DERIVING THE QUADRATIC REGRESSION EQUATION USING ALGEBRA

Sheldon P. Gordon
Department of Mathematics
Farmingdale State University of New York
2350 Broadhollow Road
Farmingdale, New York 11735
gordonsp@farmingdale.edu

Florence S. Gordon
Department of Mathematics
New York Institute of Technology
Old Westbury, New York 11768
fgordon@nyit.edu

In discussions with leading educators from many different fields [1], MAA's CRAFTY (Curriculum Renewal Across the First Two Years) committee found that one of the most common mathematical themes in those other disciplines is the idea of fitting a function to a set of data in the least squares sense. The representatives of those partner disciplines strongly recommended that this topic receive considerably more attention in mathematics courses to develop much stronger links between mathematics and the way it is used in the other fields.

This notion of curve fitting is one of the most powerful "new" mathematical topics introduced into courses in college algebra and precalculus over the last few years. But it is not just a matter of providing the students with one of the most powerful mathematics topics they will encounter elsewhere. Curve fitting also provides the opportunity to stress fundamental ideas about linear, exponential, power, logarithmic, polynomial, sinusoidal, and logistic functions to situations from all walks of life. From a pedagogical perspective, these applications provide additional ways to reinforce the key properties of each family of functions. From the students' perspective, they see how the mathematics being learned has direct application in modeling an incredibly wide variety of situations in almost any field of endeavor, making the mathematics an important component of their overall education. Fortunately, these capabilities are built into all graphing calculators and spreadsheets such as *Excel*, so it becomes very simple to incorporate the methods into courses at this level.

However, many mathematicians naturally feel uncomfortable with simply introducing these techniques into their courses just because the

available technology has the capability. In their minds, the techniques become equivalent to the students using the technology blindly without any mathematical understanding of what is happening. Typically in regression analysis, multivariable calculus is used to minimize the sum of the squares of the deviations between the data values and the equation of the mathematical model, but these derivations are inaccessible to precalculus students. An example of a treatment using calculus appears in [2].

In a previous article [3], the authors presented an algebra-based derivation of the regression equations for the best fit line $y = ax + b$ that simultaneously reinforces other concepts that one would treat in college algebra and precalculus when discussing properties of quadratic functions. In the present article, we extend the approach used in [3] to develop an algebra-based derivation of the regression equations that fit a quadratic function to a set of data. Also we indicate how this approach can be extended and broadened to encompass the notion of multivariable linear regression.

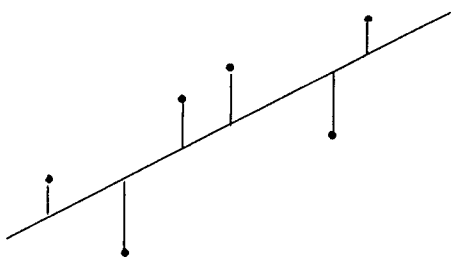

Figure 1

DERIVING THE LINEAR REGRESSION EQUATIONS

The least squares criterion used to create the regression line $y = ax + b$ that fits a set of n data points $(x_1, y_1), (x_2, y_2), \ldots (x_n, y_n)$ is that the sum of the squares of the vertical distances from the points to the line be a minimum. See Figure 1. This means that we need to find those values of a and b for which $S = \sum_{i=1}^{n} \left[y_i - (ax_i + b) \right]^2$ is minimized. In a comparable way, the least squares criterion can be applied to create the quadratic regression function $y = ax^2 + bx + c$ that fits a set of n data points $(x_1, y_1), (x_2, y_2), \ldots (x_n, y_n)$ by finding the values of the three parameters a, b, and c that minimize

$$S = \sum_{i=1}^{n} \left[y_i - \left(ax_i^2 + bx_i + c \right) \right]^2 . \qquad \textbf{(1)}$$

To simplify the notation, we omit the indices in the summation symbol.

First we consider the sum of the squares as a function of a. We can rewrite Equation (1) as $S=\sum\left[ax_i^2+\left(bx_i+c-y_i\right)\right]^2$ and expand it to obtain $S=\sum\left[a^2x_i^4+2ax_i^2\left(bx_i+c-y_i\right)+\left(bx_i+c-y_i\right)^2\right]$. Now rewrite this expression by distributing the summation and use the fact that a does not depend on the index of summation i so that it can be factored out. We get

$$\begin{aligned}S&=a^2\sum x_i^4+2a\sum x_i^2\left(bx_i+c-y_i\right)+\sum\left(bx_i+c-y_i\right)^2\\&=a^2\sum x_i^4+2a\sum\left(bx_i^3+cx_i^2-x_i^2y_i\right)+\sum\left(bx_i+c-y_i\right)^2.\end{aligned}$$

Therefore, S can be thought of as a quadratic function of a. The coefficient of a^2 is the sum of the fourth powers of the x's, so it is a positive leading coefficient and the corresponding parabola with S as a quadratic function of a opens upward. Thus, it achieves its minimum at the vertex.

