FITTING SURGE FUNCTIONS TO DATA

Sheldon P. Gordon

ADDRESS: Department of Mathematics, Farmingdale State University of New York, Farmingdale NY 11735 USA. gordonsp@farmingdale.edu

ABSTRACT: The problem of fitting a surge function to a set of data such as that for a drug response curve is considered. A variety of different techniques are applied, including using some fundamental ideas from calculus, the use of a CAS package, and the use of Excel's regression features for fitting a multivariate linear function to a set of transformed data. The results of the different approaches are contrasted and discussed.

KEYWORDS: Surge functions, drug response curves, curve fitting, least squares, multivariate regression analysis, rational functions.

INTRODUCTION

One of the "new" families of functions that are being introduced in the early years of the mathematics curriculum is the *surge* function, which is treated at the calculus level in [2, 3] and at the precalculus level in [1]. The surge function, whose graph is shown in Figure 1, has the form

$$f(x) = Ax^p e^{-bx}, \tag{1}$$

where $p > 0$ and $b > 0$, which is equivalent to $f(x) = Ax^p c^x$, $0 < c < 1$. Surge functions are used to model a variety of real-world applications, such as the response to an initial dose of a drug (the level of the medication in the bloodstream rises relatively rapidly to a peak and thereafter decays as the drug is washed out of the body by the kidneys) or the body's response to an infection. Surge functions are also used to model the results of an advertising campaign that initially causes a fast increase in sales, but which then slowly diminish. From a modeling point of view, the initial surge is

accounted for by the power function term x^p and the subsequent slow decay is accounted for by the decaying exponential term e^{-bx} or c^x.

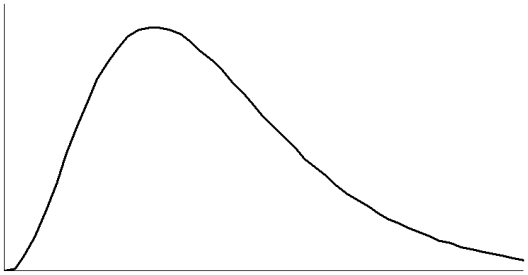

Figure 1. The graph of a surge function.

A surge function such as the one pictured in Figure 1 has a maximum and two points of inflection for $x \geq 0$. As shown in [4] using standard calculus techniques, the maximum occurs at the point where

$$x = \frac{p}{b}, \quad b \neq 0. \tag{2}$$

A second horizontal tangent occurs at the origin where $f'(0) = 0$; depending on the value of p, this could be a turning point or an inflection point. The two inflection points seen in the graph in Figure 1 occur at $x = \frac{p \pm \sqrt{p}}{b}$. This can be written in a more insightful way as

$$x = \frac{p}{b} \pm \frac{\sqrt{p}}{b}, \tag{3}$$

which indicates that the two inflection points are located symmetrically about the turning point at $x = \frac{p}{b}$; however, the inflection points do occur at different heights since the curve is not symmetric about the vertical line through the turning point.

Another extremely important theme in the modern mathematics curriculum is that of fitting a function to a set of data. All graphing calculators (as well as spreadsheet packages such as Excel) have the capability of fitting a linear, exponential, power, logarithmic and polynomial function (up to 4^{th} degree on a calculator and up to 6^{th} degree on Excel) to a set of data. Many calculators also have the capability of fitting a logistic function and a sinusoidal function to data. However, unless one uses a specialized computer package such as Mathematica or Maple, there is no readily available tool for fitting a surge function to a set of data. This issue was addressed in [4], where several different techniques were discussed, but none was particularly satisfactory in the sense of yielding good results in an easily accessible way.

In this article, we look at this issue again and consider in detail an approach that has the advantage of giving reasonably accurate results with a readily available tool, as well as the approach of applying the least squares criterion directly with the assistance of a CAS package.

Figure 2. The drug response curve for Viagra.

To illustrate the different approaches, we will use data on the mean plasma concentration level for sildenafil citrate (Viagra), a drug that our students would certainly be aware of and which would obviously pique their interest. We show the drug response curve for Viagra, as posted at the Pfizer website [5], in Figure 2. In this situation, the independent variable t is time in hours since a dose of Viagra was taken initially and the dependent variable, the mean plasma concentration level C for a group of healthy male volunteers, is measured in nanograms per milliliter (ng/ml). It is evident that the drug achieves its maximum level slightly more than an hour after it is taken and thereafter the level decays relatively slowly.

