USING THE EXCEL SOLVER FOR NONLINEAR REGRESSION

William P. Fox
Department of Defense Analysis
Naval Postgraduate School

Abstract

Most students have had some exposure to linear regression in their studies. Usually, this is limited to simple linear regression and perhaps multiple linear regression with several independent variables. Often, we need models that are nonlinear. We explain how to obtain these nonlinear models using the Excel Solver. Further, we point out the importance of the values of the initial decision variables in the Solver's schemes. We illustrate with two examples, one exponential model and one sinusoidal model with a linear trend.

Introduction

We teach a three course sequence in mathematical modeling for decision making. Our audience is mid-career military officers whose background prerequisite for our course is college algebra.

In course one we spend about five lessons in model fitting with least squares. We discuss the basic linear model, $y=mx+b$; polynomial models such as $y = a + bx + cx^2$, and multiple regression of the form: $y = a + bx + cz$. During those lessons we introduce the concept of minimizing the sum of squared error as our decision criterion as well as the concept of R^2 and a visual residual plot analysis. In course three we again need the same type models as course one so we do a review.

In course two on stochastic models we have found that the data sets we examine need more than basic models, we need nonlinear regression models. We will illustrate how we use Excel to perform two of these nonlinear regression examples, exponential models and sinusoidal regression with a linear trend.

Exponential Regression Using Excel's Solver

In this section we consider exponential regression. We'll see that if we do exponential regression in the usual way, we get an answer that is not as good as it could be. As a matter of fact, Fox (1993) showed that using the ln-ln transformations was merely an approximation to the nonlinear regression. We illustrate this again using Excel. Let's consider fitting an exponential function to the patient data taken from Neter (1996) and see what happens. First we will do the standard ln-ln fit with a transformation back into the original space and then we will compare the result we obtain using the Solver strictly to minimize the sum of squared error of the exponential function of interest. We find that we need the initial ln-ln model in order to obtain "good" initial estimates for our exponential model. We also note that in the example shown by http://archives.math.utk.edu/ICTCM/VOL13/C013/paper.html the differences in SSE are much more dramatic than our example below.

The data (from Neter, et al.,[1]) :

Days	Prognosis
Hosp	Index
2	54
5	50
7	45
10	37
14	35
19	25
26	20
31	16
34	18
38	13
45	8
52	11
53	8
60	4

