

## Section 5.8 Regression/Least Squares Approximation

Regression is a powerful technique for predicting the value of a dependent variable and estimating the values of model parameters. The method has wide application.

In interpolation we force the error to be zero at specific isolated locations hence the process can be viewed as a local procedure. In regression the error is minimized in a certain way in a global sense for data involved.

In interpolation we constructed linear systems either directly or indirectly through the use of basis functions that had unique solutions hence coefficient matrices were square and nonsingular. In this case 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  $\mathbf{A}$ .

In regression the model being constructed is over determined meaning the number of coefficients in the model is smaller than the number of data points used. That is, we are dealing with 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 over determined 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  $\text{col}(\mathbf{A})$ , we can rephrase the situation as follows:

**Determine the vector in  $\text{col}(\mathbf{A})$  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  $\text{col}(\mathbf{A})$ . Note that if the linear system is consistent, then the standard solution  $\mathbf{x}$  of  $\mathbf{Ax} = \mathbf{b}$  is in  $\text{col}(\mathbf{A})$  and hence the distance between  $\mathbf{b}$  and  $\text{col}(\mathbf{A})$  is zero. Thus the minimization problem includes the solution of consistent linear systems as a special case.

Since regression requires a minimization problem we can expect calculus to be involved, but in some instances a clever matrix algebra approach can be used.

### Regression/Least Squares Line

If a scatter plot of a data set  $S = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$  looks like it can be well approximated by a straight line  $y = mx + b$  then we want the “best” values of parameters  $m$  and  $b$  so that the error involved is minimized. 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 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  $S$ , squaring the quantities, and adding them gives the expression  $E(m, b)$  shown next:

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

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:

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

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

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

The matrix  $\mathbf{C}$  is nonsingular (See the exercises.) so (1) 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  $\mathbf{Cz} = \mathbf{d}$  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 (1). 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.

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

Table 1.

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

$$y = -3.5582x + 58.6159 \quad (2)$$

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 (2) 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

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

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

Figure 1 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.

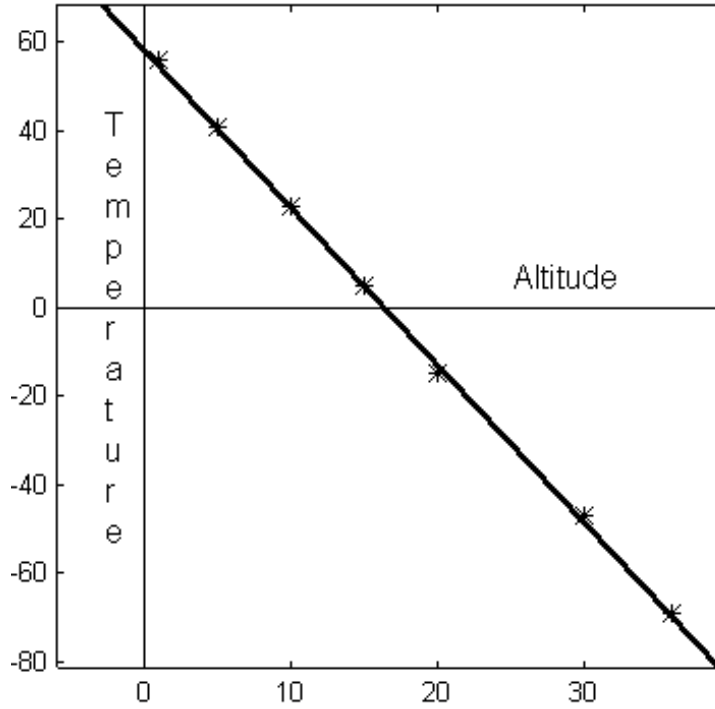


Figure 1.



