Spreadsheets in Education (eJSiE)

Volume 1 | Issue 1 Article 5

July 2003

Spreadsheets, Graphing Calculators and the Line of Best Fit

Bernie O'Sullivan
St. Luke's Anglican School

Spreadsheets, Graphing Calculators and the Line of Best Fit

Abstract
The advent of the hand-held graphing calculator has produced its own mini revolution within the Mathematics classroom. The ability for students to now perform complex mathematical procedures quickly and accurately has pushed back the boundaries of content coverage that is possible within senior Mathematics courses in much the same way as the introduction of the scientific calculator did in the late 70's. There has been much quality debate and research into the best way to incorporate this new technology, and what balance should be reached between "by hand" methods and hand-held technology. For a discussion on the incorporation of computer algebra systems (CAS), see Flynn et al [1].

One technique that can now be done, almost mindlessly, is the line of best fit. Both the graphing calculator and the Excel spreadsheet produce models for collected data that appear to be very good fits, but upon closer scrutiny, are revealed to be quite poor. This article will examine one such case. I will couch the paper within the framework of a very good classroom investigation that will help generate students' understanding of the basic principles of curve fitting and will enable them to produce a very accurate model of collected data by combining the technology of the graphing calculator and the spreadsheet.

Keywords
spreadsheets, graphs, line of best fit, calculators

Spreadsheets, Graphing Calculators and the Line of Best Fit

Bernie O'Sullivan
St Luke's Anglican School

July 23, 2003

Abstract

The advent of the hand-held graphing calculator has produced its own mini revolution within the Mathematics classroom. The ability for students to now perform complex mathematical procedures quickly and accurately has pushed back the boundaries of content coverage that is possible within senior Mathematics courses in much the same way as the introduction of the scientific calculator did in the late 70's. There has been much quality debate and research into the best way to incorporate this new technology, and what balance should be reached between "by hand" methods and hand-held technology. For a discussion on the incorporation of computer algebra systems (CAS), see Flynn *et al* [1].

One technique that can now be done, almost mindlessly, is the *line of best fit*. Both the graphing calculator and the Excel spreadsheet produce models for collected data that appear to be very good fits, but upon closer scrutiny, are revealed to be quite poor. This article will examine one such case. I will couch the paper within the framework of a very good classroom investigation that will help generate students' understanding of the basic principles of curve fitting and will enable them to produce a very accurate model of collected data by combining the technology of the graphing calculator and the spreadsheet.

Classroom Investigation

A cup of boiled water cooling to room temperature provides an excellent source of data that is exponential in nature. The Computer Based Laboratory (CBL) that can be attached to a Texas Instruments calculator will collect the "temperature versus time" data and store it to the calculator's lists. Readers can refer to the instructions in the appropriate manual for the necessary steps. (Other calculators will have their own equivalents).

Once the data has been collected, students can begin fitting curves to this data on their graphing calculator, checking the r and R^2 values. I won't spend any more time now on this side of the investigation as some of the things I will discuss regarding the spreadsheet can also be accomplished on the calculator.

Click here to obtain a copy of the data. This will enable you to perform the steps that are outlined in this article.

By use of the graph link software and cable, students can now transfer this data to their PC and save it in an appropriate file as a **text** document. To retrieve the file, it will need to be opened from within Excel. Follow the prompts and accept each of the three steps that appear on your screen. You now have your data before you.

Graph this data using Chart Wizard; choose XY (scatter), points only, with no lines (Figure 1). For better viewing, re-scale the vertical axis. Right-click the vertical axis and choose "Format …" and scale from 60 minimum to 90 maximum (Fig 2) Also rescale the size of the data markers to size 2. This is also achieved by a right-click, but on a data point.