From this graph, we can estimate a set of data points to use as our target in finding an equation for the surge function that matches the curve shown in the Pfizer graph. In particular, we estimate the following values (see Table 1) for t in hours and the corresponding concentration level C in nanograms per milliliter:

t	0.05	0.4	0.6	1.2	1.8	2.1	3	4	6	8	10	12	18	24
C	1	50	320	440	410	350	250	170	80	50	30	20	12	6

Table 1. Data on level of Viagra.

It is interesting to note that the first point shown in Pfizer's graph in Figure 2 is not at the origin although we would presume that the initial Viagra concentration level would be 0 at time 0. We will address this issue later in the article.

From the data, we conclude that the maximum concentration level of about 440 ng/ml occurs at about $t = 1.2$ hours. Furthermore, the two inflection points, which correspond to the points where the function is changing most rapidly, occur at about $t = 0.4$ and $t = 2.4$ hours after the Viagra is first taken. Since we know that the inflection points for a surge function should be symmetrically located about the turning point, we might opt to average the two deviations about $t = 1.2$ and so estimate that the inflection points are 1 hour above and 1 hour below $t = 1.2$; that is, at $t = 0.2$ and $t = 2.2$.

Alternatively, we might reason that just because the largest value of C in the table is at $t = 1.2$ hours does not necessarily mean that this is the absolute maximum value – the Viagra level might reach a higher level somewhat beyond $t = 1.2$, say at $t = 1.3$ or $t = 1.4$, and the latter value might make for a more symmetric format. We leave this possibility for the interested reader to pursue.

We now substitute the estimates for the location of the turning point ($t = 1.2$) and the inflection points ($t = 0.2$ and 2.2) into Equations (2) and (3) to get $\frac{p}{b} = 1.2$ and $\frac{p+\sqrt{p}}{b} = 2.2$ and the latter is equivalent to $\frac{\sqrt{p}}{b} = 1$. These two equations can be solved readily to get $p = 1.44$ and $b = 1.2$, so that the form of our surge function is

$$C(t) = At^{1.44}e^{-1.2t}.$$

We know that the peak concentration value of approximately 440 ng/ml occurs at about $t = 1.2$, so that $C(1.2) = A(1.2)^{1.44}e^{-1.44} = 0.30806A = 440$, and so $A = 1428.29$. The corresponding model for the surge function is therefore

$$C(t) = 1428.29t^{1.44}e^{-1.2t}. \tag{4}$$

We show the graph of this function superimposed over the data points in Figure 3 and conclude that it is a reasonably good fit for t between 0 and about 3 hours, although thereafter the surge function dies out much more rapidly than the concentration of Viagra does.

Probably the most common measure used to assess how well a function fits a set of data is the sum of the squares of the vertical deviations between the curve and the data points. The corresponding value for the surge func-

tion (4) shown in Figure 3 is 53,459.9. We will use this value for comparison in our subsequent calculations.

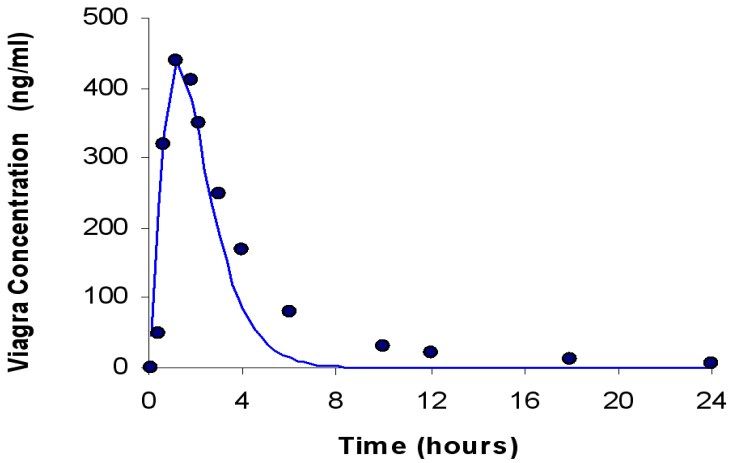

Figure 3. Viagra data and the surge function.

