

## Least Squares

In applications discussed so far, consistent linear systems have provided an important mathematical model. We previously discussed nonsingular linear systems that have a unique solution. We also briefly discussed **underdetermined linear systems**; that is, systems which have infinitely many solutions. In each of these cases the linear system  $\mathbf{Ax} = \mathbf{b}$  is consistent, or equivalently,  $\mathbf{b}$  is in the column space of  $\mathbf{A}$ . That is,  $\mathbf{b}$  is a linear combination of the columns of  $\mathbf{A}$ .

It would be convenient if we could ignore all inconsistent systems; however, we cannot. Inconsistent systems do occur, and we must deal with them. In order to accomplish this we need to alter our notion that a solution to a matrix problem satisfies the equality requirement that  $\mathbf{A}$  'times'  $\mathbf{x}$  gives  $\mathbf{b}$ . We begin this as follows.

Consider the case of an **overdetermined linear system**; that is, a linear system of equations in which there are more equations than unknowns. Such linear systems arise naturally from experiments or collections of data which consist of a large number of observations that are used to estimate a few unknowns in a mathematical model. Examples include computing the orbit of a satellite or path of a projectile, determining rate constants of various types, and, in general, calculating coefficients in a proposed model of a physical phenomenon or process. Since errors will invariably be included in such observational data, it is expected that such overdetermined linear systems will be inconsistent. Whether the resulting linear system is inconsistent or not, values of unknown coefficients of the model are required. If the linear system  $\mathbf{Ax} = \mathbf{b}$  is inconsistent, then the use of row operations will fail to yield a result. In such cases an alternative is to seek a vector  $\mathbf{z}$  so that the product  $\mathbf{Az}$  is as close to the right side  $\mathbf{b}$  as possible. Since  $\mathbf{Az}$  is a linear combination of the columns of  $\mathbf{A}$  and hence in the column space of  $\mathbf{A}$ , denoted  $\mathbf{col}(\mathbf{A})$ , we can rephrase the situation as follows:

**Determine the vector in  $\mathbf{col}(\mathbf{A})$  that is closest to  $\mathbf{b}$ .**

This implies that we must solve a minimization problem; namely, determine the minimum distance from  $\mathbf{b}$  to the subspace  $\mathbf{col}(\mathbf{A})$ . Note that if the linear system is consistent, then the standard solution  $\mathbf{x}$  of  $\mathbf{Ax} = \mathbf{b}$  is in  $\mathbf{col}(\mathbf{A})$  and hence the distance between  $\mathbf{b}$  and  $\mathbf{col}(\mathbf{A})$  is zero. Thus the minimization problem includes the solution of consistent linear systems as a special case.

A matrix approach to solving this minimization problem requires concepts we have not yet developed, but an alternative is to formulate the minimization problem in terms of calculus. Here we consider the special case in which there are just two unknown coefficients in the linear system. We use a geometric model to construct the appropriate minimization statement and then show how to reformulate the result in terms of matrices.

## Geometric Model (Special Case)

Let  $D = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$  be a data set in which the points are distinct in the sense that all the x-coordinates are different from one another. We say there is a 'linear' relationship between the x and y coordinates provided there is some line  $y = mx + b$  that either goes through each point  $(x_i, y_i)$  or comes close to all the ordered pairs in  $D$ . If each point of  $D$  lies on the same line then the values of  $m$  and  $b$  can be determined from the solution of a nonsingular linear system. If there is no line that goes through all the points in  $D$ , then an inconsistent linear system arises in trying to determine  $m$  and  $b$ . In either case we have the linear system in the unknowns  $m$  and  $b$  given by

$$\begin{array}{l} mx_1 + b = y_1 \\ mx_2 + b = y_2 \\ \vdots \\ \vdots \\ mx_n + b = y_n \end{array} \quad \text{or} \quad \begin{bmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \vdots \\ \vdots & \vdots \\ x_n & 1 \end{bmatrix} \begin{bmatrix} m \\ b \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ \vdots \\ y_n \end{bmatrix}. \quad (1)$$

Let  $\mathbf{A} = \begin{bmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \vdots \\ \vdots & \vdots \\ x_n & 1 \end{bmatrix}$ ,  $\mathbf{x} = \begin{bmatrix} m \\ b \end{bmatrix}$ , and  $\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ \vdots \\ y_n \end{bmatrix}$ . Then the linear system in (1) is  $\mathbf{Ax} = \mathbf{y}$ .

Here we investigate the case in which  $\mathbf{Ax} = \mathbf{y}$  is inconsistent. Thus geometrically, the data in  $D$  is a set of noncollinear points as shown in Figure 1.

