly. In the fifth case (set 2) the coefficients are exactly 1.0, 0.1, 0.01, 0.001, 0.0001 and 0.00001. For each of 4 cases, the Y values have patterns of offset additives. There are 21 Y values for each set, and the additives are adjusted with increasing variations. The data sets are all integers, so it is a thorough test of algorithms using double precision computations.

THE ANALYSIS OF TEST RESULTS

Trendline charts that are shown below in the analysis section are the equations with the default equation constants. Below each chart is a table giving the LINEST output for the same data set, as a comparison.

For reference purposes, the cells in the LINEST output are:

Table 11-2: Cells In The LINEST Output

Column 1	Column 2	Columns 3 to n
Coefficient n	Coefficient n-1	Next Coefficient
Standard Error of Coefficient n	Standard Error of Coefficient n-1	Standard Error of Next Coefficient
Coefficient of Determination, R squared	Standard Error of the Y estimate	
The F Statistic	Degrees of Freedom, d1	
Regression Sum of Squares	Residual sum-of-squares	

The intercept is always the rightmost column. The highest polynomial term is in column 1.

For the t test on coefficient values, test $t = \text{coefficient } n / \text{standard error of coefficient } n$, with $df = d1$.

For the F test, $F = \frac{\text{The F statistic}}{\text{deg 1} = 21 - (d1 - 1) \text{ or } 20 - d1, \text{ deg 2} = 21}$

TRENDLINE EQUATION PARAMETER VALUE INSTABILITIES

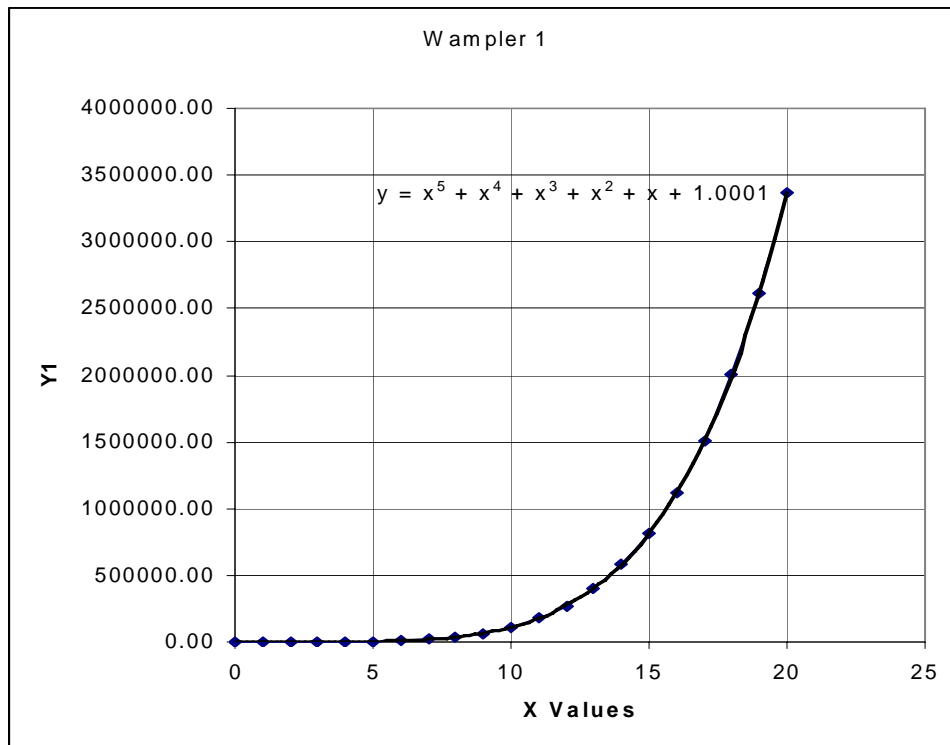
There is a major problem here, some kind of instabilities in the Trendline computation of equation parameter values. The fitted parameter value set will change from day to day when a new chart is formed, using the same data set and the same settings. I have not been able to characterize this instability.

This instability affects the testing for accuracy, and therefore it is hard to generalize on how accurate the Trendline equation coefficients really are. It is also difficult to judge if Trendline actually passes the NIST tests.

In run 1 on Wampler sets 4 and 5, the change from an expected fifth order equation to a straight line (that Microsoft describes as a better trend) occurred automatically. In run 2 (as a test check) the fifth order polynomial was retained for all five sets. The coefficient values for these sets were exactly the same for sets 1, 3, 4 and 5, which seemed odd. The values were different from run 1. The results of run 1 could not be repeated.

WAMPLER SET 1

Figure 11-7: Wampler 1 Output Chart With Trendline, Run 2



Succeeding runs on the Wampler 1 data set gave identical chart appearances. However if the equation is examined, one finds different coefficient values, that when rounded all come out to 1 for the coefficients and 1.0001 or 1 for the intercept. Table 11-4 shows some of the differences.

Table 11-3: LINEST Output For Wampler 1

1.0000000000E+00	1.0000000000E+00	1.0000000000E+00	9.999999997E-01	1.000000001E+00	9.999999988E-01
1.7766796933E-15	8.9299368384E-14	1.6050719508E-12	1.2327136480E-11	3.7385086168E-11	3.4044062170E-11
1.0000000000E+00	3.7331201085E-11	#N/A	#N/A	#N/A	#N/A
2.7000669412E+33	1.5000000000E+01	#N/A	#N/A	#N/A	#N/A
1.8814317208E+13	2.0904278617E-20	#N/A	#N/A	#N/A	#N/A

Table 11-4: Wampler 1 Trendline Coefficient Set Values

Term	Run 1 Coefficient Values	Run 2 Coefficient Values	Run of Dec 18 th , Thru the origin ⁴
X^5	0.99999999961887	1.0000000000000000	1.0000000000000000
X^4	1.00000000194466	1.0000000046566	1.0000000046566
X^3	0.999999964876784	0.999999932944774	0.999999985098838
X^2	1.00000026446872	1.00000184774398	1.00000053644180
X	0.999999270959636	0.999979496002197	0.999996662139892
Intercept	1.00000036798883	1.00005340576171	

Table 11-5: LRE Values For the Coefficients in Tables 11-3 and 11-4

Term	Trendline (Run 1)	Trendline (Run 2)	LINEST	Trendline (thru-the-origin)
X^5	10.42	16.00	14.65	16
X^4	8.71	9.33	12.87	9.33
X^3	7.45	7.17	11.52	7.83
X^2	6.58	5.73	10.56	6.27
X	6.14	4.69	10.02	5.47
Intercept	6.43	4.27	9.33	

A strange occurrence with run 2 was that the coefficient values in the middle column of table 11-4 were fully identical (digit to digit) to the Wampler 3, 4 and 5 set outcomes.

Reruns again come up with different coefficient and intercept values. Table 11-4 is just a sample. We have to conclude that with the variability and the possibility that even lower LRE values than run 2 would occur in succeeding tests, that Trendline failed the NIST Wampler 1 test.