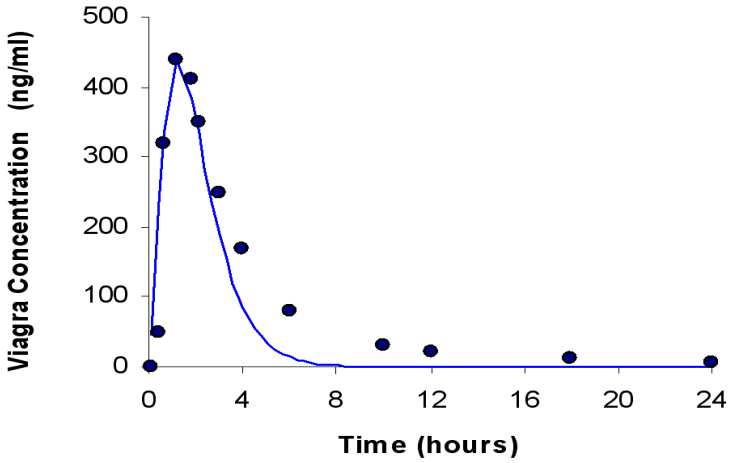

Figure 4. Another surge function with the Viagra data.

We can obtain a better approximation to a best-fit surge function by using some of the features in Mathematica to minimize the value of the sum of the squares. (The actual Mathematica session, including the specific commands used and the corresponding output, is shown in Appendix A.)

The result of this process is the quite different surge function

$$C(t) = 1100.31t^{1.6634}e^{-1.08376t}. \quad (5)$$

All three of the parameters have changed considerably and should change the function significantly compared to the surge function in Equation (4). The corresponding value for the sum of the squares associated with this function is 25,815.1, so that it is significantly smaller (about $\frac{1}{2}$ as large) than the value of 53,459.9 that we got by simply applying the calculus-based results. We show this function superimposed over the data points in Figure 4, and observe that it seems to be a better fit to more of the points than the surge function (4) in Figure 3. Nevertheless, this function is still a rather poor fit to the data, especially after about $t = 5$ hours. And, perhaps more importantly, because this approach requires the use of a specialized CAS program that might not be available to all students, it may not be the most effective approach with which to investigate other sets of data that fall into the pattern of a surge function.

USING MULTIVARIATE REGRESSION

We now consider a different way to find the equation of a surge function that fits a set of data. To do so involves using some ideas on fitting a linear function of two or more variables to a set of multivariate data. In particular, suppose we have a set of $(x_1,\ x_2,\ \ldots,\ x_n, y)$ data, where y depends on the n independent variables $x_1, x_2, \ldots, x_n$. Multivariate linear regression is a standard tool that is used to find the linear function $Y = c_0 + c_1X_1 + c_2X_2 + \ldots + c_nX_n$ that is the best fit to the data. Think of this as finding the hyperplane in $n+1$ dimensional space that comes closest to all of the points in the data set. This capability is available in many software packages, including Excel, so it is readily at hand for use. (Note that this feature is not automatically loaded when Excel is first installed; rather, it must be loaded one time as an Add-In under Excel's Tools menu – just select Analysis ToolPak, and it is thereafter available under Tools. One of the options you then get is Regression. We describe the use of this feature in Appendix B.)

One somewhat unexpected application of multivariate linear regression is in fitting a polynomial in one variable x to a set of (x, y) data. Suppose we wish to find a polynomial of degree 3, $y = a_0 + a_1x + a_2x^2 + a_3x^3$, that fits such a table of values. We can think of the polynomial expression as a linear function of x, x^2, and x^3 and then apply multivariate linear regression, where $X_1 = x$, $X_2 = x^2$, and $X_3 = x^3$. The resulting coefficients

for the constant term and for the three variables X_1, X_2, and X_3 are then the coefficients of the desired polynomial.