65 6

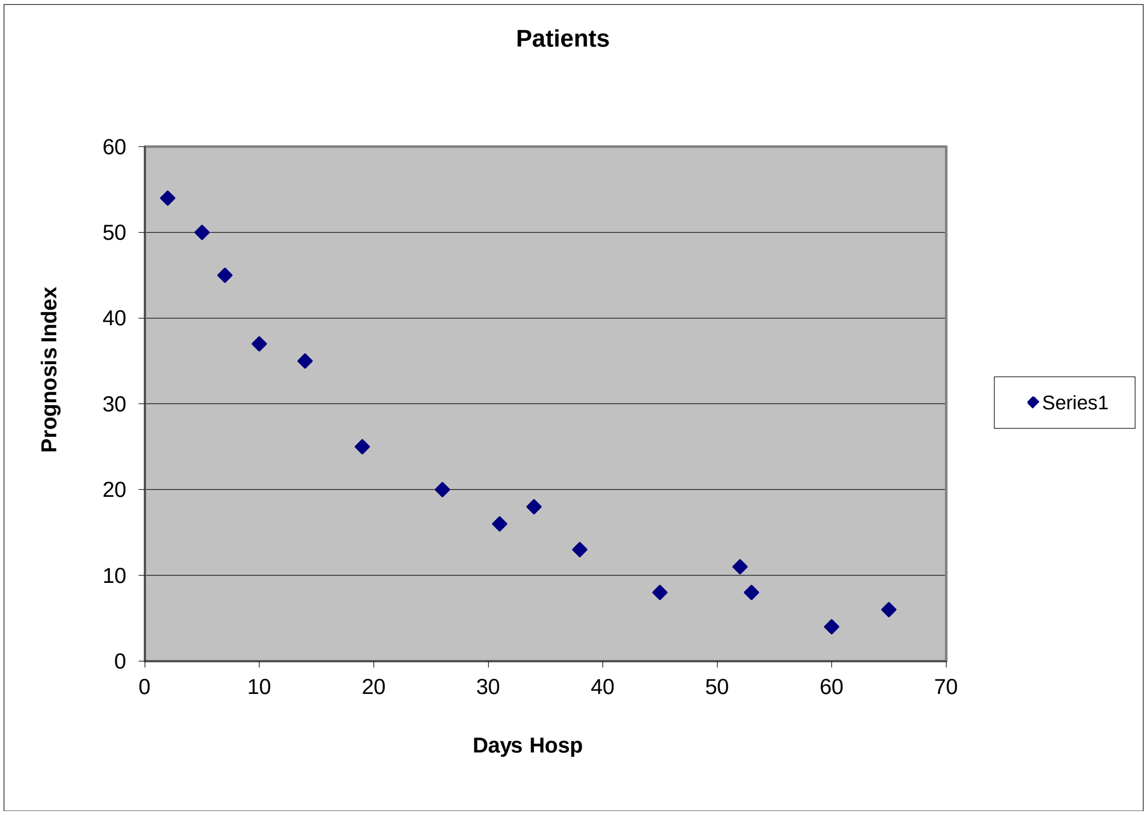

Figure 1. Scatter plot of patient data with decreasing exponential trend.

First we plot the data to examine the trends, see Figure 1.

The trend appears to a decreasing function that is slightly concave up. Our guess is perhaps an exponential model of the form, $y = g_o e^{g_1 x}$ might work well.

Prior to transforming the data in Excel, we need to know what transformation to apply. We take the natural logarithm of both sides of our model form, $y = g_o e^{g_1 x}$ to obtain the transformed model, $\ln y = \ln g_o + g_1 x$.

In Excel, we begin by taking the natural logarithm of the "y" variable to match our transformation and then we obtain a scatterplot of this transformed data, shown in Figure 2.

We note that the plot is reasonably linear so we obtain a linear regression model and use it to approximate our model. We utilize the regression feature in the Data Analysis package in Excel and obtain the summary output.

The only two pieces of useful information from this output are the intercept, 4.037159, and the slope, -0.03797. The regression model is *ln (y) = 4.037159-0.03797 x.* We need to transform this back using the laws of logs and exponentials into the real *xy* space to obtain y=*56.6646e*$^{-0.03797x}$. In this space we find we have a SSE of *56.08671*. The R^2 is about *(1-SSE/SST)* =0.995. We plot the residuals, Figure 3.

Days	ln(PI)
2	3.988984
5	3.912023
7	3.806662
10	3.610918
14	3.555348
19	3.218876
26	2.995732
31	2.772589
34	2.890372
38	2.564949
45	2.079442
52	2.397895
53	2.079442
60	1.386294
65	1.791759

