Least-Squares Data Fitting

By Dave Slomer

Have you ever had a physics lab in which you gathered data points that were supposed to lie on a line (or a parabola or whatever) only to find that they just didn't? (In fact, has your data <u>ever</u> fit what it was supposed to in an experiment? Don't worry—it's an imprecise world!) The data points in an experiment might be those in Table 1, which come very <u>close</u> to lying on the parabola $f(x) = x^2$.

x	1	2	3	4
y	1	4	9	15

Is $f(x) = x^2$ the best model for the data? Maybe. And maybe $f(x) = .9x^2$. Who's to say?

If you put that data into two **lists** in your TI-89, it can "fit" various *regression models* to that data. There are 10 different regression models, including **QuadReg**, short for **Quad**ratic **Reg**ression. The **QuadReg** model returned for the given data is $f(x) = .75x^2 + .95x - .75$. As you can see in figures 1a and 1b, the model looks like a very good fit, both graphically and analytically, since a perfect fit would be represented by an $\mathbf{R^2}$ value of exactly 1.

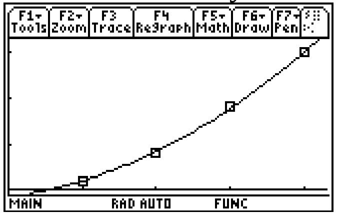
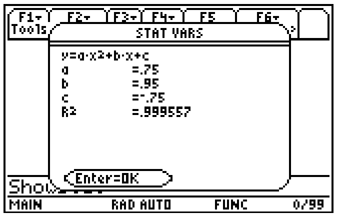


Fig. 1a



```
STAT VARS
y=a·x²+b·x+c
a          =.75
b          =.95
c          =⁻.75
R²         =.999557
```

Fig. 1b

But what if you "knew" that the data was supposed to fit a more restricted parabola, a model of the form $f(x) = ax^2$? (Suppose you had measured times for an object falling from various heights on the moon to determine its gravitational constant *a* using the model $h = at^2$.) There is a valuable mathematical technique for finding a "best" equation for a given set of data, because the technique gives <u>you</u> the choice of models. This technique is the "**Method of Least Squares**".

The Method's process applies the optimization (max-min) theory of calculus to the <u>sum</u> of the **squares** of the <u>vertical</u> distances from each given point to the "given" curve (the model). That sum is then *minimized*. This explains the use of "**least**" in the Method's name. "Least" is also what "best" really means in the previous paragraphs. It's easier to do than the process sounds, especially if you make a sketch, such as the one in figure 2.

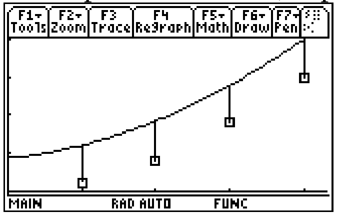


Fig. 2

- First, *plot* each given point and label it.

- Next, represent the model, $f(x) = ax^2$, by drawing a curve that *noticeably misses* the given points (so you can see the vertical distances involved in the minimization).
- Now, draw *vertical* lines from each given point to the curve.
- Then, find the *y-coordinates* of the points on the curve at the *given x-coordinates*. Express these in terms of *a*, since we don't yet know what *a* is.
- Finally, analytically construct the sum. Its format will be as follows, for our four points:

$$\boxed{\textit{Equation 1:}\quad S = (\qquad\qquad\qquad)^2 + (\qquad\qquad\qquad)^2 + (\qquad\qquad\qquad)^2 + (\qquad\qquad\qquad)^2}$$

(Inside each set of parentheses should be the vertical distance from a given data point to the corresponding point on the curve that we are fitting the data to. There are 4 terms, one for each given data point.)

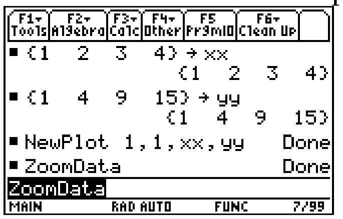
Maybe the following is obvious, maybe not:

- The sum, *S*, will be a function of *a* alone.
- Since we want to minimize that sum, we will solve the equation $\frac{dS}{da} = 0$. (Why?)
- For the value of *a* that we find, we might want to be sure that $\frac{d^2S}{da^2} > 0$ (Why?).
- We should graph *S* vs. *a* to see what is really going on in the Least Squares process.
- We must graph *f* to see how well it fits the data—a "reality check".

The process can be tedious, especially if you have a lot of data, but your '89 will help.

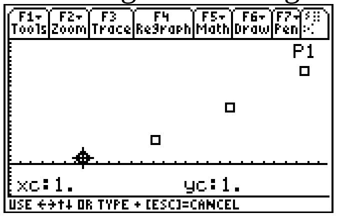
Exercise 1: Fit the data in Table 1 to $f(x) = ax^2$.