We now apply a similar approach to fitting a surge function to a set of (x, y) data, and will then apply the procedure to the data on the concentration levels of Viagra. Since the surge function we seek has the form

$$y = Ax^p e^{-bx},$$

we can take logarithms of both sides to get

$$\log(y) = \log(A) + \log(x^p) + \log(e^{-bx}) = \log(A) + p\log(x) - bx$$

and so $\log y$ is a *linear* function of x and $\log x$. Thus, we can set $Y = \log y$, $X_1 = x$ and $X_2 = \log x$ and apply multivariate linear regression to an extended table of values that also includes a column of $\log x$ values and a column of $\log y$ values. In order to take logs of the t and the C values, we need to avoid the obvious starting point where $t = 0$ and $C = 0$; we do this by making a very minor change in the values of the two variables at that point and use $t = 0.05$ instead of 0 and $C = 10$ instead of 0. We presume that the researchers at Pfizer did the same in producing the graph in Figure 2 on their website; otherwise, it is far more natural to use $(0, 0)$ as the starting point. For the Viagra data, we then have the extended Table 2.

C	t	$\log t$	$\log C$
10	0.05	-1.30103	1
50	0.4	-0.39794	1.69897
320	0.6	-0.22185	2.50515
440	1.2	0.079181	2.643453
410	1.8	0.255273	2.612784
350	2.1	0.322219	2.544068
250	3	0.477121	2.39794
170	4	0.60206	2.230449
80	6	0.778151	1.90309
50	8	0.90309	1.69897
30	10	1	1.477121
20	12	1.079181	1.30103
12	18	1.255273	1.079181
6	24	1.380211	0.778151

Table 2. Original and transformed data.

When we "hit" this set of transformed data with the multivariate regression features of Excel, we get the linear regression equation

$$Y = 2.3190 - 0.1242X_1 + 0.7613X_2,$$

which is equivalent to

$$\log C = 2.3190 - 0.1242t + 0.7613 \log t.$$

We can eliminate the logs algebraically by undoing the original transformation using powers of 10 and so obtain

$$\begin{aligned} 10^{\log C} = C &= 10^{2.3190 - 0.1242t + 0.7613 \log t} \\ &= 10^{2.3190} 10^{-0.1242t} 10^{0.7613 \log t} \\ &= 208.45 (10^{-0.1242})^t 10^{\log t^{0.7613}} \\ &= 208.45 (0.7513)^t t^{0.7613}. \end{aligned}$$

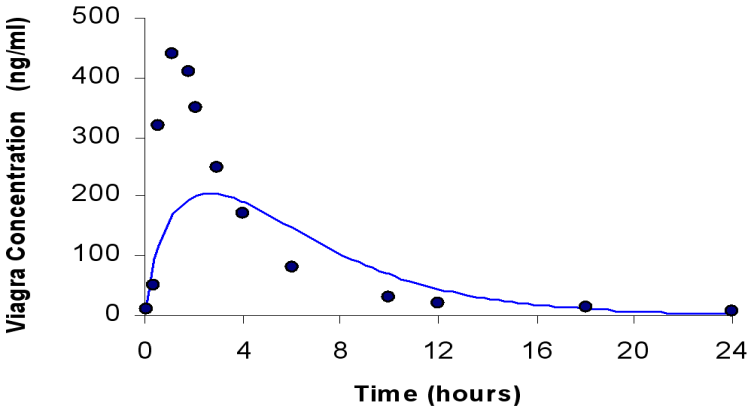

Figure 5. The surge function based on multivariate regression.

The base 0.7513 in the exponential term can be converted to an appropriate power of e by solving the exponential equation $e^{-b} = -0.7513$, which leads to $b = 0.2859$. Thus, we have the surge function $C(t) = 208.45t^{0.7613}e^{-0.2859t}$, which is shown superimposed over the data points in Figure 5. It is obviously a very poor fit to the data, other than at the very beginning and the very end. Furthermore, the associated value for the sum of the squares is 192,812.4, which is considerably larger than anything we had before and which therefore corroborates our visual conclusion that

the fit is extremely poor. Yet, the logic leading up to this result seems reasonable in the sense that the multivariate regression process produces the best-fit plane to the transformed data and therefore leads us to expect a much better fit. Let's see what has gone wrong.