⁴ For Wampler 1, the through-the-origin set had one subtracted from all the Y values

WAMPLER 2

Figure 11-8: Wampler 2 Output Chart With Trendline, Run 1

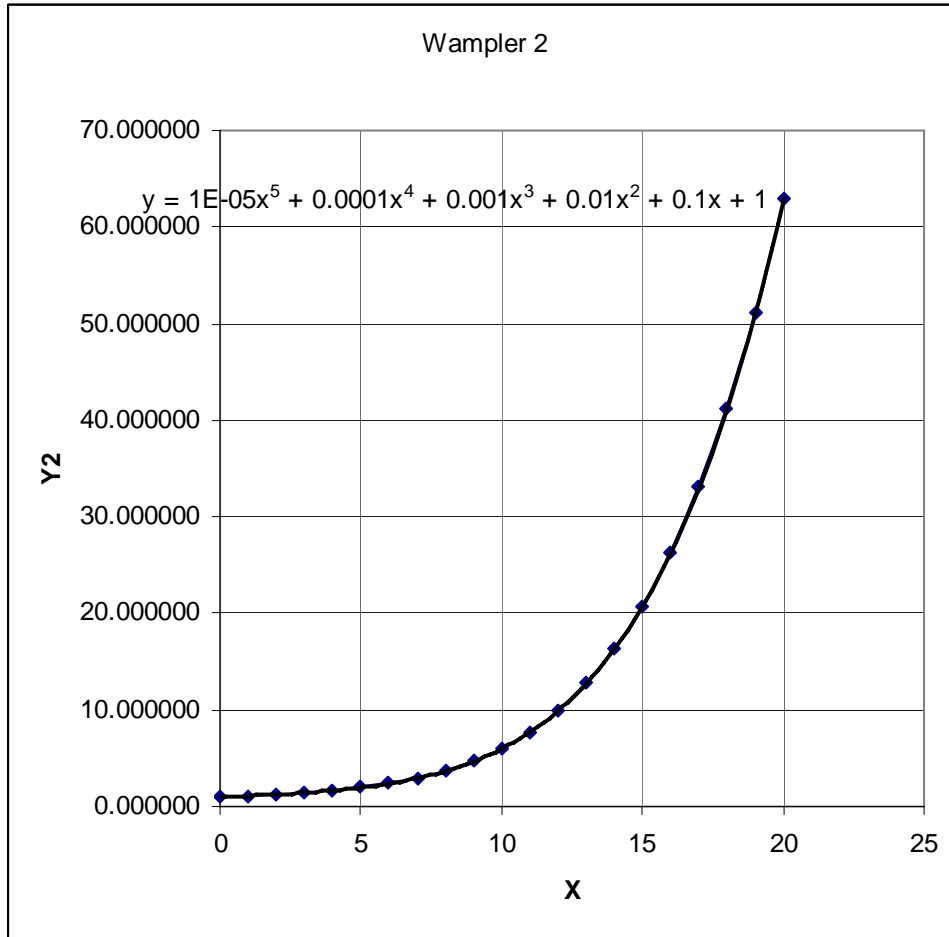


Table 11-6: LINEST Output For Wampler 2

1.0000000000E-05	1.0000000000E-04	1.0000000000E-03	1.0000000000E-02	1.0000000000E-01	1.0000000000E+00
4.7182452138E-20	2.3714815848E-18	4.2625145538E-17	3.2736600142E-16	9.9281825840E-16	9.0409224578E-16
1.0000000000E+00	9.9138725745E-16	#N/A	#N/A	#N/A	#N/A
1.3436287035E+33	1.5000000000E+01	#N/A	#N/A	#N/A	#N/A
6.6029185837E+03	1.4742730414E-29	#N/A	#N/A	#N/A	#N/A

Table 11-7: LRE Values For the Coefficients, Wampler 2

Term	Trendline (fig 11-2)	LINEST (table 11-6)
X^5	10.08	>11
X^4	9.28	>11
X^3	8.86	>11
X^2	8.40	>11
X	8.76	>11
Intercept	8.95	>11

In this particular case, the trendline equation is the same as the regression equation. Trendline passed Wampler 2. Given the magnitude of the power terms with respect to the Y values, I would not expect in repeat runs that the LRE values for the coefficients and intercept would be less than 5.

WAMPLER 3

Figure 11-9: Wampler 3 Output Chart With Trendline, Run 1

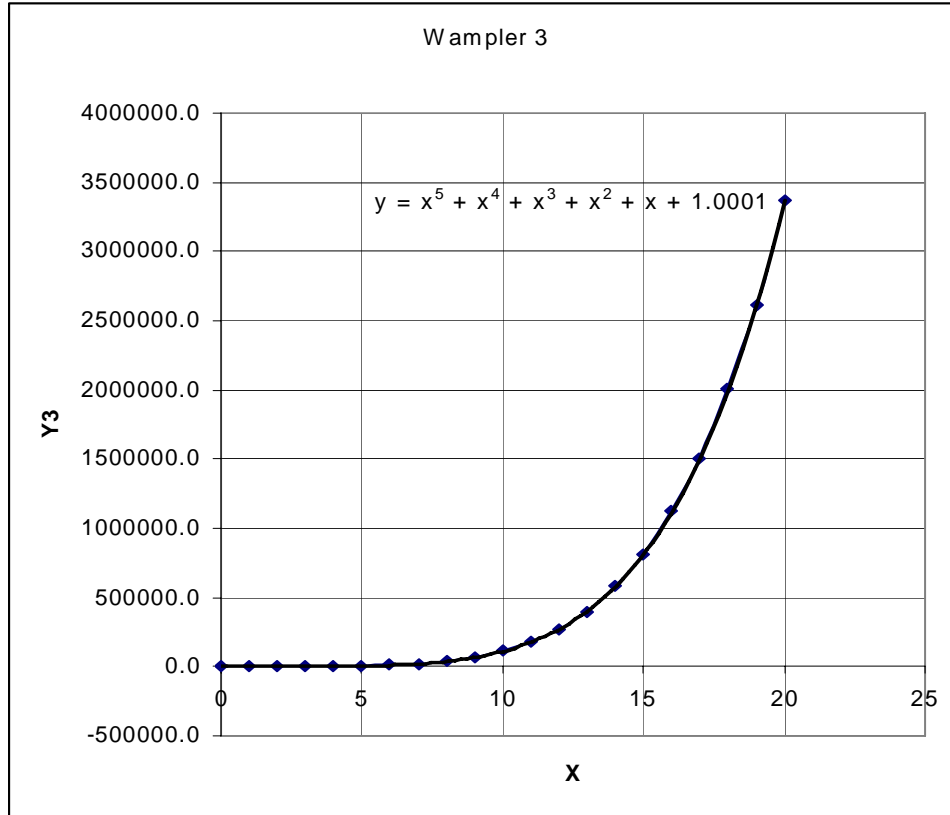


Table 11-8: LINEST Output For Wampler 3

1.0000000000E+00	1.0000000000E+00	9.9999999999E-01	1.0000000000E+00	9.9999999991E-01	1.0000000000E+00
1.1232485468E-01	5.6456651217E+00	1.0147550755E+02	7.7934352433E+02	2.3635517347E+03	2.1523262468E+03
9.9999555903E-01	2.3601450238E+03	#N/A	#N/A	#N/A	#N/A
6.7552445824E+05	1.5000000000E+01	#N/A	#N/A	#N/A	#N/A
1.8814317208E+13	8.3554268000E+07	#N/A	#N/A	#N/A	#N/A

Table 11-9: LRE Values For the Coefficients, Wampler 3, Run 1

Term	Trendline (fig 11-3)	LINEST (table 11-8)
X^5	9.60	>11
X^4	9.33	>11
X^3	7.17	>11
X^2	5.73	>11
X	4.69	>11
Intercept	4.27	>11

Again the 1.0001 intercept shows up, when the LINEST value is exactly one. Trendline fails the Wampler 3 test on the basis of table 11-9.