First, press **F6** (**2nd F1**) on the home screen to clear all one-letter variables. Turn off or clear all **Y=** functions. Put the data into 2 lists, **xx** and **yy**, *scatterplot* it, and view the graph. To do so, on the home screen, give the commands **{1,2,3,4}→xx** and then **{1,4,9,15}→yy (note the commas)**. Then give the commands **NewPlot 1,1,xx,yy** and **ZoomData**. Refer to figure 3a. (You may know other ways to accomplish this.) The graph screen will show a scatterplot like the one being traced in figure 3b.



```
■ {1  2  3  4} → xx
                {1  2  3  4}
■ {1  4  9  15} → yy
               {1  4  9  15}
■ NewPlot 1,1,xx,yy     Done
■ ZoomData              Done
ZoomData
MAIN      RAD AUTO    FUNC    7/99
```

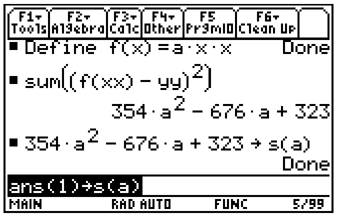
Fig. 3a



```
xc:1.          yc:1.
USE ←→↑↓ OR TYPE + [ESC]=CANCEL
```

Fig. 3b

With the preliminaries finished, begin the analysis. First, define the model, **f(x)**, and the "sum function", **s(a)**. To do so, key in the commands in figure 4. Note the shortcut for the 3rd command, highlighted in the command line.



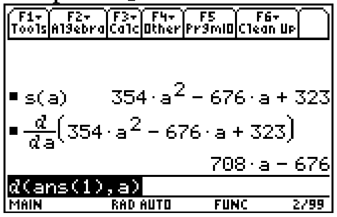
```
■ Define f(x) = a·x·x                Done
■ sum((f(xx) − yy)²)
                   354·a² − 676·a + 323
■ 354·a² − 676·a + 323 → s(a)
                                   Done
ans(1)→s(a)
```

Fig. 4

The ’89 can evaluate a function at every element of a list. So, the command **f(xx)** would return <u>four</u> function values, one for each *x*-coordinate stored in list **xx**. Also, **f(x)–yy** would compute and display <u>four</u> vertical distances. Thus, the **sum** command in figure 4 does a <u>lot</u> of behind-the-scenes algebra. It painlessly computes the nasty sum of squares that you constructed in Equation 1, above.

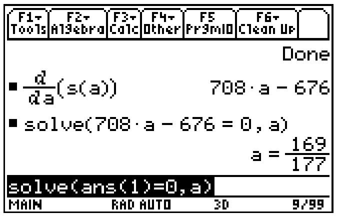
To help put it all together, look at the graph of **s**, the sum-of-squares function that we need to minimize. Turn off the scatterplot [on the **Y=** screen, move the cursor above **y1** to the **Plot1** line and press **F4**] and define **y1 = s(x)** (since the ’89 can only graph functions of **x)**. Because of the enormity of the coefficients (refer to figure 4), don’t expect a pretty graph if you **Zoom Dec**imal or **Zoom St**andar**d**! Tracing in one of those windows shows that **ymax** should be in the hundreds.

Even without a graph of **s**, figure 4 shows that **s** is a concave-up parabola. (Why?) We are looking for the *x*-coordinate of the vertex. (Why?) We could find it via menus on the graph screen, but do it on the home screen, via the commands in figures 5a and 5b. Note the shortcuts highlighted in the command lines. [We do this on the home screen to prepare for Exercises 2 through 5, which deal with regression models with up to <u>six</u> parameters (not just **a**). For those, there will be no graph screen menu help in finding the minimum point.]



```
■ s(a)        354·a² − 676·a + 323
■ d/da(354·a² − 676·a + 323)
                          708·a − 676
d(ans(1),a)
```

Fig. 5a



```
                                 Done
■ d/da(s(a))              708·a − 676
■ solve(708·a − 676 = 0, a)
                            a = 169/177
solve(ans(1)=0,a)
```

Fig 5b

We now know that the sum is minimized when **a** is 169/177. The Least Squares parabola for the data is $\boxed{f(x)=\frac{169}{177}x^2}$.

As a “reality check”, define **y1 = f(x)** and graph. That function, our model for the given data, should *appear* to hit or just barely miss each data point, just as in figure 1a, but it might not actually have exactly hit <u>any</u> (Why?). Of course, the window you are in may have some bearing on these appearances.

To summarize, we:

- **defined f(x)** to be our model, **a*x**$^{\mathbf{2}}$.
- **sum**med the squares of the vertical distances between model and data.
- **stored** that sum into **s(a)**.
- took the **derivative** of **s** with respect to **a**, set it equal to 0, and **solve**d for **a**.