## Matrix Computations to Obtain the Normal Equations

The normal system of equations was derived using calculus. Here we show how to develop the normal equations directly from the system of equations derived from the expressions  $y_i = mx_i + b$  which have the following matrix equation.

$$\begin{bmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \vdots \\ \vdots & \vdots \\ x_n & 1 \end{bmatrix} \begin{bmatrix} m \\ b \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ \vdots \\ y_n \end{bmatrix} \Leftrightarrow \mathbf{Az} = \mathbf{y} \quad (3)$$

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  $\mathbf{C}$  and right side  $\mathbf{d}$  of the normal system of equations. Each of the entries of  $\mathbf{C}$  and  $\mathbf{d}$  can be expressed as a dot product:

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

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

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

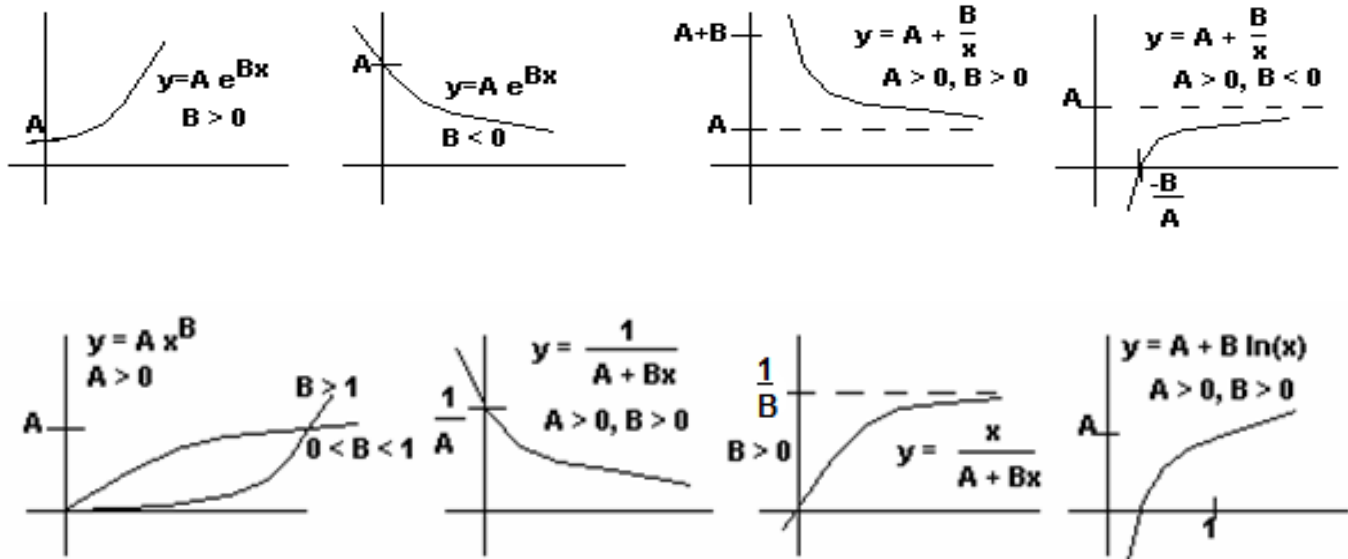
Comparing these expressions to the entries of  $\mathbf{C}$  and  $\mathbf{d}$  and recalling that the entries of the product of a pair of matrices can be expressed as dot products we see that

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

Hence the normal system of equations is obtained from the original system  $\mathbf{Az} = \mathbf{y}$  by multiplying both sides by  $\mathbf{A}^T$  to give  $\mathbf{A}^T \mathbf{A} \mathbf{z} = \mathbf{A}^T \mathbf{y}$ . 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.

## Transformations to Linear/Linearization

In many circumstances, a least squares line model is not appropriate. Growth and decay phenomena often obey exponential laws and still other phenomena are modeled by logarithmic laws, and in some cases higher degree polynomials provide good models. A scatter plot of the data can be useful in determining the type of model to construct. Below are some graphs of families of curves that can be considered. (Note that these curves are monotone.)