WAMPLER 4

Figure 11-10: Wampler 4 Output Chart With Trendline, Run 1

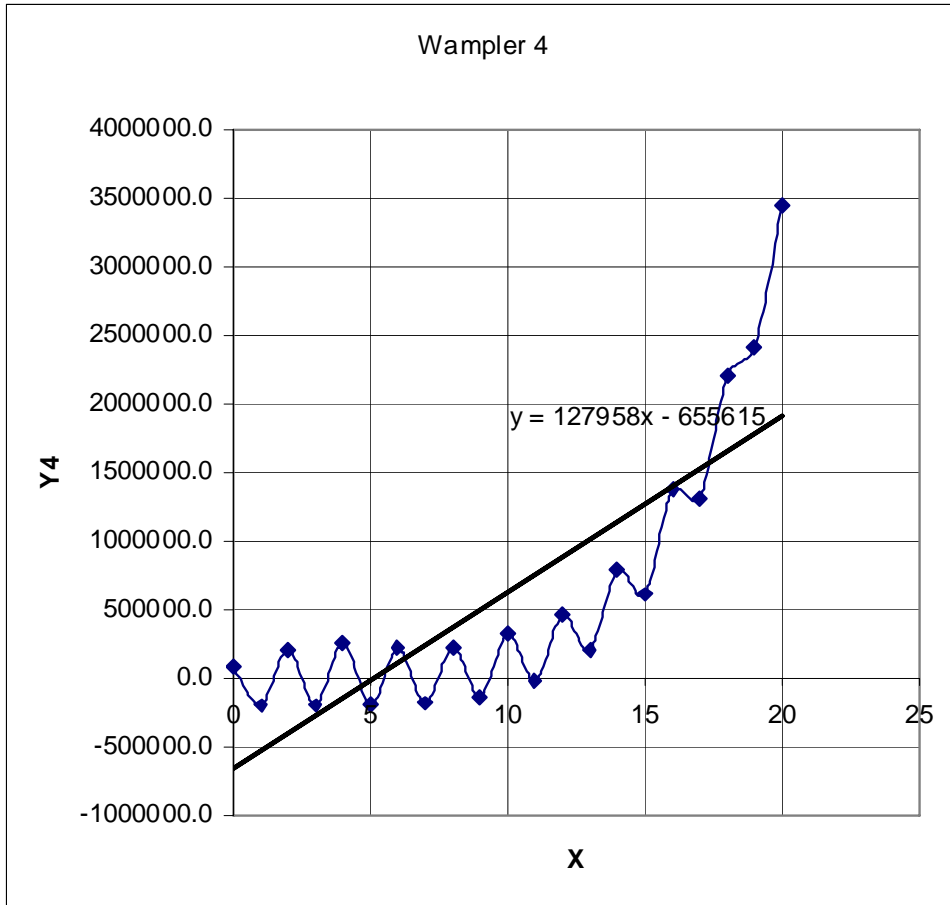


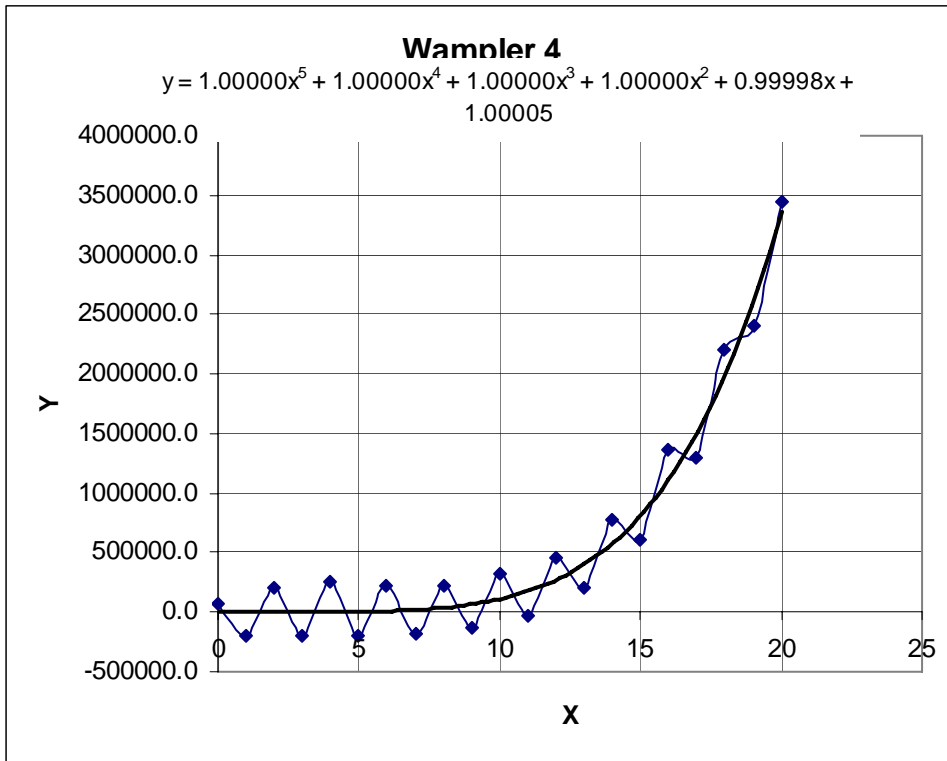
Table 11-10: LINEST Output For Wampler 4

1.0000000000E+00	1.0000000000E+00	9.9999999965E-01	1.0000000027E+00	9.9999999264E-01	1.0000000036E+00
1.1232485468E+01	5.6456651217E+02	1.0147550755E+04	7.7934352433E+04	2.3635517347E+05	2.1523262468E+05
9.5747844083E-01	2.3601450238E+05	#N/A	#N/A	#N/A	#N/A
6.7552445824E+01	1.5000000000E+01	#N/A	#N/A	#N/A	#N/A
1.8814317208E+13	8.3554268000E+11	#N/A	#N/A	#N/A	#N/A

Here we have a major departure from the regression, where Trendline has dropped to the straight line projection mode. The data, even with the added “noise” of Wampler 4, still shows the characteristic upward “trend”, which is lost in the straight line output.