The line that comes closest to all the data in  $D$  is called the **line of 'best fit'**. The technique to determine the line of best fit is known as the **method of least squares** since [we adopt the criterion that we want to minimize the sum of the squares of the vertical distance from a data point to the](#)

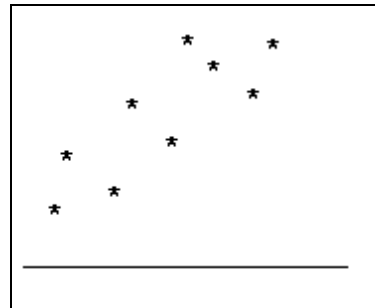


Figure 1.

[line  \$y = mx + b\$ .](#) The vertical distance, often called the **deviation**, from a point  $(x_i, y_i)$  to the line  $y = mx + b$  is given by the expression  $(mx_i + b - y_i)$ , which is just the difference of the y-coordinates of the data point and the point on the line when  $x = x_i$ . Computing this difference for each point in

D, squaring the quantities, and adding them gives the expression  $E(m,b)$  shown next:

$$E(m,b) = (mx_1 + b - y_1)^2 + (mx_2 + b - y_2)^2 + \dots + (mx_n + b - y_n)^2$$

This expression is called the **sum of the squares of the deviations**. The line of best fit is obtained by determining the values of  $m$  and  $b$  that minimize the sum of the squares of the deviations given in the expression  $E(m,b)$ .

Using summation notation we have

$$E(m,b) = \sum_{i=1}^n (mx_i + b - y_i)^2 .$$

To use calculus to obtain the values that minimize  $E(m,b)$  we proceed as follows. Compute the partial derivative of  $E(m,b)$  with respect to  $m$  and the partial with respect to  $b$ , set them equal to zero and solve for  $m$  and  $b$ . We obtain

$$\frac{\partial E(m,b)}{\partial m} = 2 \sum_{i=1}^n (mx_i + b - y_i) \cdot x_i = 0$$

$$\frac{\partial E(m,b)}{\partial b} = 2 \sum_{i=1}^n (mx_i + b - y_i) = 0$$

This gives us two equations in the unknowns  $m$  and  $b$  which can be simplified and rearranged into the following form:

$$\begin{aligned} m \sum_{i=1}^n x_i^2 + b \sum_{i=1}^n x_i &= \sum_{i=1}^n x_i y_i \\ m \sum_{i=1}^n x_i + b \sum_{i=1}^n 1 &= \sum_{i=1}^n y_i \end{aligned}$$

Note that  $\sum_{i=1}^n 1 = n$ , thus this linear system in matrix form is given by

$$\begin{bmatrix} \sum_{i=1}^n x_i^2 & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & n \end{bmatrix} \begin{bmatrix} m \\ b \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^n x_i y_i \\ \sum_{i=1}^n y_i \end{bmatrix} . \quad (2)$$

Let  $\mathbf{C} = \begin{bmatrix} \sum_{i=1}^n x_i^2 & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & n \end{bmatrix}$ ,  $\mathbf{d} = \begin{bmatrix} \sum_{i=1}^n x_i y_i \\ \sum_{i=1}^n y_i \end{bmatrix}$ . Then (2) is given by matrix equation

$\mathbf{C}\mathbf{x} = \mathbf{d}$ . The matrix  $\mathbf{C}$  is nonsingular (See Exercise 7.) so (2) has a unique solution. The resulting values of  $m$  and  $b$  are, respectively, the slope and  $y$ -intercept of the line of best fit. The linear system  $\mathbf{C}\mathbf{x} = \mathbf{d}$  is called the **normal system of equations** for the line of best fit. **When we want to compute the line of best fit for a particular data set we can immediately construct the normal system of equations and then find its solution.** There is no need to repeat the minimization steps employed to obtain (2). We illustrate the technique in Example 1.

**Example 1.** Various airlines publish a table showing how the temperature (in °F) outside an airplane changes as the altitude (in 1000 feet) changes. This data is given in Table 1. Determine the line of best fit to this data set.

$x$ (Altitude in 1000's)	1	5	10	15	20	30	36
$y$ (Temperature in °F)	56	41	23	5	-15	-47	-69

Table 1.

Using (2) we construct the normal system of equations:

$$n = 7$$

$$\sum_{i=1}^7 x_i = 1 + 5 + 10 + 15 + 20 + 30 + 36 = 117$$

$$\sum_{i=1}^7 x_i^2 = 1^2 + 5^2 + 10^2 + 15^2 + 20^2 + 30^2 + 36^2 = 2947$$

$$\sum_{i=1}^7 y_i = 56 + 41 + 23 + 5 - 15 - 47 - 69 = -6$$

$$\sum_{i=1}^7 x_i y_i = (1)(56) + (5)(41) + (10)(23) + (15)(5) + (20)(-15) + (30)(-47) + (36)(-69) = -3628$$

$$\begin{bmatrix} \sum_{i=1}^7 x_i^2 & \sum_{i=1}^7 x_i \\ \sum_{i=1}^7 x_i & 7 \end{bmatrix} \begin{bmatrix} m \\ b \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^7 x_i y_i \\ \sum_{i=1}^7 y_i \end{bmatrix} \quad \text{or} \quad \begin{bmatrix} 2947 & 117 \\ 117 & 7 \end{bmatrix} \begin{bmatrix} m \\ b \end{bmatrix} = \begin{bmatrix} -3628 \\ -6 \end{bmatrix}$$