In Figure 6, we show the plot of the points $(\log t, \log C)$. Notice that, other than the left-most point $(-1.5, 1)$, the remaining points are clustered relatively tightly and mostly display a clear pattern. This suggests that the results we get for the regression equation might be very sensitive to small changes in the values of the coordinates at the left-most point in the sense that this point may have a disproportionate effect on the coefficients in the regression equation.

Moreover, the left-most point is what we estimated to avoid the problem with taking logs of 0. And, because it involves a negative value for t, a relatively minor change in the value of t near 0 would likely result in a major change in the value of $\log t$. In addition, when you look back at Figure 2 (the Pfizer website graph), it is evident that this initial point is the one for which it is hardest to estimate an accurate value.

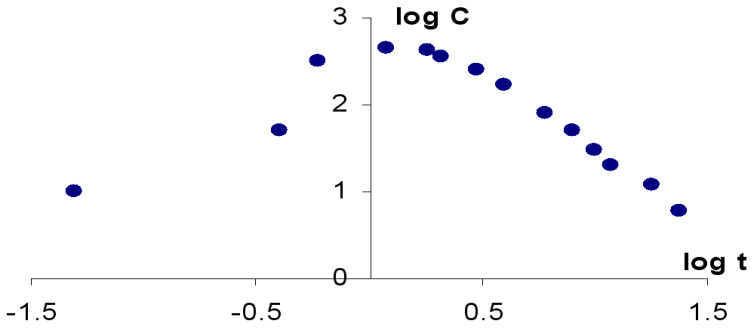

Figure 6. Plot of log C vs. $\log t$.

Let's see just how much of an effect we get by trying a slightly different estimate for the value of t for this point. Instead of using $t = 0.05$, suppose we try $t = 0.10$ and maintain the value $C = 10$. The resulting surge function is $C(t) = 619.44t^{0.8236}e^{-0.2924t}$. The value for the coefficient has changed dramatically from $A = 208.45$; the power in the power function term has changed a bit from $p = 0.7613$ to $p = 0.8236$; and the multiple in the exponential term has changed fairly minimally from $b = -0.2859$ to $b = -0.2924$. However, the corresponding value of the sum of the squares is now 700,741.1, which is almost four times as large as the previous value of 192,812.4 and the resulting surge function is a far poorer fit to this data. More significantly, a relatively small change in the estimate of the point

near the origin clearly results in a huge change in the results.

One way to circumvent this issue is to realize that, by its very nature, every surge function must pass through the origin provided that the power p in the power function term is positive. As a consequence, it might make sense to ignore the point near the origin altogether and see what happens if we use only the remaining points for the multivariate regression analysis. When we do this, the corresponding linear function is $Y = 2.4276 - 0.092X_1 + 0.225X_2$, which is equivalent to $\log C = 2.4276 - 0.092t + 0.225 \log t$.

When we undo the logarithmic transformation by taking powers of 10, we eventually get the surge function

$$C(t) = 267.67(0.8091)^t\, t^{0.225} = 267.67t^{0.225}e^{-0.2118t}.$$

The associated value for the sum of the squares is 153,650.7. This is a substantial improvement over the two preceding surge functions using multivariate regression with estimates of the point near the origin. However, it is still considerably larger than the value of 53,459.9 we initially obtained using the calculus argument, let alone the value of 25,815.1 that resulted from the Mathematica routine for minimizing the sum of the squares.

Incidentally, there is another statistical measure used to assess how well a multivariate linear function fits a set of data; it is the *coefficient of multiple determination* and is denoted by R. It is the extension of the correlation coefficient r to multivariate data. For the three functions we have created using multivariate regression, the corresponding values are $R = 0.8768$, $R = 0.8581$, and $R = 0.8988$, respectively. While all three values are statistically significant, the fact that they are all quite close to one another means that we should not make a definitive call on which of the three surge functions is the best fit based solely on the value of R.