- **graphed** the least-squares function determined by **a**, to make sure it fit the data.

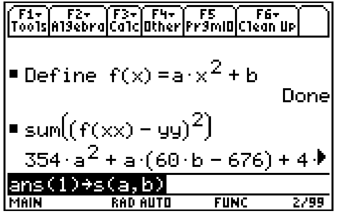
Exercise 2: Fit the data points in Table 1 to the Least Squares parabola of the form $f(x) = ax^2 + b$.

The procedure will be the same until you take the derivative, because there will be <u>two</u> variables in the sum—**a** and **b**—and this is probably new to you. But functions of two variables can be optimized in much the same way as functions of one variable. The concept is called *partial derivatives*, whose symbols (e.g., $\frac{\partial S}{\partial a}$) look a lot like those for normal derivatives (the TI-89 uses normal derivative notation).

To make a long story shorter than it should be, you will be setting both partial derivatives of *S* equal to 0, solving $\frac{\partial S}{\partial a} = 0$ and $\frac{\partial S}{\partial b} = 0$. While based on mathematics beyond the scope of many calculus courses, the mechanics of computing partial derivatives are almost as simple as normal derivatives.

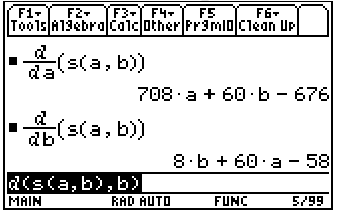
To find $\frac{\partial S}{\partial a}$, for example, "pretend" that *b* is constant and take the derivative of *S* with respect to *a*. Then let *a* be constant and take the derivative of *S* with respect to *b*. You will have a system of *two* linear equations in *two* variables, *a* and *b*, which usually isn't too hard to handle, but the coefficients will be nasty. Your '89 can help.

The entire process is shown below in the home screen commands in figures 6a through 6d, <u>but first clear all the one-letter variables</u> (**F6**). The command line in figure 6c is too long to fit. It says **solve(ans(2)=0 and ans(1)=0 , {a,b})** in TI-89 syntax. In mathematical terms, it says, "Solve the system of equations defined by the <u>last **two**</u> commands [the partial derivative calculations in figure 6b] with respect to **a** and **b**."



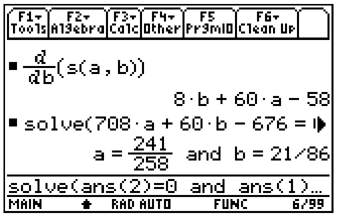
```
■ Define f(x)=a·x^2+b
                              Done
■ sum((f(xx)-yy)^2)
  354·a^2+a·(60·b-676)+4·▸
ans(1)→s(a,b)
```

Fig. 6a



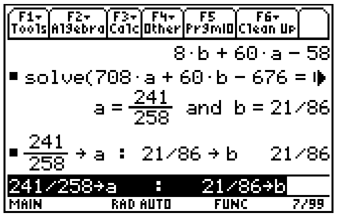
```
■ d/da(s(a,b))
                   708·a+60·b-676
■ d/db(s(a,b))
                      8·b+60·a-58
d(s(a,b),b)
```

Fig. 6b



```
■ d/db(s(a,b))
                      8·b+60·a-58
■ solve(708·a+60·b-676=▸
            a=241/258 and b=21/86
solve(ans(2)=0 and ans(1)…
```

Fig. 6c



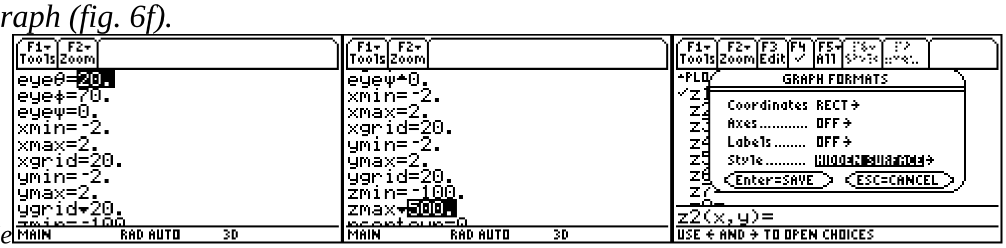
```
                      8·b+60·a-58
■ solve(708·a+60·b-676=▸
            a=241/258 and b=21/86
■ 241/258→a : 21/86→b      21/86
241/258→a  :  21/86→b
```

Fig. 6d

After 7 commands, we have a solution: $\boxed{f(x)=\frac{241}{258}x^2+\frac{21}{86}}$ is the Least Squares parabola of the form $f(x)=ax^2+b$ that fits the data. [The TI-89 is a *<u>fine</u>* piece of machinery, isn't it? (Do the process by <u>hand</u> to <u>really</u> appreciate what it did for you!)]