Solving for  $m$  and  $b$  we find that (to four decimal places)  $m = -3.5582$  and  $b = 58.6159$ . Hence the line of best fit is

$$y = -3.5582x + 58.6159 \quad (3)$$

Figure 2 shows both the data from Table 1 and the line of best fit. Note that this line comes very close to all the data, which is a strong indication that there is a linear relationship between altitude and temperature.

The line of best fit can also be used as a mathematical model to estimate either the temperature at a given altitude or the altitude at which a specific temperature occurs. For instance to estimate the temperature at an altitude of 40,000 ft we set  $x = 40$  in (3) to obtain  $y = -3.5582(40) + 58.6159 = -83.7121$ , or approximately  $-84^\circ\text{F}$ . In a similar fashion, if we want to estimate the altitude at which the temperature is  $-30^\circ\text{F}$ , we set  $y = -30$  and solve for  $x$ . We find that

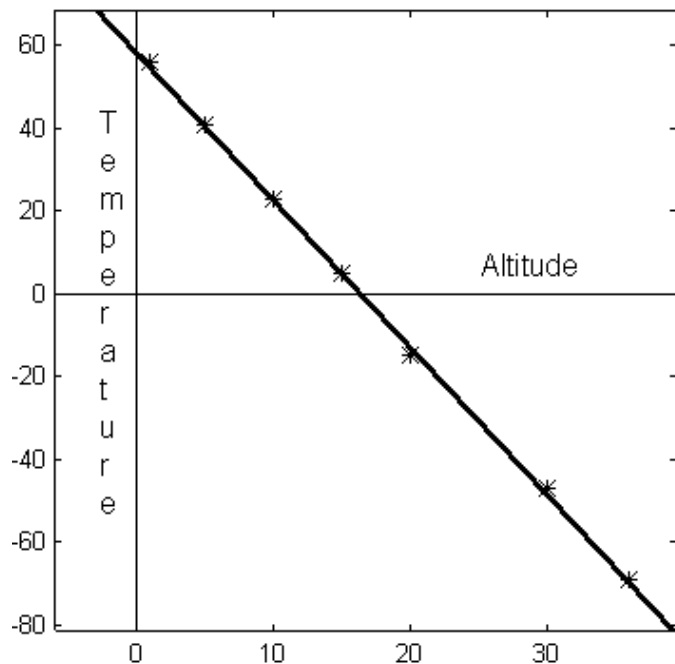


Figure 2.

$$x = \frac{-30 - 58.6159}{-3.5582} = \frac{-88.6159}{-3.5582} \approx 24.9047$$

so a temperature of  $-30^\circ\text{F}$  occurs at approximately 25,000 ft. ■

Another mathematical model can be obtained by interchanging the roles of the  $x$  and  $y$  data. Thus geometrically this would amount to plotting the temperature horizontally and the altitude vertically. The model obtained in this manner is different from that obtained in Example 1.

It is quite possible that a given data set is not well approximated by a line. In such instances the method of least squares<sup>1</sup> can be extended to obtain functions of other shapes that come close to all the points in the set.

### Matrix Computations to Obtain the Normal Equations in (2)

The normal system of equations in (2) was derived using calculus. Here we show how to develop the normal equations directly from the original system of equations

$$\begin{bmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \vdots \\ \vdots & \vdots \\ x_n & 1 \end{bmatrix} \begin{bmatrix} m \\ b \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ \vdots \\ y_n \end{bmatrix} \Leftrightarrow \mathbf{Ax} = \mathbf{y} \quad (4)$$

given in (1). The development we present is not an alternate verification that we have obtained the line closest to all the data in the least squares sense, rather it provides an easy matrix formulation of the normal system of equations. We proceed with a set of observations regarding the entries of the matrix **C** and right side **d** of the normal system of equations. Each of the entries of **C** and **d** can be expressed as a dot product:

$$\sum_{i=1}^n x_i^2 = [x_1 \ x_2 \ \dots \ x_n] \cdot \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \quad \sum_{i=1}^n x_i y_i = [x_1 \ x_2 \ \dots \ x_n] \cdot \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

$$\sum_{i=1}^n x_i = [1 \ 1 \ \dots \ 1] \cdot \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \quad \sum_{i=1}^n y_i = [1 \ 1 \ \dots \ 1] \cdot \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

$$n = [1 \ 1 \ \dots \ 1] \cdot \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$$