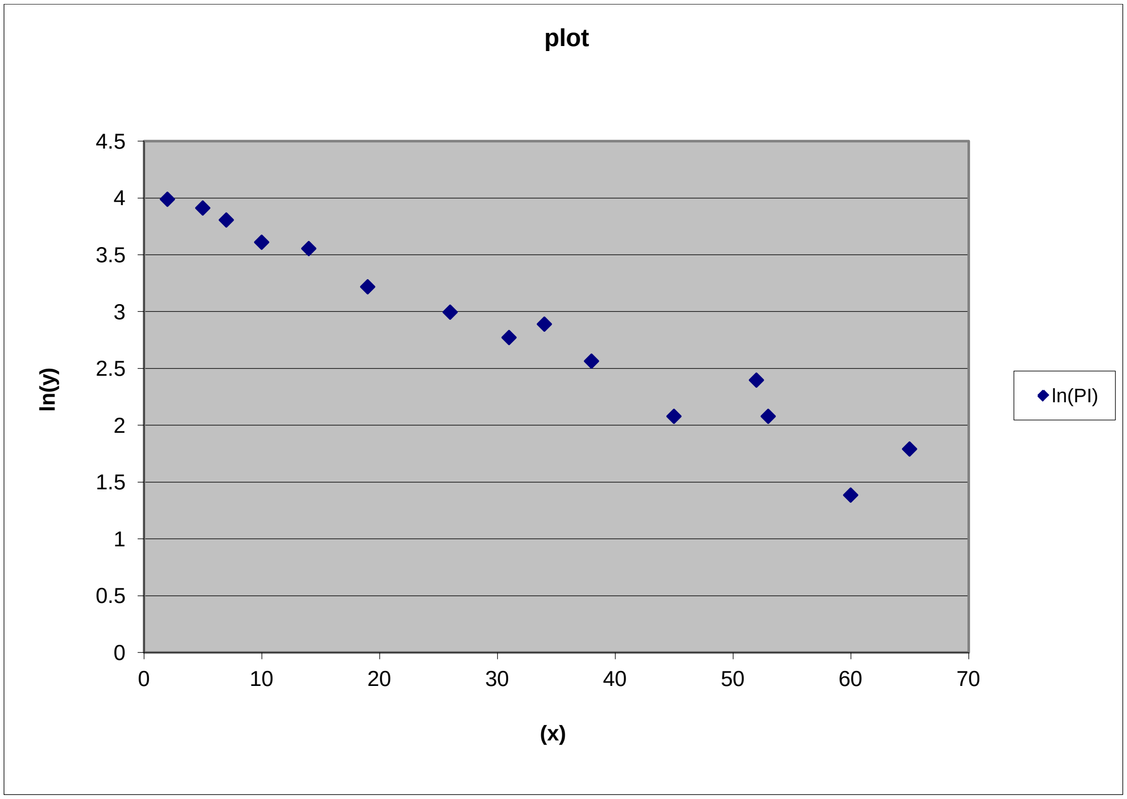

Figure 2. Plot of year versus ln(Prognosis Index) indicating a line.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.97728
R Square	0.955076
Adjusted R	0.951621
Standard E	0.179379
Observatio	15

ANOVA

	df	*SS*	*MS*	*F*	*ignificance F*
Regressior	1	8.892955	8.892955	276.3791	3.86E-10
Residual	13	0.418296	0.032177		
Total	14	9.311251			

	Coefficients	*tandard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*Ipper 95.0%*
Intercept	4.037159	0.084103	48.00247	5.08E-16	3.855465	4.218853	3.855465	4.218853
X Variable	-0.037974	0.002284	-16.62465	3.86E-10	-0.042909	-0.033039	-0.042909	-0.033039

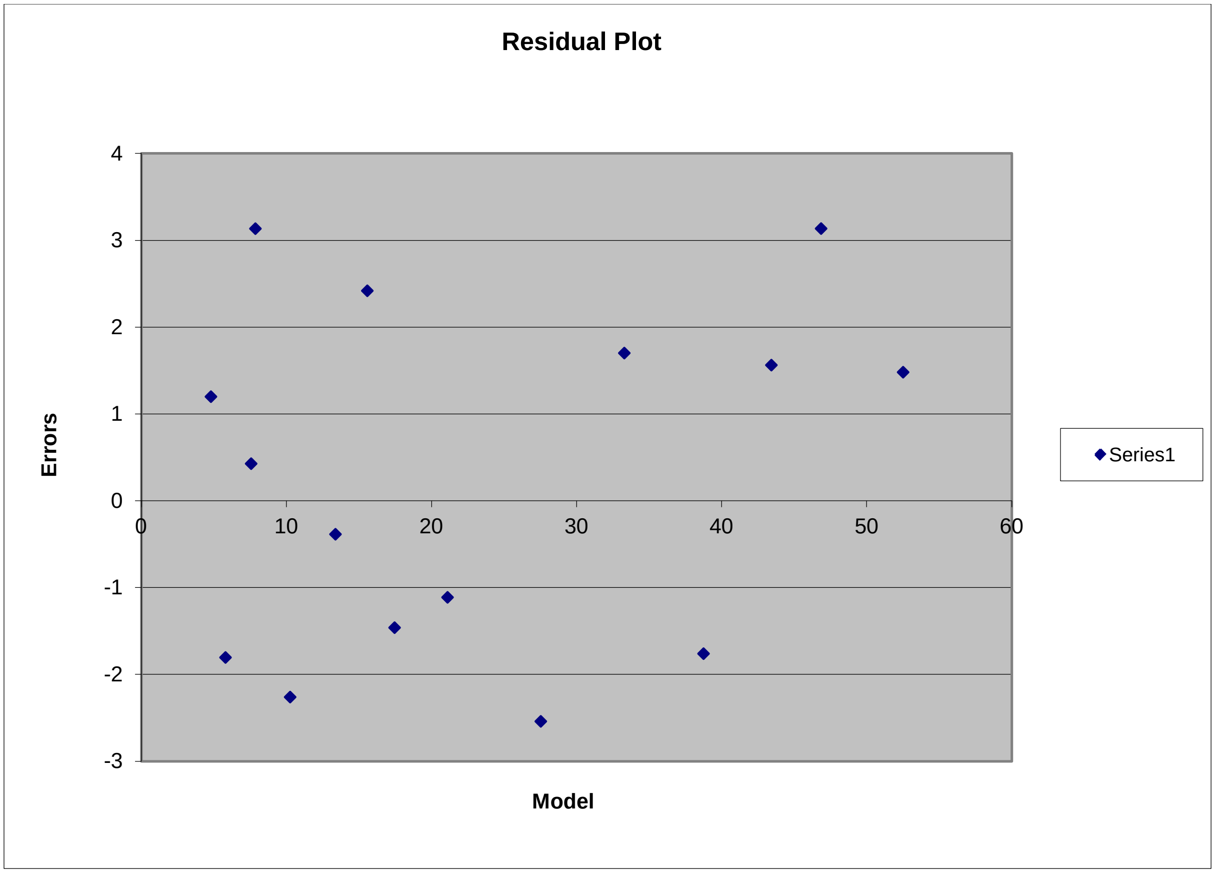

Figure 3. Residual plot from our approximate model.