We now use the fact that the vertex of any parabola $Y=AX^2+BX+C$ occurs at $X=\frac{-B}{2A}$. (We use capital letters here because a, b, c, x, and y have different meanings in the present context.) Thus, the minimum value for S occurs at

$$a=\frac{-2\sum\left(bx_i^3+cx_i^2-x_i^2y_i\right)}{2\sum x_i^4}.$$

We cross-multiply to obtain $a\sum x_i^4=-\sum\left(bx_i^3+cx_i^2-x_i^2y_i\right)$. Since b and c do not depend on the index of summation i, we have

$$a\sum x_i^4+b\sum x_i^3+c\sum x_i^2=\sum x_i^2y_i. \qquad (2)$$

Once the x_i and y_i values are given in the data points, all of the summations are just sums of known numbers and are constant. Thus, this is a linear equation in the three unknown parameters a, b, and c.

Now, look again at Equation (1) for S. We can think of it as a quadratic function of b and perform a similar analysis. When we re-arrange the terms, we get

$$\begin{aligned}S&=\sum\left[bx_i+\left(ax_i^2+c-y_i\right)\right]^2\\&=\sum\left[b^2x_i^2+2bx_i\left(ax_i^2+c-y_i\right)+\left(ax_i^2+c-y_i\right)^2\right]\\&=b^2\sum x_i^2+2b\sum\left(ax_i^3+cx_i-x_iy_i\right)+\sum\left(ax_i^2+c-y_i\right)^2.\end{aligned}$$

We can think of S as a quadratic function of b. Again, the leading coefficient is positive, so the parabola opens upward and therefore achieves its minimum at

$$b = \frac{-2\sum\left(ax_i^3 + cx_i - x_i y_i\right)}{2\sum x_i^2}.$$

We cross-multiply and collect terms to obtain

$$a\sum x_i^3 + b\sum x_i^2 + c\sum x_i = \sum x_i y_i \ . \qquad \textbf{(3)}$$

Notice that this is another linear function of a, b, and c since all of the sums are known quantities.

Finally, we look at Equation (1) for a third time and think of S as a function of c, so that

$$\begin{aligned} S &= \sum\left[c + \left(ax_i^2 + bx_i - y_i\right)\right]^2 \\ &= \sum\left[c^2 + 2c\left(ax_i^2 + bx_i - y_i\right) + \left(ax_i^2 + bx_i - y_i\right)^2\right] \\ &= c^2\sum 1 + 2c\sum\left(ax_i^2 + bx_i - y_i\right) + \sum\left(ax_i^2 + bx_i - y_i\right)^2. \end{aligned}$$

Because summing the number 1 n times yields n, this equation is equivalent to

$$S = nc^2 + 2c\sum\left(ax_i^2 + bx_i - y_i\right) + \sum\left(ax_i^2 + bx_i - y_i\right)^2.$$

We can think of S as a quadratic function of c with a positive leading coefficient. Consequently, the parabola opens upward and has its minimum at

$$c = \frac{-2\sum\left(ax_i^2 + bx_i - y_i\right)}{2n}.$$

We cross-multiply and collect terms to obtain

$$a\sum x_i^2 + b\sum x_i + cn = \sum y_i, \qquad \textbf{(4)}$$

which is the third linear function of a, b, and c.

The three equations (2)-(4) in three unknowns

$$\begin{aligned} a\sum x_i^4 + b\sum x_i^3 + c\sum x_i^2 &= \sum x_i^2 y_i \\ a\sum x_i^3 + b\sum x_i^2 + c\sum x_i &= \sum x_i y_i \\ a\sum x_i^2 + b\sum x_i + cn &= \sum y_i. \end{aligned}$$

are called the *normal equations*. Their solution gives the coefficients a, b, and c in the equation of the quadratic regression function.

WHEN DOES THE SOLUTION BREAK DOWN?

Before going on, let's consider the conditions under which this system of equations does not have either a unique solution or any solution. This occurs when the coefficient matrix is singular, and hence its determinant is zero. Thus,

$$\begin{vmatrix} \sum x_i^4 & \sum x_i^3 & \sum x_i^2 \\ \sum x_i^3 & \sum x_i^2 & \sum x_i \\ \sum x_i^2 & \sum x_i & n \end{vmatrix} = 0$$

For simplicity, let's investigate when this occurs with $n = 3$ points. We substitute $\sum x_i = x_1 + x_2 + x_3$, $\sum x_i^2 = x_1^2 + x_2^2 + x_3^2$, and so forth into the above expression. Using Derive to expand and factor the resulting expression, we find that the determinant is equal to

$$(x_1 - x_2)^2 (x_1 - x_3)^2 (x_2 - x_3)^2 = 0 .$$

Thus, the system does not have a unique solution if $x_1 = x_2$, $x_1 = x_3$, or $x_2 = x_3$; that is, if any two or more of the three points are in a vertical line.