Several repeats of the test failed to reproduce figure 11-10. All the repeats were as shown in Figure 11-11.

Figure 11-11; Wampler 4 Output Chart With Trendline, Additional Runs

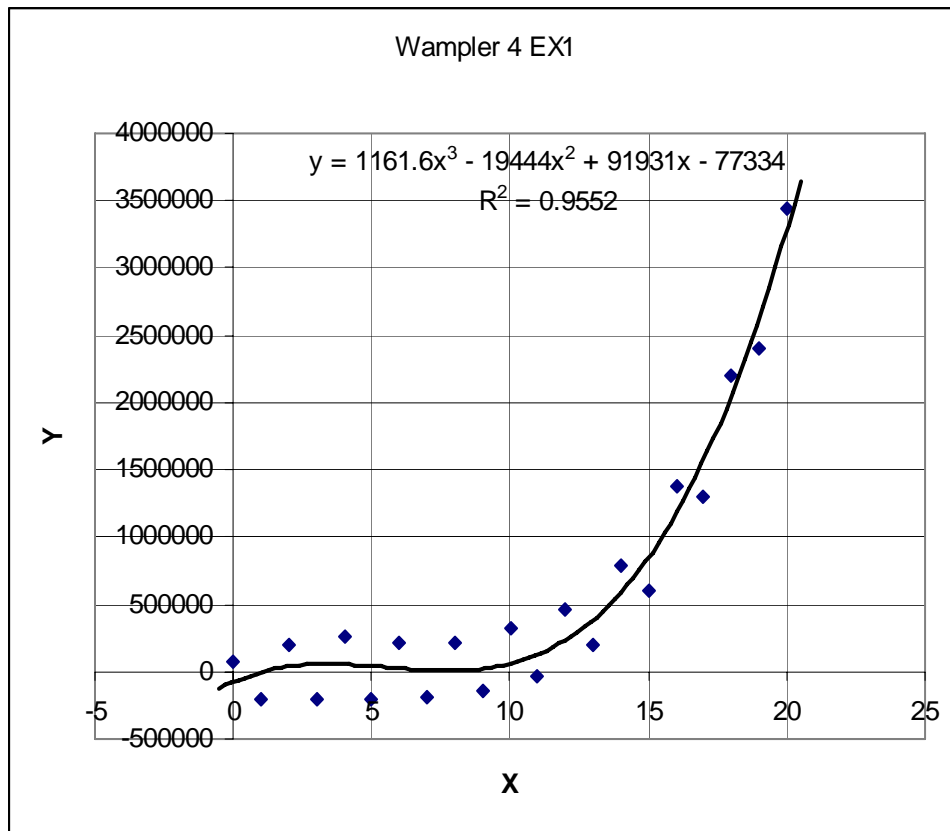


Succeeding runs generally had coefficient values about the same as shown in figure 11-11.

If the form of figure 11-10 appears, then we can say that Trendline fails the Wampler 4 test. If however the next tester gets repeats of the form of figure 11-11, he would conclude that Trendline passed Wampler 4.

A straight line projection of the form of figure 11-10 is a misleading trend. It fails to recognize the increasing rise of Y values beyond 15. A better trend would be to reduce the polynomial to 3 as shown in figure 11-12.

Figure 11-12: Wampler 4 Output Chart With A Cubic Trendline



The trendline did not drop to the linear form shown in figure 11-10, but retained the upward projection of the left area points. Consequently by intervention, we have found a better “trend”.

WAMPLER 5

Figure 11-13: Wampler 5 Output Chart With Trendline, Run 1

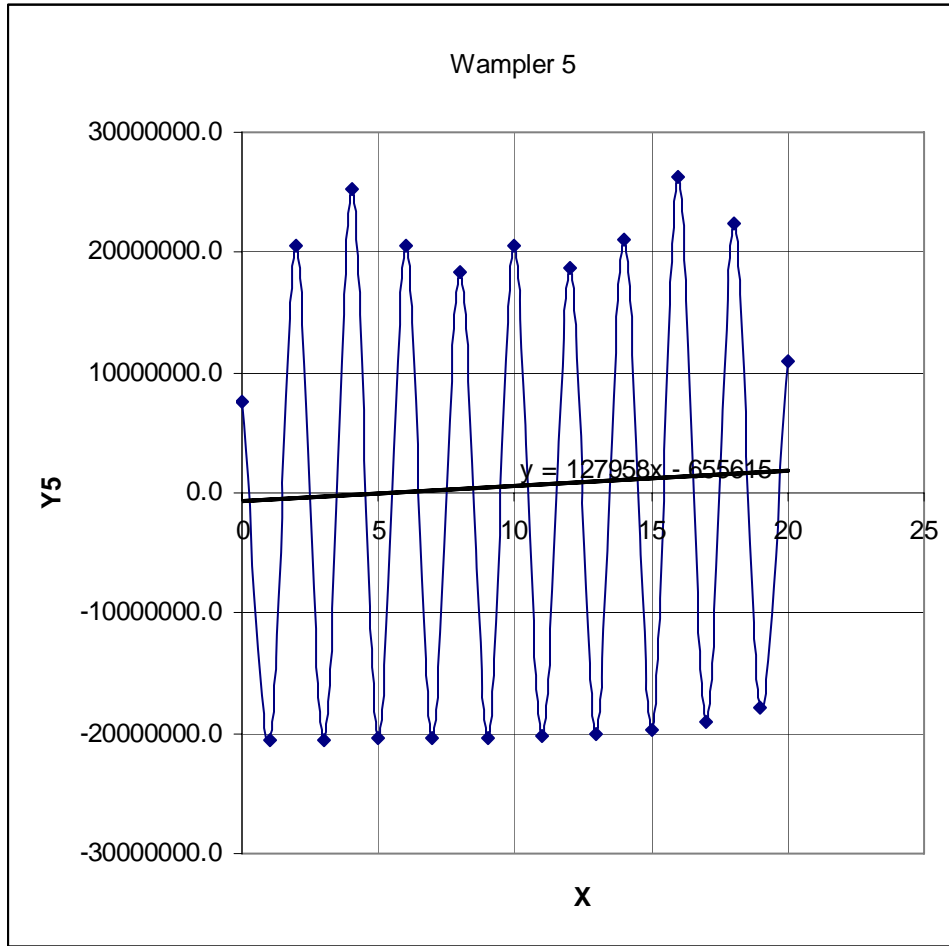


Table 11-11 LINEST Output For Wampler 5

9.9999999996E-01	1.0000000019E+00	9.9999996488E-01	1.0000002645E+00	9.9999927096E-01	1.0000003680E+00
1.1232485468E+03	5.6456651217E+04	1.0147550755E+06	7.7934352433E+06	2.3635517347E+07	2.1523262468E+07
2.2466892157E-03	2.3601450238E+07	#N/A	#N/A	#N/A	#N/A
6.7552445824E-03	1.5000000000E+01	#N/A	#N/A	#N/A	#N/A
1.8814317208E+13	8.3554268000E+15	#N/A	#N/A	#N/A	#N/A

Here again we have a major departure from the regression in figure 11-13, where Trendline dropped to the straight line projection mode. However this was not repeatable. All repetitions were like figure 11-14.