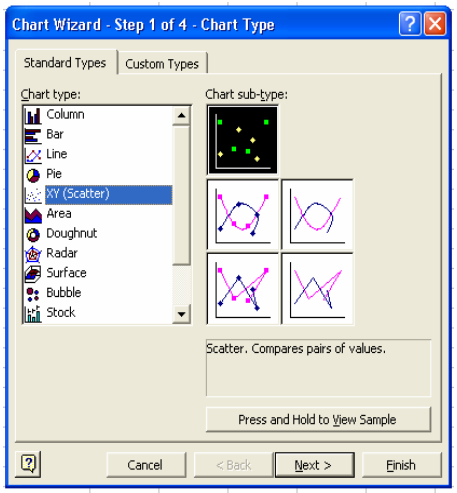

Figure 1: Selecting X-Y Scatter

A trendline is easily added if we right-click on one of the data points and choose "Add Trendline…". Choose to fit an exponential curve. Under the "Option" tab, choose to show the equation and the R-squared value. The computer yields an equation of $y = 86.442e^{-0.0003x}$ with an R^2 value of 0.9923.

Many high school students seek to use the R^2 value as a measure of "how well the model fits the data". Students may use an arbitrary figure of 70%. Any line with an R^2 greater than this is said to be a good model (the source of this 70% is unknown and puzzling)[1]. MacGillivray [2] provides two excellent contrasting examples of the danger involved with R^2. In the first, a very high R^2 is obtained from an inappropriate model while the second example produces a low R^2 value (19.8%) from an appropriate model.

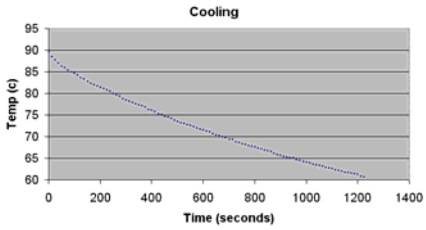

Figure 2: Graph of the Data

The very high R^2 value for my data would convince most students that this was in fact a good model of the data, but does R^2 tell the whole story? Your graph (Figure 3) shows us that the residuals (difference between the original data and the trendline) produce a distinctive pattern. This patterning in the residuals indicates that this model is not as good a fit as it first seemed. The residuals from a good fit

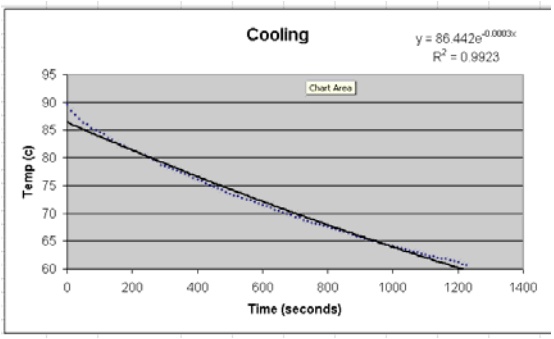

Figure 3: Line of Best Fit

[1] One of the forms for R2 is that R2 = 1 – RSS/SST, where RSS is the sum of the squares of the residuals and SST is the sum of the squares of the differences between the y-observations and their overall average. In this form, R2 is often interpreted colloquially as the "percentage of variation explained by the least squares regression line".

should be randomly scattered around zero when plotted against the x-variable (explanatory/independent variable).

This would be due to the assumption that the y-values (response/dependent variable) should be distributed normally around their expected value due to things such as experimental error[2] Figure 4 shows a chart of the residuals graphed against their x-value. This patterning indicates that the model does not accurately account for the variation in the data. For a model to be a good representation of the data, the sum of the deviations squared should be as small as possible.

Students should now check the sum-of-squared-deviations value for their model. If the time and temperature are in columns A and B, enter the equation generated by the computer into column C as follows; "=86.442*EXP(-0.0003*A1). The value "e" is represented in Excel by EXP(number), where the number in the brackets is the power "e" is being raised to. We can display our residuals squared in column D with the formula "= (B1-C1) ^2". Fill both of these formulae to the bottom[3].