Now, we go the Solver. Our model is $y = g_o e^{g_1 x}$ which has two decision variables g_o and g_1. Our model to minimize is $SSE = \sum (y_i - (g_0 e^{g_1 x}))^2$.

First, we assume the choice of decision variables values does not matter and picked g_o =1 and g_1 =1. The result was not good. We then go back and start with our transformed parameters as guesses: g_o =56.6646 and g_1 = - 0.03797 and obtain a final SSE of 49.459 and R^2 of 0.9959 with final parameters for our model leading to $y=58.60656e^{-0.039586x}$.

Sinusoidal Data with a Linear Trend Regression with Excel's Solver

Given the following CO2 data and the scatterplot below, see Figure 4.

Years	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Metric-Tons CO2	20	15	10	18	24	18	13	21	28	22	19	25	32	26	21	29	35	28	22	32

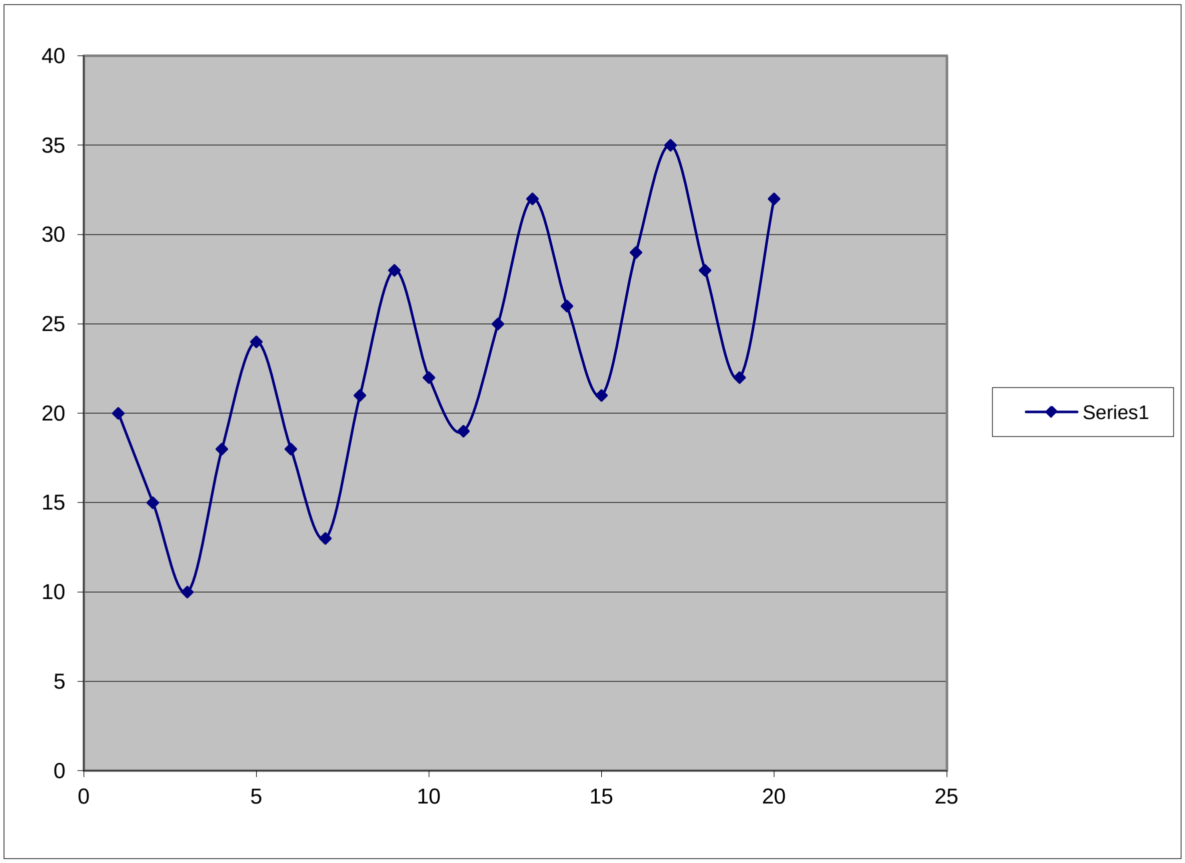

Figure 4. Scatterplot of CO2 data showing possible trend.

We analyze the trends seen from the plot. Our analysis of the data from the plot is that the trend is an increasing oscillating function over time. Our goal is predicting the next three months of CO2.