Let  $S = \{(x_i, y_i) : i=1,2,\dots,n\}$ . To attempt to model the data in set  $S$  by one of the function forms  $y = F(A,B,x)$  shown above we proceed as in the linear case. Namely minimize the square root of the sum of the squares of the deviations which is given by

$$\sqrt{\sum_{i=1}^n (y_i - F(A,B,x))^2}$$

(In practice to simplify the algebra we just minimize the sum of the squares of the deviations.) When we take the partial derivative of the expression with respect to  $A$  and then with respect to  $B$  and set each of these equal to zero we are lead to two equations to solve for the  $A$  and  $B$ . **Unfortunately with the function forms given above the system of equations is nonlinear. In order to avoid having to solve a nonlinear system we often perform a transformation on the data set so that the transformed data set can be fit with a straight line.** We say we perform a **transformation** on the equation that **linearizes** it; that is, converts it into a linear combination of functions to which we can apply an approach which requires that we solve a linear system of equations.

**Example:** Suppose you choose to use the power law  $y = ax^b$ . Taking the logarithm of both sides we get

$$\log(y) = \log(ax^b) = \log(a) + b \log(x)$$

(We could have used the natural logarithm.) For data set  $S = \{(x_i, y_i) : i=1,2,\dots,n\}$  we get the system of equations

$$\log(y_i) = \log(a) + b \log(x_i), i = 1, 2, \dots, n.$$

**What restriction does this place on the data set?**

Next let  $\mathbf{A} = \log(a)$  and  $\mathbf{B} = b$  and we can write the matrix form of the over determined linear system as

$$\begin{bmatrix} 1 & \log(x_1) \\ 1 & \log(x_2) \\ \vdots & \vdots \\ \vdots & \vdots \\ 1 & \log(x_n) \end{bmatrix} \begin{bmatrix} A \\ B \end{bmatrix} = \begin{bmatrix} \log(y_1) \\ \log(y_2) \\ \vdots \\ \log(y_n) \end{bmatrix}$$

Let  $\mathbf{C} = \begin{bmatrix} 1 & \log(x_1) \\ 1 & \log(x_2) \\ \vdots & \vdots \\ \vdots & \vdots \\ 1 & \log(x_n) \end{bmatrix}$ ,  $\mathbf{d} = \begin{bmatrix} \log(y_1) \\ \log(y_2) \\ \vdots \\ \log(y_n) \end{bmatrix}$ , and multiply both sides of the preceding equation by  $\mathbf{C}^T$  to obtain

the 2 by 2 linear system  $\mathbf{C}^T \mathbf{C} \begin{bmatrix} A \\ B \end{bmatrix} = \mathbf{C}^T \mathbf{d}$  to solve for A and B. Finally we solve for parameters  $\mathbf{a}$  and  $\mathbf{b}$ ;

$\mathbf{a} = 10^A$  and  $\mathbf{b} = B$  and we have the model equation  $\mathbf{y} = \mathbf{a}\mathbf{x}^{\mathbf{b}}$ . Note: the linearized approach does not necessarily provide the same values for the parameters as solving the nonlinear model.

**Example:** The following table lists weight W and the wingspan L for birds of a particular species.

x = W (kilograms)	0.5	1.5	2.0	2.5	3.0
y = L (metters)	0.77	1.10	1.22	1.31	1.40

Develop a regression model of the form  $\mathbf{y} = \mathbf{a}\mathbf{x}^{\mathbf{b}}$  and approximate the wingspan for a bird weighing 3.2 kilograms.