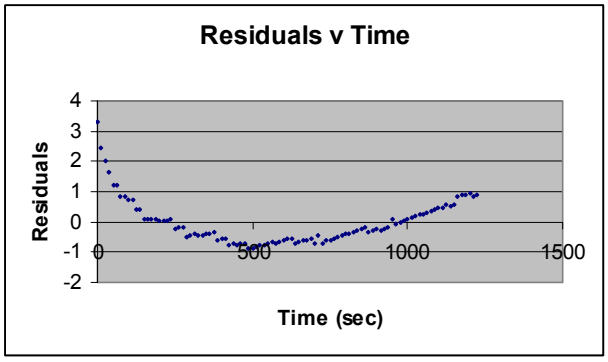

Figure 4: Residuals

Use the Auto Sum Wizard to calculate the sum of column D and place this value at the top of your page for easy reference when needed (Figure 5). My score of 53.27 is not too bad, but there is still a great deal of improvement that can be obtained (see the Appendix at the end of this paper).

	A	B	C	D	E	F	G
1							
2	Time (sec)	Temp	Computer trend Line	Dev Squared	Sum of Deviations Squared		53.27
3	0	89.73	86.442	10.81094			

Figure 5: Calculation of Deviation

[2] But experimental error would not be the only reason for deviation from the mean. If there is a deterministic relationship between the two variables, something like circumference = pi*diameter, then deviations would result from experimental error (such as having a figure that is not a perfect circle or measurement error). But in other situations, where there is not a deterministic relationship, the values of the y-variable are determined by the model and random (individual) deviation. For example, we can find a linear model that relates the height of adult sons with the height of the father. This model will do a very good job of predicting the mean height of sons when the father is 6ft tall (oh, I should be using centimetres I suppose!); however, because of individual variation in the heights of sons, the model would be less reliable for predicting the height of an individual. Here the distribution of points about the line is not really a result of what I would call experimental error.

[3] A fast method to fill down a formula is to highlight both cells and double click the fill-down handle found in the bottom right corner. This repeats the formula down to the last contiguous cell in that column, or to the last contiguous cell in the column to the left if your current column has no other values.

Even though the data is obviously exponential, the exponential trendline does not necessarily give the best fit. Copy the original data to a new sheet (right click the tab, check the "Make a Copy" box and choose where in the workbook you want the new sheet placed). Repeat the steps above to fit a polynomial trendline of order 2, i.e. a quadratic. As can be seen below, the quadratic gives a better R^2 value (0.9987) as well as a better sum of deviations squared (9.35). For most exponential data this will be the case. The exception is when the data has exceptionally steep growth or decay.

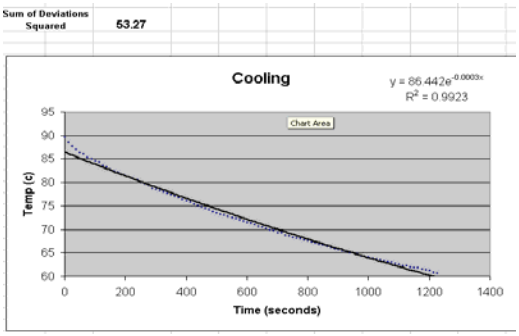

Figure 6: Exponential Model

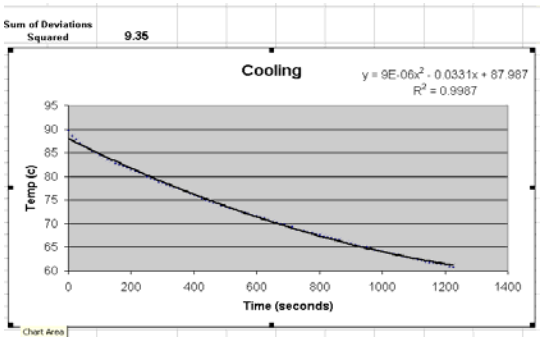

Figure 7: Quadratic Model

Having the sum of deviations squared as 9.35 is excellent, but we can still do a little better. By use of scrollbars[4] and the "Solver" function[5], we can lessen the sum of deviations squared even further. Copy your data to a new sheet.