If we use the standard [1,1,1,1,1] as initial estimates for our model in the Solver, we obtained final estimates of:

Decision variables

a	1.567461
b	1.125106
c	1.766401
d	0.80018
e	14.61877

Our model *y = 1.567461*sin(1.125106 x + 1.766401)+0.80018 x + 14.61877* yields a model that we overlay in Figure 5 and we see does not fit well although it does oscillate with a linear trend even with a *SSE* of 382.2 and an R^2 that appears good at 0.966. The visualization gives the lack of validity of the model.

Neter et al. mention without much comment that sometimes a default like a vector of all 1's will not converge properly to the better result. This is the case above. The model does not capture the exact trend we seek. We must obtain better initial estimates.

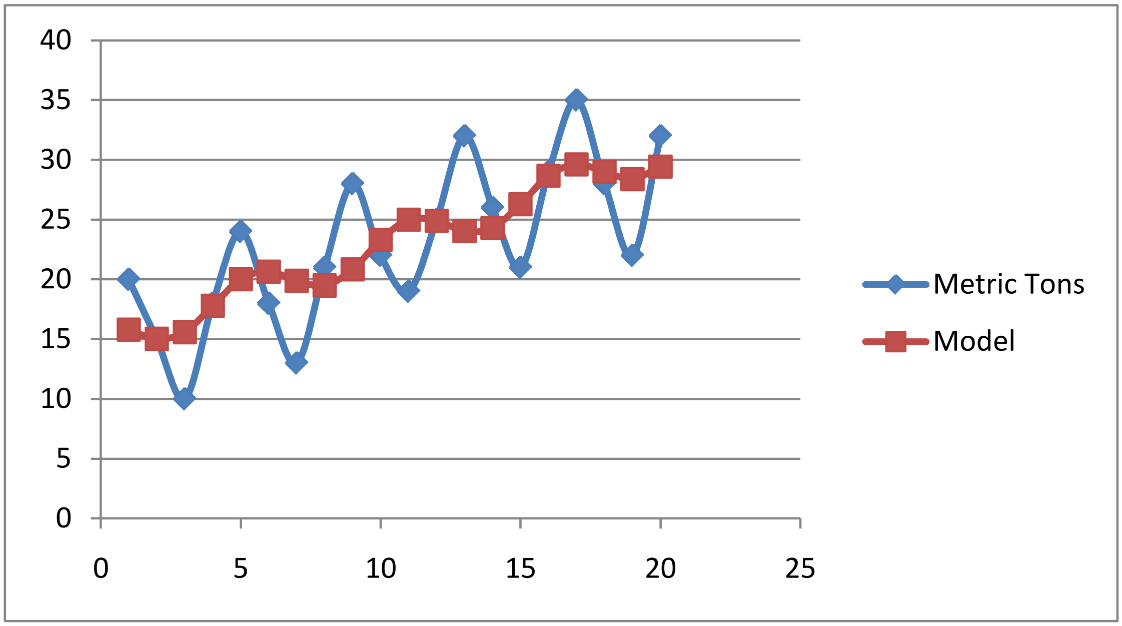

Figure 5. Plot of data and model from [1,1,1,1,1] showing a poor visual fit.

A quick review of trigonometric definitions for amplitude, period, and phase shift as well as a quick review of slopes of lines and intercepts enable the modeler to obtain a better initial guess. In this case, in order to do better, we go back to the original scatter plot, Figure 4, and we estimate the parameters of the model. The amplitude values (peaks) are from about 10 to 20 and provide an estimate of 5 for the amplitude. The period, *b*, is found by using the formula, $b=\frac{2\pi}{p}$. Weestimate $b=\frac{2\pi}{4}=1.57$.

We see a shift, *c*, of approximately -3 from a typical sine function as it is about 2 units to the right of the origin. The slope of the line through the oscillation, *d*, is about 1. Lastly, we use the midpoint of the sum of the peaks to approximate *e* as about 15. We point out that using these final parameter values previously obtained by the model did not constitute better initial estimates. Starting over using these new approximate starting values improves the model results. Our new starting vector is [5,1.57,-3,1,15]. We use these as our initial values in the Solver to obtain these values and model:

decision variables

a	-6.31644
b	1.573548
c	-3.05807
d	0.875956
e	13.69287

$$y=-6.31644*\sin(1.575486x3.05807)+0.875956x+13.69287.$$

In Figure 6, we show the data with our fit. Our *SSE* is 11.577 (a vast reduction from the previous 383.2) and an R^2 of 0.998977. More importantly, the figure attests to the model capturing our trends.