The resulting equation shown on the chart may be a reasonable projection (left and right) of the data, given the “scatter” in the data. It is not a valid regression. A projection is not a regression, but it may be computed Y values for new X values that are below or above the given data set.

Figure 11-14: Wampler 5 Output Chart With Trendline, Succeeding Runs

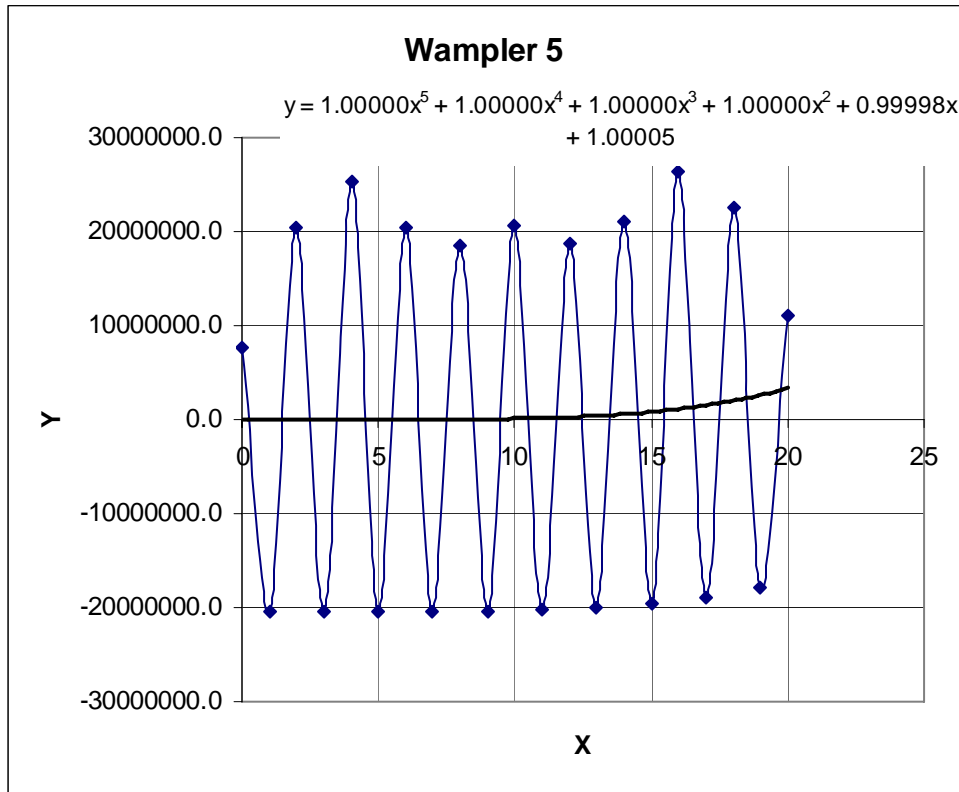


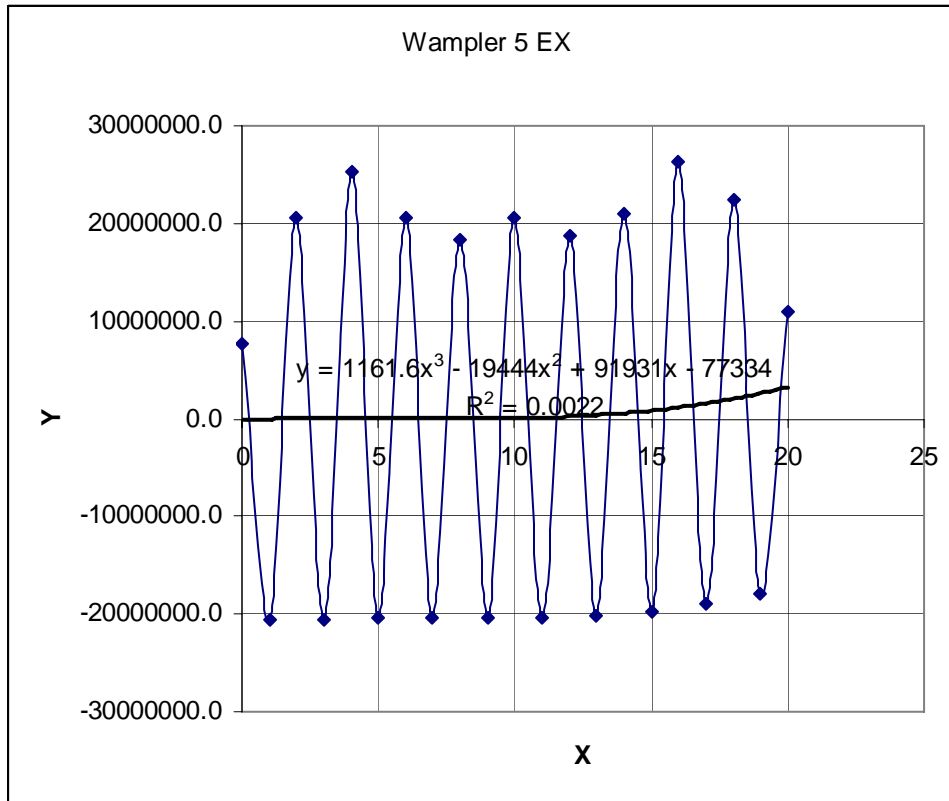
Table 11-12: Coefficient Values, Wampler 5, Additional Run

Term	Coefficient Values	LRE Values	LINEST LRE Values
X^5	1.000000000000000	16.00	10.40
X^4	1.00000000046566	9.33	8.72
X^3	0.999999932944774	7.17	7.45
X^2	1.00000184774398	5.73	6.58
X	0.999979496002197	4.69	6.13
Intercept	1.00005340576171	4.27	6.43

In general terms we can say that the Trendline calculation on the Wampler 5 data tends to do poorly on the Wampler data, resulting in the lowest LRE value of 4. LINEST does not do much better, but gives us the lowest LRE value of about 6.