At this time it would be advisable to check what add-ins you currently have installed in Excel. Add-ins can be accessed through the **Tools** button on your task bar. Recent versions of Excel provide only the more frequently used add-ins and I suggest that these are all you are likely to need. (Analysis Tool Pack, Analysis Tool Pack-VBA, Conditional Sum Wizard, Internet Assistant-VBA, Lookup Wizard, Solver Add-in). Check these boxes and follow the appropriate prompts.

Scrollbars can be found under **View>Toolbars>Forms**. Click this icon and draw a scrollbar onto your sheet. Right-click the object, choose "Format Control". Leave the settings as they are. Click in the "Cell Link" box and choose a cell from your sheet to display your scrollbar value in (Figure 8). Click "OK" then click anywhere on the desktop to activate your settings. You will now be able to change the value of your designated cell. Repeat these steps twice more (Figure 9).

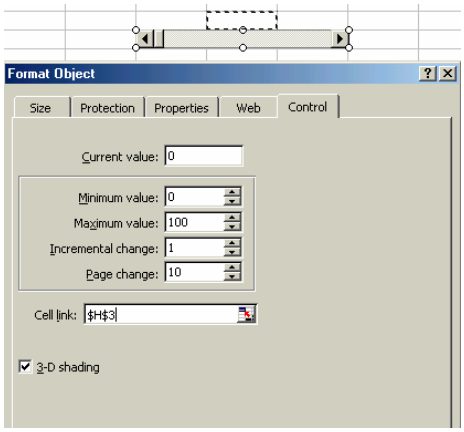

Figure 8: Scrollbar Properties

[4] Scrollbars are embedded objects within an Excel worksheet. They are used to change the value of a cell. If this cell is linked to a chart, we have an animated graph that can show us immediately the changes that occur as our parameters vary.

[5] The solver function is an "Add-in". Add-ins provide functions and interfaces that are not available in a normal Excel installation.

We will be using these three scrollbars to produce an exponential data set in the form $y = a*b^x + k$. Our goal will be for our students to change the value of these scrollbars and reduce the sum of deviations squared to their lowest possible value. To achieve this, we need to be able to adjust the range through which the bar will scroll without having to re-format the bar. This is easily done:

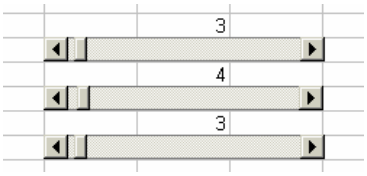

Figure 9: Adding Scrollbars

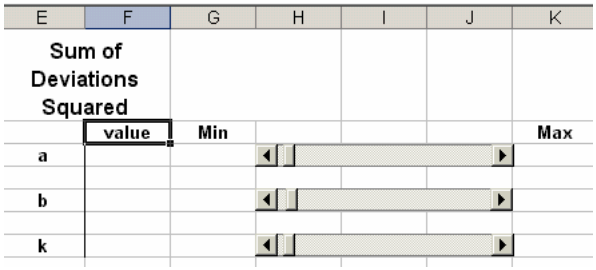

Figure 10: Controlling Parameters

Set up some cells to display your values, as shown in Figure 10. Also, pick up your scrollbars and move them to cover the cells they are linked to (right-click the bar to show the resizing handles to enable you move them). We need these values in the formulae we will be using, but they are not the final values we want to see, so let's cover them up.

In cells F3, F5 and F7 we need formulae that will produce results that fall between our desired maximum and minimum values as entered in cells G3, G5 and G7 and K3, K5 and K7. Since we don't wish to develop a new formula every time we change the maximum or minimum, this formula will need to respond to changes in the values entered in columns G and K. If we consider my values for "a", its scrollbar is linked to cell I3, and when we did "Format Control", it was set to range from 0 to 100. Therefore our formula in cell F3 becomes "= (K3-G3)/100*I3+G3". Play around with the minimum and maximum values and see that our value will always range between those you set. Repeat this formula for F5 and F7. To start the investigation, set the minimum and maximum values for all three scrollbars at 0 and 100. We will change these to more appropriate values later.