This still leaves one rather perplexing question. How can all three of these surge functions based on multivariate linear regression be so much poorer fits than the one based on calculus, let alone the one obtained using the computer search method? After all, multivariate linear regression is supposed to produce the *best* fit! The key is that it does produce the best *linear* fit to the set of transformed $(t, \log t, \log C)$ data. If all we did was to stop there, we would indeed have the best possible fit. However, we started with (t, C) data and, in the process of transforming it via logarithms, we stretched the data values in a non-linear way. After we got the corresponding multivariable linear regression equation, we undid the original transformation, which entails another non-linear stretch, but this time the inverse transformation is applied to the function, not to the data. So, although the three regression planes we obtained were the best linear

fits to the three different sets of transformed data, the corresponding surge functions are not necessarily the best, or even extremely good, fits to the original data. They may be good fits if the original data fall very closely into a surge function pattern; however, if the data is not *extremely* close to such a pattern, the resulting function based on multivariate regression may be a surprisingly poor fit.

A comparable situation arises with the curve fitting routines in calculators and in Excel; rather than directly fitting an exponential, logarithmic, or power functions to a set of data, these routines transform the data (either a semi-log plot or a log-log plot), find the regression line for the transformed data, and then undo the transformation algebraically. In the process, one obtains the best possible line for the transformed data, but in the process of undoing the transformation, a nonlinear stretch takes place and the resulting function is not necessarily the best fit within that family of functions.

FITTING A RATIONAL FUNCTION TO THE DATA

Gordon and Gordon [4] also discuss the possibility of fitting a rational function of the form

$$C(t) = \frac{at^2}{t^4 + b^2}, \tag{6}$$

to a set of data on drug concentration levels over time, where a and b are two constants. The quadratic term in the numerator is needed to reflect the curvature of the drug data at and near the origin; the quartic term in the denominator reflects the fact that the data eventually die out as time progresses.

In [4], it was found that such a rational function was actually a considerably better fit than a surge function is to a set of data on the drug concentration level for a form of L-Dopa used to treat patients with Parkinson's disease. Let's see how well such a function fits the data on Viagra. We again use Mathematica to perform a direct least squares fit with a rational function of the form in Equation (6). The resulting function is

$$C(t) = \frac{1822.6t^2}{t^4 + 3.19129^2}.$$

where the corresponding value for the sum of the squares is 28,468.6. We note that this value is slightly larger than the value of 25,815.1 we obtained before for the best fitting surge function. So, in this case, the rational function gives slightly poorer accuracy. In Figure 7, we show both the best surge function and this best rational function of the form in Equation (5) to compare the relative fits. From this, we see that the rational function

(the darker curve) spikes to a considerably higher level than either the data or the surge function do; however, it dies out more slowly than the surge function and so is a better fit to the data after about $t = 12$ hours.

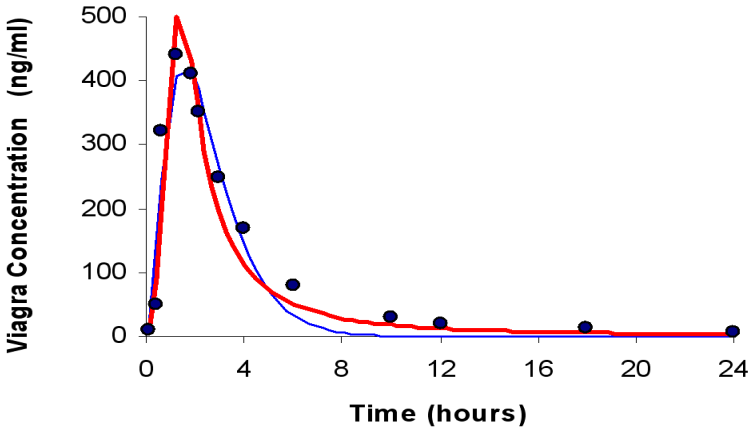

Figure 7. The rational function (darker curve) vs. the surge function.

CONCLUSIONS

In general, the focus of this kind of investigation should not be on simply finding a function that is the best possible fit to a set of data, but rather on finding a function that is a reasonable fit and whose properties provide insight into the situation being modeled. A surge function provides that kind of insight in the sense that the power function term models the initial impetus and the exponential term models the eventual exponential-type decay of the drug concentration levels. The rational function certainly fits the data well, but there are considerably less compelling interpretations for why it follows the desired behavior pattern – the term at^2 does model the initial impetus and the $\frac{1}{t^4+b}$ term does die out relatively quickly, but the latter is nowhere as convincing a description as exponential decay.