(After doing some calculus that may be beyond the scope of your calculus class, you really need to look at a graph to SEE that it all DOES make sense! Make **y1=f(x)**, turn back on the scatterplot, and view the graph. You should see a "good fit" to the data, similar to Figure 1a.)

To try to see that the partial derivatives of a function of two variables just might actually find some sort of minimum on the function's graph, set **MODE** *for* **Graph** *to* **3D**, *then define* **z1=s(x,y)**, *set the window and* **Graph Format** (**F1 8**) *as in figure 6e, and look at the graph (fig. 6f).*

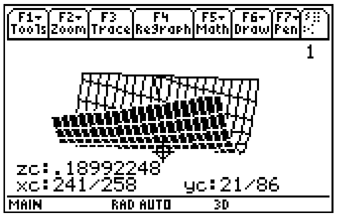


```
eyeθ=20.
eyeφ=70.
eyeψ=0.
xmin=-2.
xmax=2.
xgrid=20.
ymin=-2.
ymax=2.
ygrid=20.
zmin=-100.
```

```
eyeψ=0.
xmin=-2.
xmax=2.
xgrid=20.
ymin=-2.
ymax=2.
ygrid=20.
zmin=-100.
zmax=500.
ncontour=0
```

```
Coordinates RECT →
Axes........ OFF →
Labels...... OFF →
Style....... HIDDEN SURFACE →
z2(x,y)=
```

Fig. 6e



```
zc:.18992248
xc:241/258     yc:21/86
```

Fig. 6f

It sure <u>looks</u> like (**a**,**b**) *<u>could</u> be a low point (minimum) on the graph of* **z1**, *doesn't it? And don't the equation, (partial) derivatives, and graph of* **z1** *somewhat resemble those of parabolas, which, in two dimensions, have unique maxima or minima? Least Squares <u>does</u> work in* **3D**, *too!*

Exercise 3: Repeat the process for the parabola $f(x)=ax^2+bx+c$. This time, you will have *<u>three</u>* equations in *<u>three</u>* variables.

Similar to Exercise 2, after finding the 3 partial derivatives, could give the command **solve(ans(1)=0 and ans(2)=0 and ans(3)=0 , {a,b,c})** to find **a**, **b**, and **c**. Did you get what the '89 did in the second paragraph of this activity? Guess the '89 <u>does</u> use the Least Squares Method for some of its regression models!

Exercise 4: Visit the internet at http://lib.stat.cmu.edu/DASL/Datafiles/appliancedat.html, one of the TI-Interactive data pages, which contains data for appliance sales [in thousands] between 1960 and 1985. Put the data into 2 lists on your '89. Then use either the built-in regression models to find one that fits best and devise your own model.

Your own model may as well be something the ’89 can’t do, such as a 5th degree polynomial—most of that data wiggles at <u>least</u> that much. Refer to figure 8 for a scatterplot of the dishwasher data, where the thick graph is that of the Least Squares **<u>5th</u>** degree polynomial ($f(x) = ax^5 + bx^4 + cx^3 + dx^2 + ex + g$) and the thin graph is that of the ‘89’s **QuartReg** model (4th degree). Which looks like a better fit?

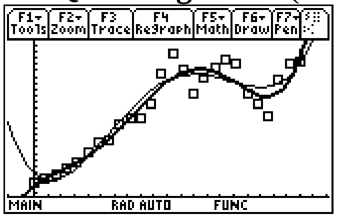


Fig. 8

(If you have access to TI-Interactive software and a TI GraphLink cable, you can go to the web page mentioned earlier and load the data into your calculator, but you may prefer to just type it in.)

Since the data only goes up to 1985, see if you think your model (or the ‘89’s, for that matter) can come close to predicting the number of dishwashers sold in 2000. To do so, you would compute **f(40)** and/or **RegEq(40)** if you choose to make 1960 “year 0”, as was done below. Do you think either answer is reasonable? Why? Which model (if either) do you think would give the best estimate of sales for 1986? 1990? 1959? 1955?

This process of trying to predict results <u>outside</u> the range of data is called ***extra**polation*. Write a paragraph or two about *extrapolation* based on your findings in the previous paragraph.

Exercise 5: “Lose” 6 years’ data from the appliance data, such as from 1970 ***<u>through</u>*** 1975. Fit the same 5th degree model to the data and see how well the new model can predict what the sales were in each of the “lost” 6 years. This process of trying to predict results <u>inside</u> the range of data is called ***inter**polation*. Write a paragraph or two about *interpolation*, comparing and contrasting with *extrapolation*.

Calculus Generic Scope and Sequence Topics: Applications of Derivatives

NCTM Standards: Number and operations, Algebra, Geometry, Problem solving, Connections, Communication, Representation