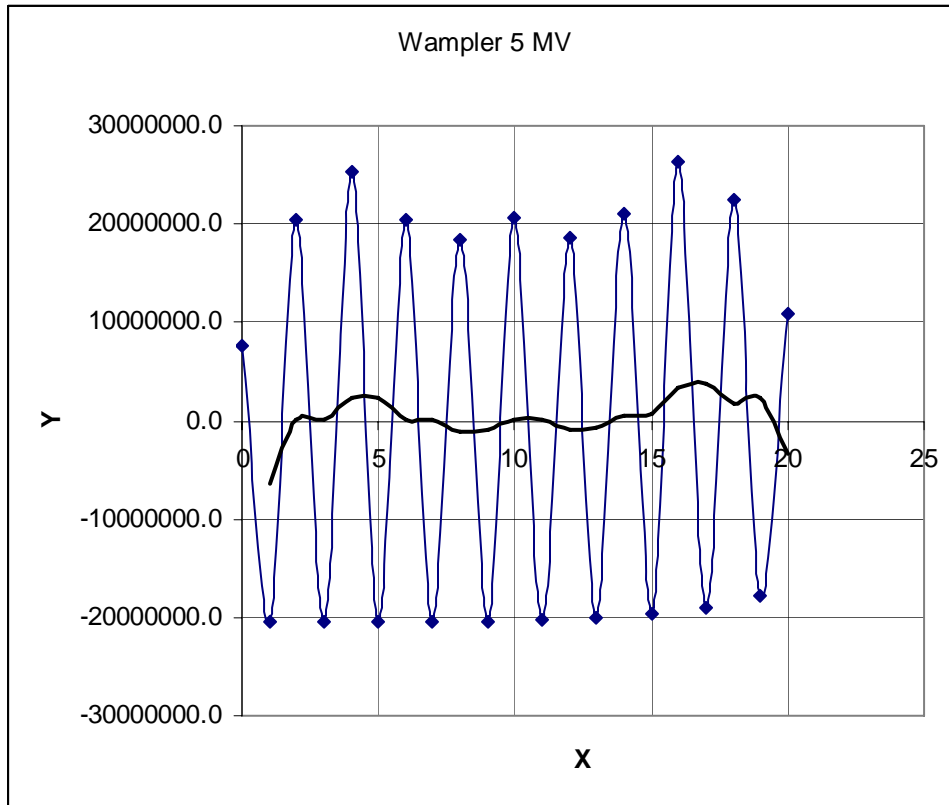
If a cubic is specified as the trendline we get figure 11-15, which retains an upward trend in spite of the very large noise component. However we cannot adequately judge the fit or equation parameter values from only the R squared value.

Figure 11-15: Wampler 5 Output Chart With A Cubic Trendline



If we try a moving average with a two period average, we get figure 11-16.

Figure 11-16: A Moving Average Trendline chart



A moving average chart here is a somewhat better view of trends, since it points out the major influences that large deviations from a central “mean” can have on any quantitative prediction of a trend. The inherent lag of the moving average does not make it a very good method of projection.

TRENDLINE EXPONENTIAL FUNCTION

NIST NELSON DATA SET

The NIST Nelson data set and equation with parameter values to be fitted is essentially an exponential fit problem. The equation however is of the form:

$$\log[y] = b_1 - b_2 * x_1 * \exp[-b_3 * x_2] + e$$

By rearranging we can calculate a new Y value (W), since the true value of b_1 is known.

$$W = -(\log(y) - b_1) / x_1$$

$$W = b_2 e^{-(b_3 * x_2)}$$

Which is now a true exponential that can be fitted by the Exponential Trendline.

However, the reduced NELSON data in this form ends up with large clusters of Y values at the X values of 180, 225, 250 and 275, and a simple curve can't be drawn.

NIST MISRLA DATA SET

The equation is of the form:

$$y = b_1 * (1 - \exp[-b_2 * x]) + e$$

The MISRLA set deviates slightly from a simple exponential. If we change the Y to (b1 – Y), the resulting modification forms an exponential function on the right side. The modified input then is:

Table 11-13: Modified MISRLA Data Set

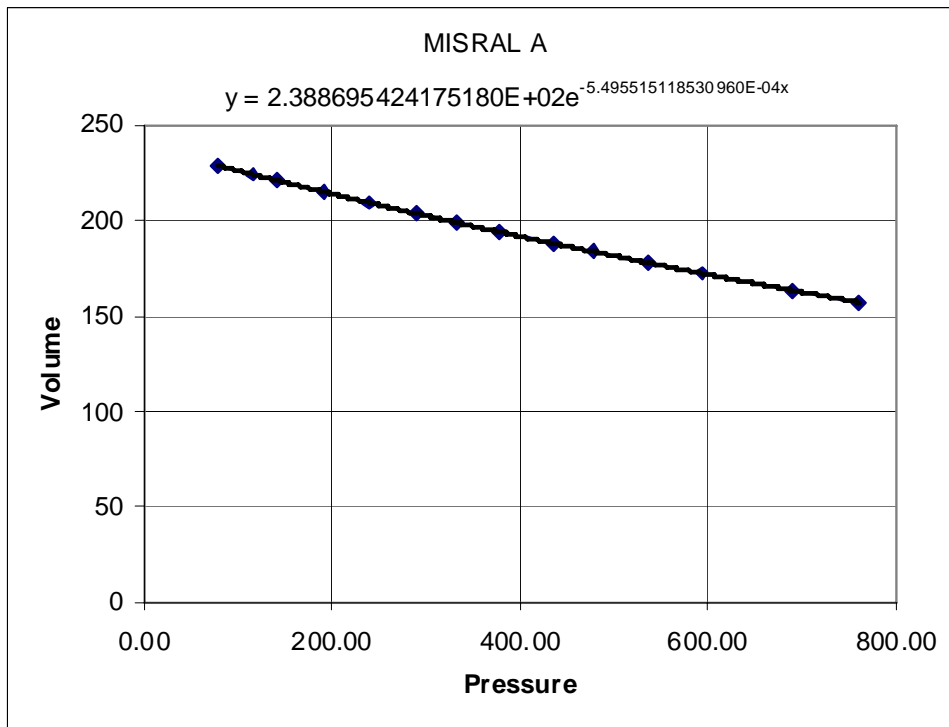
Y	X
228.8721	77.60
224.2121	114.90
221.0021	141.10
215.0121	190.80
209.3321	239.90
203.7621	289.00
198.9221	332.80
194.1221	378.40
188.1821	434.80
183.8921	477.30
177.9321	536.80
172.5421	593.10
163.4721	689.10
157.1621	760.00

And the modified function is

$$Y' = b1 * \exp(-b2 * X)$$

The corresponding data is plotted in figure 11-16, along with the Trendline equation.

Figure 11-17: NIST MISRAL A Data Set



If we compare the Trendline coefficients with the NIST baseline values:

- a. We have a LRE value of 3.52 for the coefficient and 2.96 for the exponent.
- b. If we correct for the small change in the NIST parameter values as a result of the function modification, the LRE values for the coefficient is 3.93 and 3.21 for the exponent.

Since the Trendline model is in error, the LRE values indicate the extent of the Excel error.