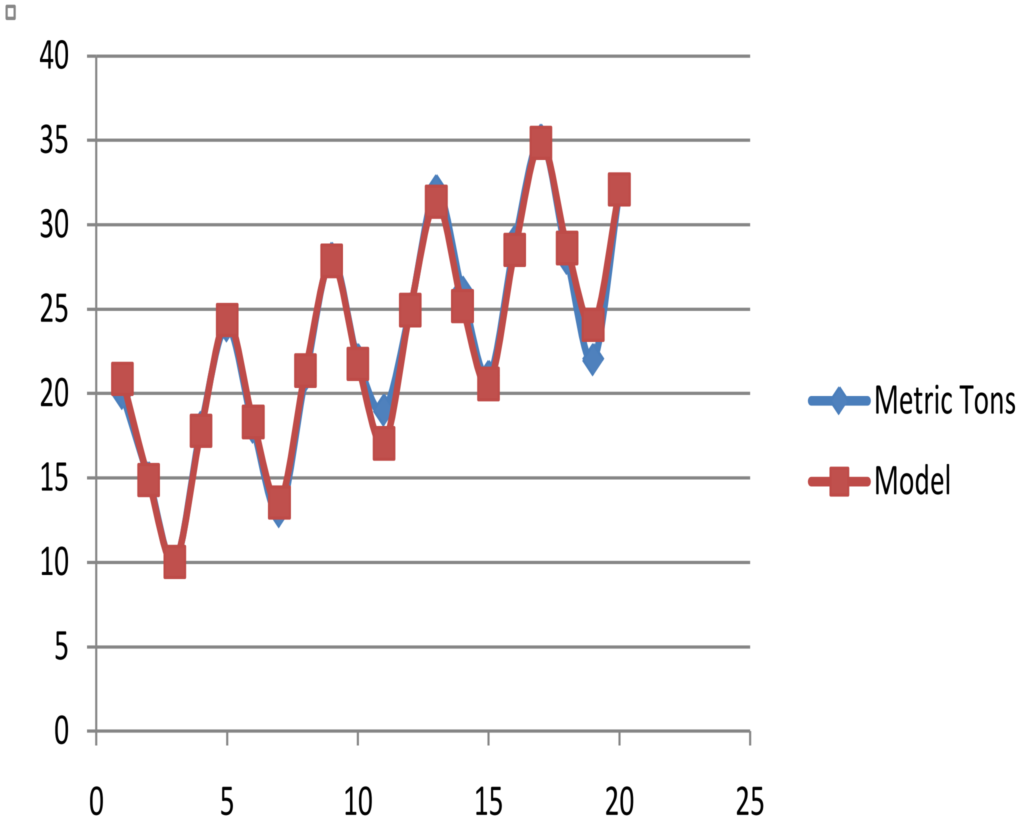

Figure 6. Plot of CO2 data and our model from the initial estimates [5,1.57,-3,1,15].