Just one quick diversion here. As can be seen above, formulae in Excel refer to cells on the sheet. This can be messy to debug later, as references such as K3 and G3 have no specific meaning to our task at hand. Since we want our exponential formula to be in the form $y = ab^x + K$, let's name cells F3, F5 and F7 as "a", "b", and "k" respectively. Click in cell F3 then click in the "name box" next to the formula bar. (Figure 11)

Change the F3 reference to "a" and hit enter to activate this name. All references to "a" in any formula in the entire workbook will now use the value found in F3. Repeat this process for values "b" and "k". (Note: we cannot use "c" as a name for a cell as Excel

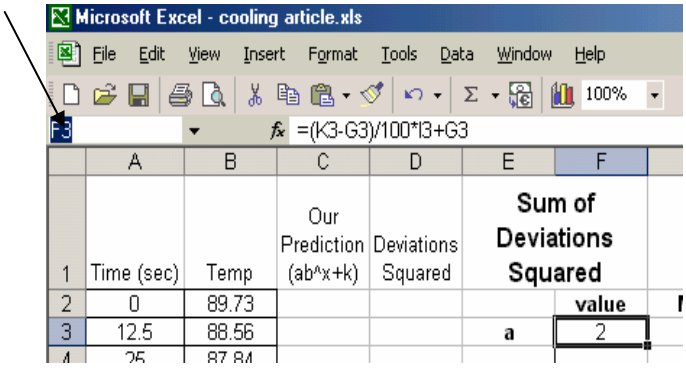

Figure 11: Adding Names

has this reserved for column references in Visual Basic).

At last we can enter our formula into C2 as follows, "= a*b^A2 + k". Double click the fill down handle to repeat this all the way to the bottom. Set up the deviations squared column as before and show the sum of this column at the top of your page for easy reference. Now let the fun begin.

Again, use Chart Wizard to graph the original data. Classroom discussion will be an excellent teaching tool that will help students set appropriate limits for "a", "b" and "k". To see a graph of our prediction, highlight the values in our prediction column (column C) and move your curser to the edge of the highlighted region. Your cursor should change to a four-headed arrow. Click and drag the data to your graph and drop it in. If you already had reasonable values for a, b and k, the graph will appear on your screen. If you cannot see it, keep adjusting your max and min values. (Figure 12)

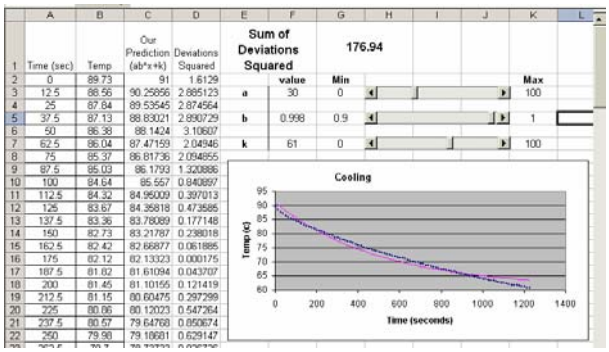

Figure 12: Adjusting parameters a, b and k

As students manipulate the values of "a", "b", and "k", the sum of deviations squared will reach its lowest value for the current max and min settings. Students can then change these limits and fine tune their result further. A classroom competition as to who can get the lowest score always eventuates. This is a competition that the teacher is sure to win, if he or she uses the "Solver" function which we added in earlier. Before you astound your students with your unbelievably low score, be sure to engage in a discussion as to what effect each parameter has on the shape of their graph.

Solver is found under the **Tools** menu. Set the target cell to wherever your sum is displayed (in my case, G1), choose the "min" option and set the cells to be changed as F3, F5, and F7. These must be separated by commas (Figure 13). Hit the Solve button and the best combination of a, b and k will be calculated. Accept the prompt that follows and astound your students with your score (5.25 for me). The only problem with the Solver function is that the formulae in F3, F5 and F7 are removed, but this is only a small inconvenience.