Comparing

these expressions to the entries of **C** and **d** and recalling that the entries of the product of a pair of matrices can be expressed as dot products we see that

---

<sup>1</sup> Recall, we have considered only a special case; a line of best fit to a data set.

$$\mathbf{A}^T \mathbf{A} = \begin{bmatrix} \sum_{i=1}^n x_i^2 & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & n \end{bmatrix} = \mathbf{C}, \quad \mathbf{A}^T \mathbf{y} = \begin{bmatrix} \sum_{i=1}^n x_i y_i \\ \sum_{i=1}^n y_i \end{bmatrix} = \mathbf{d}.$$

Hence the normal system of equations is obtained from the original system  $\mathbf{Ax} = \mathbf{y}$  by multiplying both sides by  $\mathbf{A}^T$  to give  $\mathbf{A}^T \mathbf{Ax} = \mathbf{A}^T \mathbf{y}$ . This matrix formulation of the normal system of equations, whose solution leads to the slope and y-intercept of the line of best fit, is useful in computer calculations.

### Least Square Quadratics

To determine a best fit or least squares quadratic  $y = ax^2 + bx + c$  to the set of data  $D = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$  we first construct the linear system corresponding to the equations  $y_i = ax_i^2 + bx_i + c$ . In matrix form this linear system is

$$\begin{bmatrix} x_1^2 & x_1 & 1 \\ x_2^2 & x_2 & 1 \\ \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots \\ x_n^2 & x_n & 1 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ \vdots \\ y_n \end{bmatrix}$$

Let  $\mathbf{A}$  denote the coefficient matrix,  $\mathbf{x}$  denote the column of coefficients  $[a \ b \ c]^T$ , and  $\mathbf{y}$  denote the right side. Then the least squares quadratic is determined from the solution of the normal system of equations given by  $\mathbf{A}^T \mathbf{Ax} = \mathbf{A}^T \mathbf{y}$ . Find the least squares quadratic for the data in the following table.

**Example 2:** Find the least quadratic for the data in the following table.

$x_i$	1	2	3	4	5
$y_i$	4.5	5.1	4.3	2.5	1

The coefficient matrix  $\mathbf{A}$  is

$$\mathbf{A} = \begin{bmatrix} 1^2 & 1 & 1 \\ 2^2 & 2 & 1 \\ 3^2 & 3 & 1 \\ 4^2 & 4 & 1 \\ 5^2 & 5 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 \\ 4 & 2 & 1 \\ 9 & 3 & 1 \\ 16 & 4 & 1 \\ 25 & 5 & 1 \end{bmatrix}$$

The column of unknowns is  $\mathbf{x} = \begin{bmatrix} a \\ b \\ c \end{bmatrix}$  and the right side is  $\mathbf{y} = \begin{bmatrix} 4.5 \\ 5.1 \\ 4.3 \\ 2.5 \\ 1 \end{bmatrix}$ . So we have

over determined system  $\mathbf{Ax} = \mathbf{y}$ . It follows that the normal system of equations is computed as  $\mathbf{A}^T \mathbf{Ax} = \mathbf{A}^T \mathbf{y}$ . We have the normal system

$$\begin{bmatrix} 979 & 225 & 55 \\ 225 & 55 & 15 \\ 55 & 15 & 5 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} 128.6 \\ 42.6 \\ 17.4 \end{bmatrix}$$

whose solution is  $\begin{bmatrix} a \\ b \\ c \end{bmatrix} \approx \begin{bmatrix} -0.3714 \\ 1.2686 \\ 3.7600 \end{bmatrix}$ . Thus the least squares quadratic model for

this data is  $-0.3714x^2 + 1.2686x + 3.76$ . Figure 3 shows the graph of the quadratic model for the data and the data set.

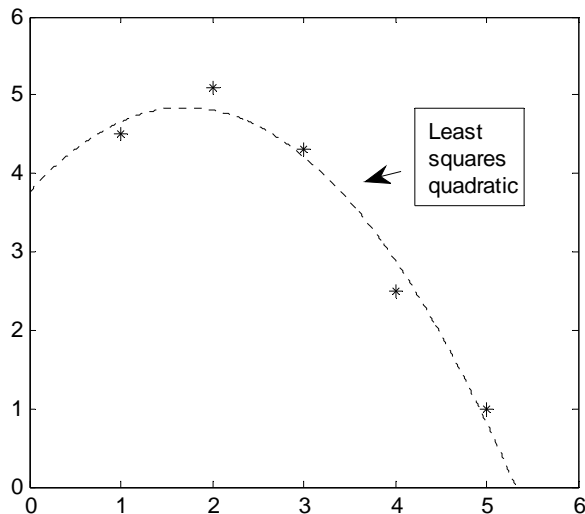


Figure 3.