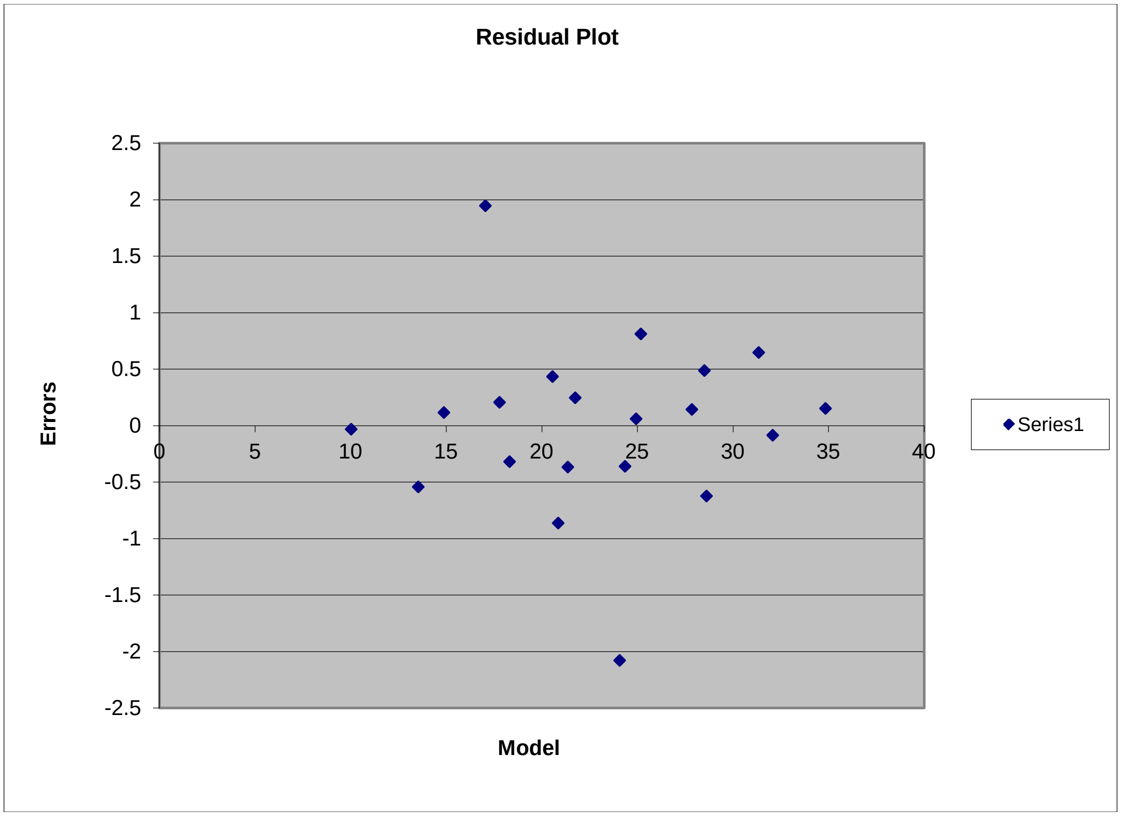

Figure 7. Residual plot from our model showing no pattern.

Class Example

Recently we were examining the following data for casualties due to IED devices in Afghanistan.

The following data represents 9 years of IED casualties from 2001-2009.

We want to predict the casualties for 2010-12 with our model. We then take the real 2010 outcome and compare it to our model to see how well the model performed.

Our data is increasing and curved (see Figure 8), so in our analysis in class we build and compare both a polynomial model, $y = b\,x + c\,x^2$, and an exponential model, $y = a\,e^{b\,x}$. The polynomial model is a review from previous classes on basic regress and the students obtain the model, $y = -20.7\,x + 7.27\,x^2$ with a *SSE =6774* and an R^2 of approximately 0.979. Additionally the residual plot shows no trend. The prediction for $x=10$ is 520.

Next, we want to fit the model, $y = a\,e^{b\,x}$. Using Excel 2010's Solver and its GRG Nonlinear solver and starting values for "a" and "b" of 1, we find the optimal values are $a=7.077$ and $b=0.4602$. This gives the model, $y = 7.077\,e^{0.4602\,x}$. The R^2 is *0.9965* and the residual plot show no trends (see Figure 9). Our prediction for $x=10$ is *705.51*.

Year	2001	2002	2003	2004	2005	2006	2007	2008	2009
# IED Casualties	4	25	26	27	73	130	184	263	451

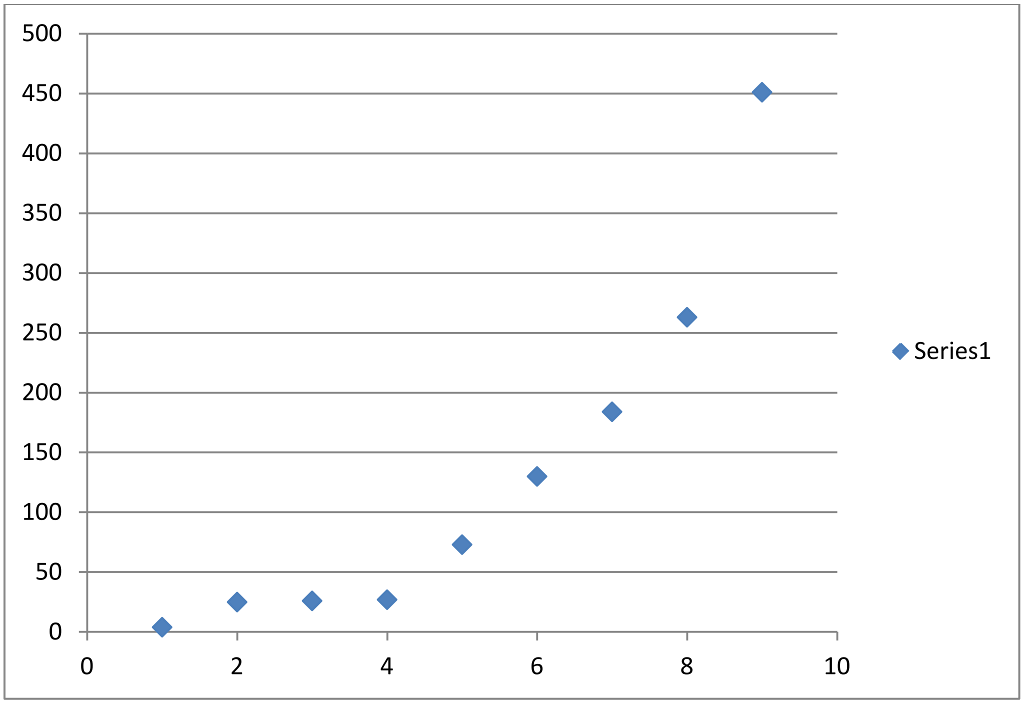

Figure 8. Scatterplot of IED casualties.