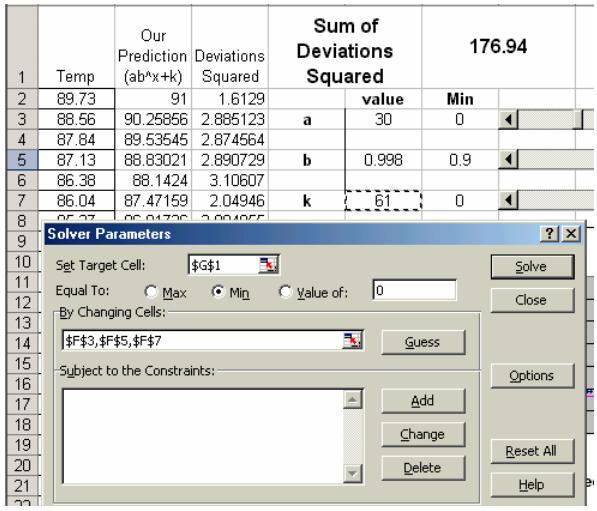

Figure 13: Using Solver

So, why go to so much trouble to use scrollbars and Solver to get such a small return on our efforts? (5.25 compared to 9.35 for the quadratic model developed by the computer) The power of a model comes from its ability to accurately predict. The

quadratic would be as useful as the model we developed for interpolation, but obviously inappropriate for extrapolation (for increasing values of time at least).

So, how good is our model for extrapolation? How does it compare with those generated by the computer? Copy the data to three new sheets and repeat the process of plotting the points and fitting a trendline, exponential, quadratic and using scrollbars, but only plot the first 50 points as your data source and see how accurate your models are at predicting the other 50 values. (Note: you will not be able to name your values a, b and k in your new scrollbar sheet as these are already in use. Named cells are global variables in the active workbook). Figure 14 shows how the three models compared when predicting the most extreme data point.

Max Time 1225 sec.	**Temp 60.7 °**	**% Error**
Exponential	53.8 °	11.8%
Quadratic	71.2 °	17.2%
Scrollbars	63.9 °	5.3%

Figure 14: Comparison of Models

The scrollbar-generated formula is clearly the best option for prediction of results outside the collected range. How would things change if we instead used the middle 50 values as our data source and extrapolated outside this range, above and below? This provides a starting point for a useful in-class discussion. Students can then complete this activity on their own, as homework or even as part of an assignment. The comparison in percentage errors between the quadratic model and our scrollbar model makes for some interesting discussion also.

Acknowledgements

The author wishes to thank Gloria Barrett, North Carolina School of Science and Mathematics, Durham NC, for her assistance with the statistical content of this paper.

References

[1] Flynn, P., Berenson, L., & Stacey, K. (2002). Pushing the pen or pushing the button: A catalyst for debate over future goals for mathematical proficiency in the CAS age. *Australian Senior Mathematics Journal*, Vol 16-2.

[2] MacGillivray, H. (1999). R[Squar]ED – DANGER! *Teaching Mathematics*. (Journal of the Queensland Association of Mathematics Teachers) Vol 24-4.

Appendix: Least Residuals

The reason the residuals are not the smallest for sums of squared deviations is as follows.

The line of best fit for a straight line through points is obtained from the minimum critical point of

$$S = \sum_{i=1}^{N} \{mx_i + c - y_i\}^2$$

The line of best fit for an exponential fit through points is obtained from the minimum critical point of

$$S = \sum_{i=1}^{N} \{ae^{bx_i} - y_i\}^2 \qquad (1)$$

This is particularly difficult to do mathematically and program and so $y = ae^{bx}$ is replaced by its logarithmic equivalent

$$\ln y = \ln a + bx$$

or

$$Y = A + bx \qquad \text{(where } Y = \ln y \text{ and } A = \ln a\text{)}$$

which is now linear. The Excel program then finds the minimum critical point of

$$\begin{aligned} S &= \sum_{i=1}^{N} \{bx_i + A - Y_i\}^2 \\ &= \sum_{i=1}^{N} \{bx_i + \ln a - \ln y_i\}^2 \end{aligned} \qquad (2)$$

The critical points for functions (1) & (2) are different, hence the critical point for (2) does not give the smallest sum-of-deviations-squared for equation (1) which is required.