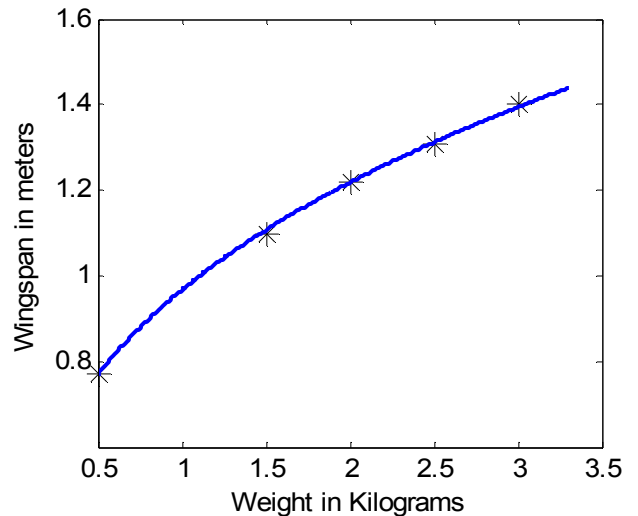
Using the notation above we have

$$\mathbf{C} = \begin{bmatrix} 1 & \ln(0.5) \\ 1 & \ln(1.5) \\ 1 & \ln(2.0) \\ 1 & \ln(2.5) \\ 1 & \ln(3.0) \end{bmatrix}, \mathbf{d} = \begin{bmatrix} \ln(0.77) \\ \ln(1.10) \\ \ln(1.22) \\ \ln(1.31) \\ \ln(1.40) \end{bmatrix} \text{ and } \mathbf{C}^T \mathbf{C} \begin{bmatrix} A \\ B \end{bmatrix} = \mathbf{C}^T \mathbf{d}$$

$$\rightarrow \begin{bmatrix} 5.0000 & 2.4204 \\ 2.4204 & 3.1718 \end{bmatrix} \begin{bmatrix} A \\ B \end{bmatrix} = \begin{bmatrix} 0.6393 \\ 0.9747 \end{bmatrix}$$

Solving we get  $\begin{bmatrix} A \\ B \end{bmatrix} = \begin{bmatrix} -0.0331 \\ 0.3326 \end{bmatrix}$ . So  $\mathbf{a} = e^{-0.03331} \approx$

0.9672 and  $\mathbf{b} = 0.3326$  and we have model  $\mathbf{y} = 0.9672 \mathbf{x}^{0.3326}$ . It follows that if a bird weighs 3.2 kilograms, the model predicts the wingspan to be  $\mathbf{y} = 0.9672 (3.2)^{0.3326} \approx 1.42$  meters.



+++++

**Example:** Suppose you choose to use the exponential law  $\mathbf{y} = \mathbf{a}\mathbf{b}^{\mathbf{x}}$ . Taking the logarithm of both sides we get

$$\log(\mathbf{y}) = \log(\mathbf{a}\mathbf{b}^{\mathbf{x}}) = \log(\mathbf{a}) + \mathbf{x} \log(\mathbf{b})$$

(We could have used the natural logarithm.) For data set  $S = \{(x_i, y_i) : i=1,2,\dots,n\}$  we get the system of equations

$$\log(y_i) = \log(a) + x_i \log(b), \quad i = 1, 2, \dots, n.$$

**What restriction does this place on the data set?**

Next let  $\mathbf{A} = \log(a)$  and  $\mathbf{B} = \log(b)$  and we can write the matrix form of the over determined linear system as

$$\begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \mathbf{A} \\ \mathbf{B} \end{bmatrix} = \begin{bmatrix} \log(y_1) \\ \log(y_2) \\ \vdots \\ \vdots \\ \log(y_n) \end{bmatrix}. \quad \text{Let } \mathbf{C} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}, \mathbf{d} = \begin{bmatrix} \log(y_1) \\ \log(y_2) \\ \vdots \\ \vdots \\ \log(y_n) \end{bmatrix}, \text{ and multiply both sides of the preceding}$$

equation by  $\mathbf{C}^T$  to obtain the 2 by 2 linear system  $\mathbf{C}^T \mathbf{C} \begin{bmatrix} \mathbf{A} \\ \mathbf{B} \end{bmatrix} = \mathbf{C}^T \mathbf{d}$  to solve for A and B. Finally we

solve for parameters  $\mathbf{a}$  and  $\mathbf{b}$ ;  $\mathbf{a} = 10^{\mathbf{A}}$  and  $\mathbf{b} = 10^{\mathbf{B}}$  and we have the model equation  $\mathbf{y} = \mathbf{a}\mathbf{b}^{\mathbf{x}}$ . Note: the linearized approach does not necessarily provide the same values for the parameters as solving the nonlinear model.

+++++

Another function form that is useful is  $\mathbf{y} = \mathbf{a} + \mathbf{b} \log(\mathbf{x})$ .