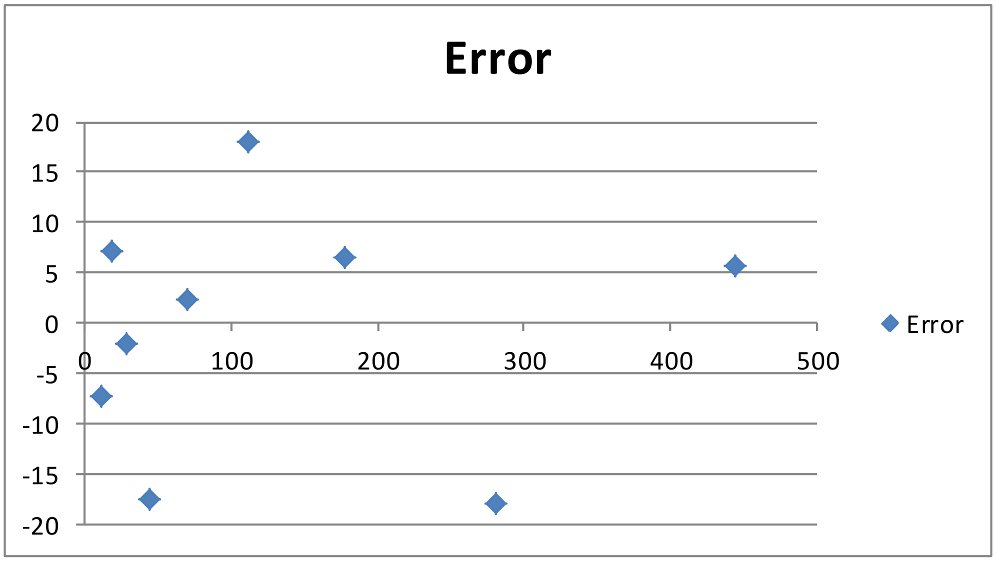

Figure 9. Residual plot shows no trend.

The students find that there were actually 630 casualties in 2010. The polynomial model under estimated the actual value by 17.4% while the exponential model over estimated 12.3%. The students conclude in this example that the exponential model is better not only because the R^2 was higher and the *SSE* was smaller but also they would prefer to over estimate casualties than under estimate them.

Conclusion

We have shown how to use the Solver to perform the minimization of the sum of squared error. Further, we have shown the importance of starting values used by the Solver to obtain the final models. Additionally, we have shown the importance of visualization not relying only on *SSE* and R^2 to determine model adequacy. In other papers (Fox, 2011) should that good starting points are critical in using Maple, a powerful computer algebra system to achieve similar results.

So why use Excel at all when there are other packages available? The main reasons are:

- Excel is readily available and very inexpensive (often it is included with the computer when it is purchased).

- Although the Solver takes care of finding the parameters, there is pedagogical value is setting up the function for optimization. We think that students get a better feel for the process using Excel.

- It is fun to set up and solve!

- The Solver might be the only available package and data might require a nonlinear model.

References

1. Neter, J, Kutner, M. H., Nachtsheim, C. J., and W. Wasserman. (1996). Applied Linear Statistical Models. 4th Ed. Irwin Press. Chicago, Il.

2. Fox, William P. (1993). The Use of Transformed Least Squares in Mathematical Modeling, *COED Journal*, Vol. III, No. 1, pages 25-31.

3. Fox, William P., Frank R. Giordano, Steve Horton, and Maurice Weir. (2009) A First Course in Mathematical Modeling, 4th Edition, Cengage Publishing: Brooks/Cole Publishing Co., Belmont, CA. (xvii + 620 pages, ISBN-13:978-0-495-01159-0)

4. Fox, William P. and Frank Giordano (2010). Workshop in Mathematical Modeling for Montana Green. Helena, Montana.

Biographical Information

Dr. William P. Fox received his BS degree from the United States Military Academy at West Point, New York, his M.S. at the Naval Postgraduate School, and his Ph.D. at Clemson University. He has taught at USMA, at Francis Marion University, and is currently a professor at the Naval Postgraduate School. He has authored two textbooks on mathematical modeling. He has over one hundred technical and educational articles, presentations, and workshop presentations. He serves as the contest director for COMAP's Collegiate Mathematical Contest in Modeling (MCM) and contest director for the High School Mathematical Contest in Modeling (HiMCM). His interests include applied mathematics, optimization (linear and nonlinear), mathematical modeling, statistical models for medical research, and computer simulations.