So the key is the realization that all we are producing is a mathematical model. There are different routes to developing such a model, not only the ones we discussed here. For instance, there may be a more sophisticated pharmacokinetics model based on differential equations, but that is beyond the scope of what we are considering here and likely beyond the scope of the students we serve in introductory courses. In the final analysis, what is most important is not the keystrokes used to produce a model, but an understanding of the mathematics underlying that model and the

development of the judgment necessary to decide how well the function actually fits the data and how the model gives us an understanding of the process.

ACKNOWLEGEMENT

The work described in this article was supported by the Division of Undergraduate Education of the National Science Foundation under grants DUE-0089400, DUE-0310123, and DUE-0442160. The author also appreciates the assistance provided by Brian Winkel, editor of *PRIMUS*, with Mathematica.

REFERENCES

1. Gordon, Sheldon P., Florence S. Gordon, *et al.* 2004. *Functioning in the Real World: A Precalculus Experience,* 2nd *Ed*. Boston: Addison-Wesley.

2. Hughes-Hallett, Deborah, Andrew Gleason, *et al.* 2002. *Calculus, 3rd Ed*. New York: John Wiley & Sons, New York.

3. Hughes-Hallett, Deborah, Andrew Gleason, *et al.* 1999. *Applied Calculus*, New York: John Wiley & Sons.

4. Gordon, Sheldon P. and Florence S. Gordon. 2003. A Spoonful of Medicine Makes the Mathematics Go Down. *The AMATYC Review*. 24: 9-24.

5. http://www.pfizer.com/download/uspi_viagra.pdf

APPENDIX A: PERFORMING DIRECT CURVE FITTING IN MATHEMATICA

Direct least squares fitting of surge function to data.

```
data = {{.05, 1}, {.4, 50}, {.6, 320}, {1.2, 440}, {1.8, 410},
       {2.1, 350}, {3, 250}, {4, 170}, {6, 80}, {10, 30}, {12, 20},
       {18, 12}, {24, 6}}
m[x_, a_, p_, b_] = a x^p Exp [-b x]
ss[a_, p_, b_] = Sum [(data[[i, 2]] - m[data[[i, 1]], a, p, b])^2,
       {i,1,Length[data]} ]
sm = FindMinimum[ss[a, p, b], {a, 1}, {b, 1}, {p, 2}]
             {23851.6, {a → 1133.57, b → 1.10895.  p → 1.69499}}
mf[x_] = m[x,a,p,b]/.FindMinimum[ss[a,p,b],{a,1},{b, 1},{p,1}][[2]]
             1133.57 e^(-1.10895 x) x^1.69499
```

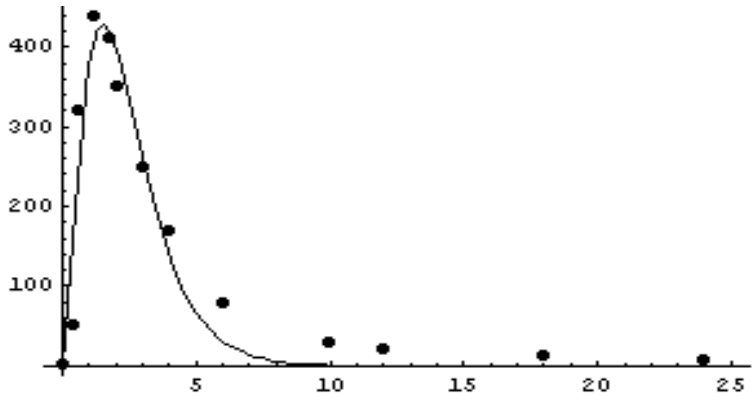

Figure 8. Plot, in Mathematica, of data with fitted curve.

APPENDIX B: PERFORMING MULTIVARIATE REGRESSION IN EXCEL

With Excel's Analysis ToolPak installed, enter the data values for the dependent variable C in Column A, say, and those for the dependent variable t in Column B and then create lists of values for $\log t$ in Column C and $\log C$ in Column D, as is done in Table 2. Then click on Tools, followed by Data Analysis, and finally Regression and OK. This will bring up the Excel dialog box shown in Figure 9. In this dialog, the first box asks you to Input Y Range; the Y-values are the values for the desired dependent variable, here $\log C$, which are in Column D. The second box asks you to Input X Range; the X-values for t and $\log t$ are in Columns B and C. Next, select the first option, Output Range, under Output options; this will give the cells in which all the regression analysis output will appear. Select a collection of cells that are empty, say from A20 to I47.