Next, let's consider what happens with $n = 4$ points. The resulting expression for the determinant of the coefficient matrix can be reduced to

$$-(x_1 - x_2)^2 (x_1 - x_3)^2 (x_2 - x_3)^2 - (x_1 - x_2)^2 (x_1 - x_4)^2 (x_2 - x_4)^2$$
$$-(x_2 - x_3)^2 (x_2 - x_4)^2 (x_3 - x_4)^2 - (x_1 - x_3)^2 (x_3 - x_4)^2 (x_1 - x_4)^2 .$$

This expression is equal to zero when $x_1 = x_2 = x_3$, or $x_1 = x_3 = x_4$, or $x_2 = x_3 = x_4$, or $x_1 = x_2 = x_4$, or $x_1 = x_2$, and $x_3 = x_4$, or $x_1 = x_3$ and $x_2 = x_4$, or $x_1 = x_4$, and $x_2 = x_3$. That is, the system of normal equations does not have a unique solution if three of the four points are on a vertical line or if two pairs of the points are on vertical lines. In other words, there must be at least three distinct values of the independent variable. Presumably, the same condition will prevail if there are five or more points; however, neither the authors nor Derive were able to perform the corresponding algebraic manipulations.

EXAMPLE OF USING THE REGRESSION EQUATIONS

We illustrate the use of the three normal equations in the following example.

Example. Find the equation of the quadratic function that fits the points

$(-2,5)$, $(-1,-1)$, $(0,-3)$, $(1,-1)$, $(2,5)$, and $(3,15)$ by solving the normal equations. (Note that these points all lie on the parabola $y=2x^2-3$, so we have a simple target to aim for.)

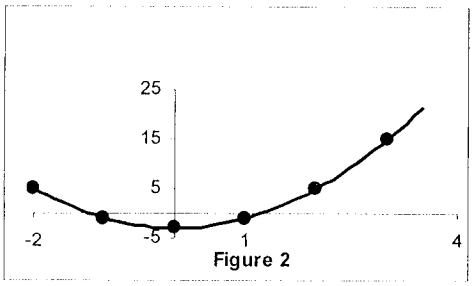

Figure 2

We begin with the scatterplot of the data with the parabola superimposed, as shown in Figure 2. We calculate the various sums needed in the three normal equations by completing the entries in the accompanying table.

x	-2	-1	0	1	2	3	$\sum x = 3$
y	5	-1	-3	-1	5	15	$\sum y = 20$
x^2	4	1	0	1	4	9	$\sum x^2 = 19$
x^3	-8	-1	0	1	8	27	$\sum x^3 = 27$
x^4	16	1	0	1	16	81	$\sum x^4 = 115$
xy	-10	1	0	-1	10	45	$\sum xy = 45$
x^2y	20	-1	0	-1	20	135	$\sum x^2y = 173$

The coefficients a, b, and c are the solutions of the normal equations (2), (3), and (4). We substitute the values for the various sums and set $n=6$ to get the system of equations.

$$115a+27b+19c=173$$
$$27a+19b+3c=45$$
$$19a+3b+6c=20.$$

As expected the solution to this system is $a=2$, $b=0$ and $c=-3$, so that the corresponding quadratic function is $y=2x^2-3$. The students may like to see that the regression features of the calculator or a software package such as *Excel* give the same results.

We note that the derivation shown above for the normal equations for the quadratic regression function can obviously be extended to derive similar sets of normal equations for higher degree regression polynomials.

A MORE GENERAL CONTEXT

We can look at these ideas in a somewhat different context that is a special case of a much more general approach. Instead of considering fitting

a function $y = f(x)$ of one variable to a set of data, we can alternatively think of fitting a linear function of several variables

$$Y = A_1X_1 + A_2X_2 + \ldots + A_kX_k + B$$

to a set of multivariate data $(x_{1i}, x_{2i}, \ldots, x_{ki}, y_i)$, for $i = 1$ to n. Geometrically, we think of this as fitting a hyperplane in $k+1$ dimensions to a set of n points in that space via the least squares criterion. Suppose that we seek to fit a quadratic function $y = ax^2 + bx + c$ to a set of bivariate data. Instead of thinking of y as a *quadratic* function of x, we can think of y as a *linear* function of the two variables x and x^2. It turns out that the three normal equations that we derived above for the quadratic fit are precisely the same set of three equations in three unknowns that would result from fitting a multivariate linear function to the data. The same notions extend to higher degree polynomial fits to a set of bivariate data.

Acknowledgement The work described in this article was supported by the Division of Undergraduate Education of the National Science Foundation under grants DUE-0089400 and DUE-0310123. However, the views expressed are not necessarily those of either the Foundation or the project.

REFERENCES

1. W. Barker and S. Ganter, Voices of the Partner Disciplines, MAA Reports, Mathematical Association of America, Washington, DC. (2004).
2. S. M. Scariano, and W. Barnett II, "Contrasting Total Least Squares with Ordinary Least Squares - Part I: Basic Ideas and Results", *Mathematics and Computer Education*, Vol. 37, No. 2, pp. 141-158 (Spring 2003).
3. S. P. Gordon and F. S. Gordon, "Deriving the Regression Equations Without Using Calculus", *Mathematics and Computer Education*, Vol. 38, No. 1, pp. 64-68 (Winter 2004).