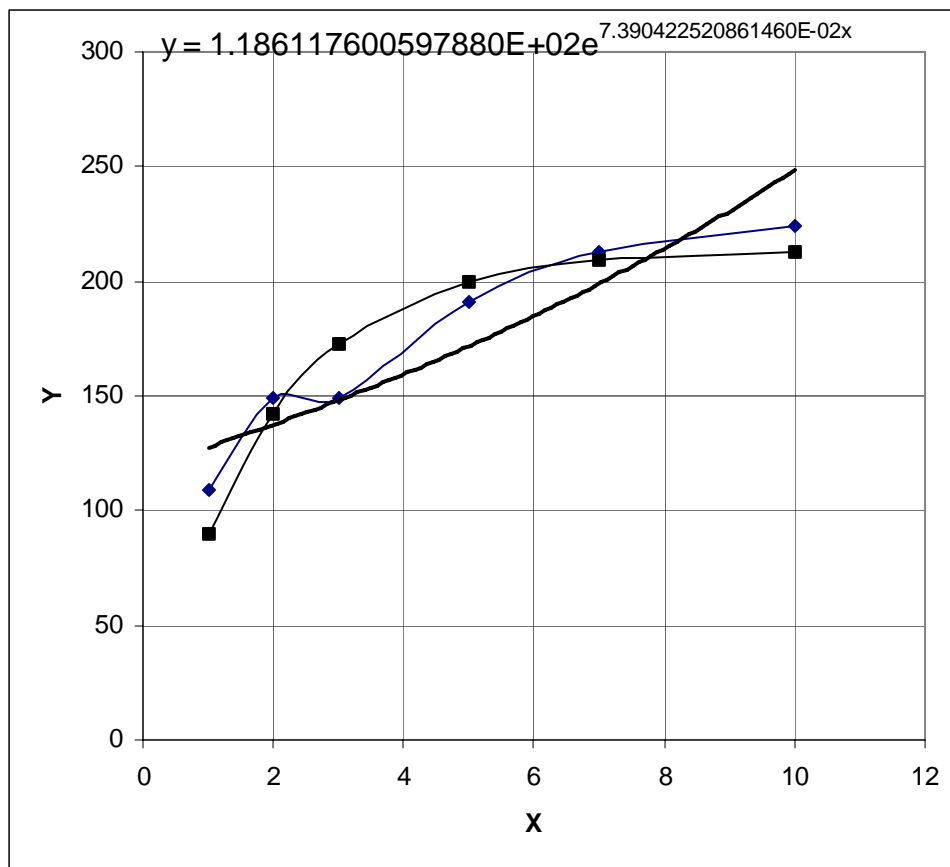
NIST BOXBOD DATA SET

The equation is:

$$y = b1*(1-\exp[-b2*x]) + e$$

The source data is uneven, and fits to any exponential class are poor. Figure 11-18 shows the data points as diamonds, the NIST fit to the above equation as squares, and the Trendline equation (as shown on the figure) as the solid black line. The Trendline is fitted to the diamond points.

Figure 11-18: NIST BoxBOD Data Set



The NIST coefficients are $b1 = 213.80940889$ and $b2 = 0.54723748542$. The first Trendline coefficient shows that Trendline is in the “ballpark, but both the sign and the value of the exponential coefficient are not in the same “ballpark”.

The source data is too irregular to try and modify the data so that the Trendline equation would be applicable.

EXCEL 2003-EXCEL 2007 EQUATION DISPLAY FAULT

In a recent Email (Hargreaves 2008), he cites the problem. “It (i.e. the trendline equation in Excel 2007) reports wildly incorrect values for polynomial trend-line parameters compared to Excel 2003 and PSI-PLOTv8.5.

The base here is an XY data set of 513 points. The following charts are plots of the data with the Trendline (2003) calculated model shown to 15 decimals.

Figure 11-19: Hargreaves Data Set, With a “Through-the-Origin” Model.

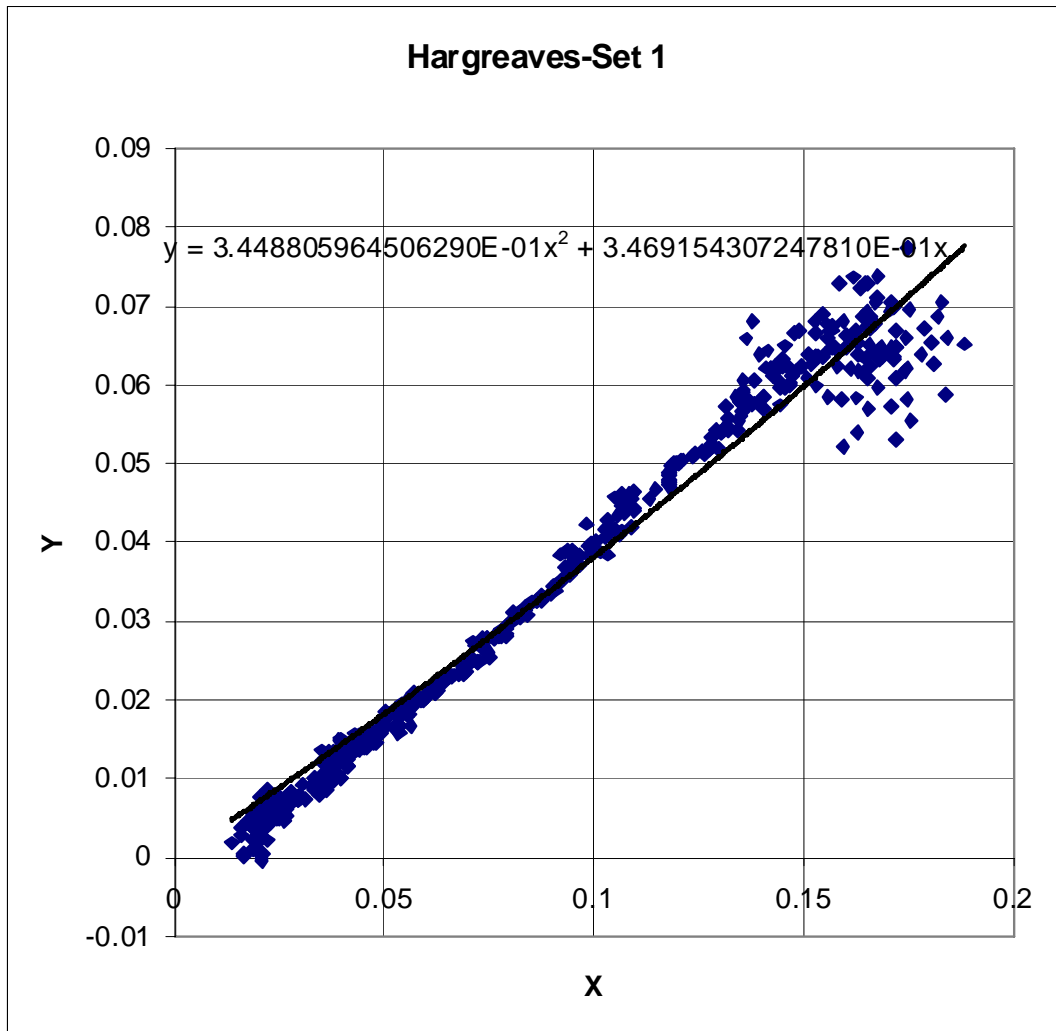
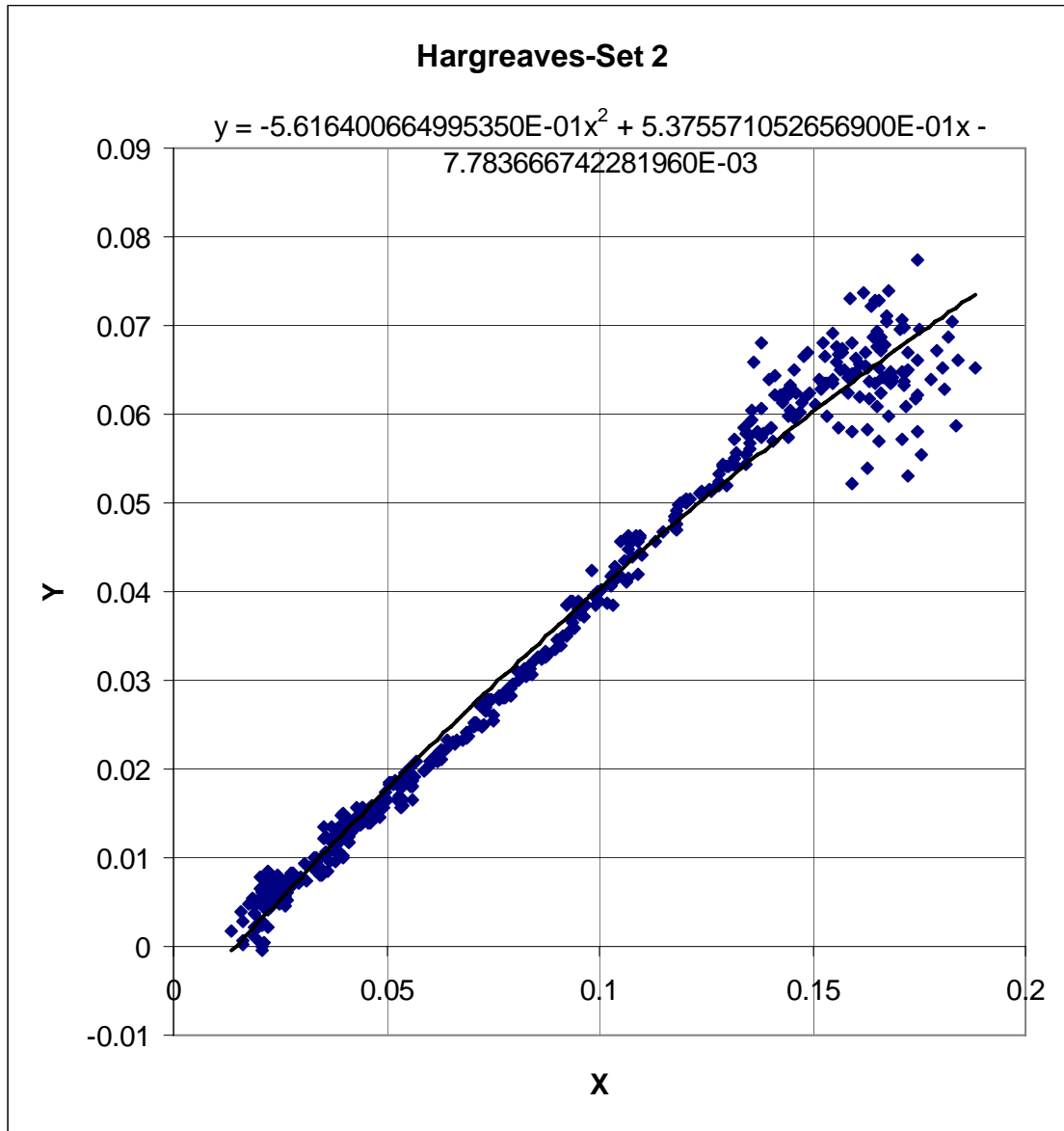