When you click on OK, Excel will perform the complete regression analysis and print the results in the cells that were indicated. Sample results are shown in Figure 10. Of all the output results, the only ones that are of significance to this discussion are the values for the regression coefficients in Rows 36-38 and possibly the value for the coefficient of multiple determination R in Row 23. In particular, the constant coefficient is 2.428, the coefficient of the first independent variable t is -0.0920 and the coefficient of the second independent variable $\log t$ is 0.2251, leading to the regression equation $Y = 2.428 - -0.0920X_1 + 0.2251X_2$, which is equivalent to $\log C = 2.428 - 0.0920t + 0.2251 \log t$.

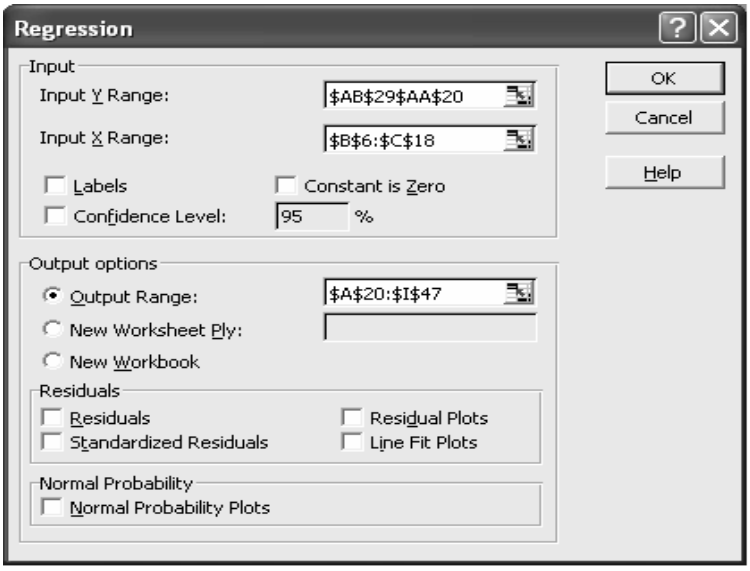

Figure 9. Excel dialog box for regression analysis.

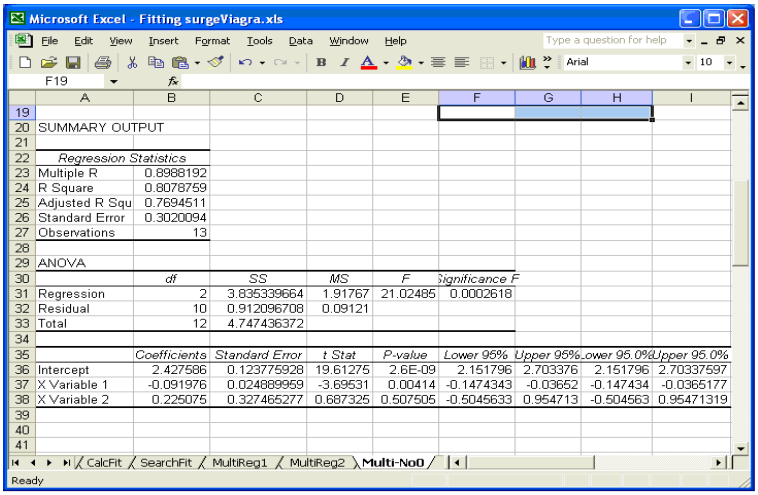

Figure 10. Excel's regression output display.

BIOGRAPHICAL SKETCH

Sheldon Gordon is Professor of Mathematics at Farmingdale State University of New York. He is a member of a number of national committees involved in undergraduate mathematics education and is leading a national initiative to refocus the courses below calculus. He is the principal author of *Functioning in the Real World* and a co-author of the texts developed under the Harvard Calculus Consortium.