Figure 11-20: Hargreaves Data Set, With an “Intercept” Model.



In this case the intercept is small, but the linear and quadratic coefficients are considerably different from that of figure 11-19. Once the intercept constraint is removed, the apparent offset of the data below an x value of 0.1 and the cluster of points above an x value of 0.15, have an important effect.

The issue here is not about the differences between figures 11-19 and 11-20, but is an issue on how Excel 2007 handles this problem.

As Hargreaves (2008) stated :

“Office 2007 (with *.xls or *.xlsx), the Polynomial equation is improperly displayed (wrong parameters) displayed when forced through zero (trendline option). After saving and reopening it loses highest order term when file is reopened if equation is forced through zero.”

Table 11-14: The Basic Problem (from Hargreaves 2008)

	This xls file via Excel 2007	This xls file via Excel 2003
After file is opened:	$y = 0.3469x$	$y = 0.3449x^2 + 0.3469x$
	$R^2 = 0.973$	$R2 = 0.973$
After removing "force through zero"	$y = -0.5616x^2 + 0.5376x - 0.0078$	$y = -0.5616x^2 + 0.5376x - 0.0078$
	$R^2 = 0.982$	$R2 = 0.982$
After reestablishing "force through zero"	$y = -0.5616x^2 + 0.3469x$	$y = 0.3449x^2 + 0.3469x$
	$R^2 = 0.973$	$R2 = 0.973$

Column 3 of table 11-14 matches the Trendline equations shown in figures 11-19 and 11-20.

(NOTE: I was not able to get my Excel 2007 to work here at this time. There is a "loading" fault that blocks the proper removal of Excel 2003 and loading Excel 2007 options in my Windows XP. Therefore I could not repeat his findings on Excel 2007. DAH)

CONCLUSIONS

Trendline passes the Wampler 2 fit test. Trendline marginally fails the Wampler 1, 3, 4 and 5 fit tests.

A definite failure would be LRE values consistently less than 3. In the case of the Wampler data, minimum LRE values were in the 4-5 range, which may be acceptable to the user, but not from a statistics viewpoint. Even LINEST values drop to the 6 LRE level, indicating that the Wampler test is indeed a hard test.

Excel 2007 Trendline has an interface problem with Excel 2003 data, resulting in incorrect equation parameter values. This is a fault in Trendline.

THE BOTTOM LINE

THE BOTTOM LINE IS THAT YOU CANNOT RELY JUST ON THE R SQUARED VALUE TO JUDGE FIT. THE COEFFICIENT T VALUES ARE NOT SHOWN IN TRENDLINE.

YOU CANNOT PROPERLY JUDGE A FIT USING ONLY TRENDLINE AND DOING A VISUAL COMPARISON OF THE TRENDLINE CURVE TO THE DATA POINTS.

THE CENTRAL ISSUE IS ABOUT THE DEVIATIONS OF THE DATA FROM THE MODEL, AS TO WHETHER THE DEVIATIONS ARE TRULY RANDOM (NOISE) OR DUE TO EQUATION MISFIT.

YOU CANNOT RELY ON THE TRENDLINE OR THE COEFFICIENT VALUES TO GIVE YOU A VALID REGRESSION.

THAT BEING A CORRECT MODEL HAS NO BEARING ON R SQUARED VALUES. TRULY RANDOM NOISE MAY GIVE LOW R SQUARED VALUES, BUT HIGH COEFFICIENT T VALUES.