Chapter 6. Orthogonality

6.5 The Method of Least Squares

Note. In this section we consider linear systems of equations (which, ironically, sometimes involve functions which are *not* linear) which do not have solutions. We introduce a way to find a "best" approximate solution. You have likely encountered this idea before. It is covered in our Introduction to Probability and Statistics (MATH 1530) class as "regression" (see my online notes at `http://faculty.etsu.edu/gardnerr/1530/Chapter5.pdf`). It is also covered as a multivariable optimization problem in Calculus 2 (MATH 2110); see the last example in my notes for "Extreme Values and Saddle Points" where the formula for the regression line is derived using partial derivatives: `http://faculty.etsu.edu/gardnerr/2110/notes-12e/c14s7.pdf`.

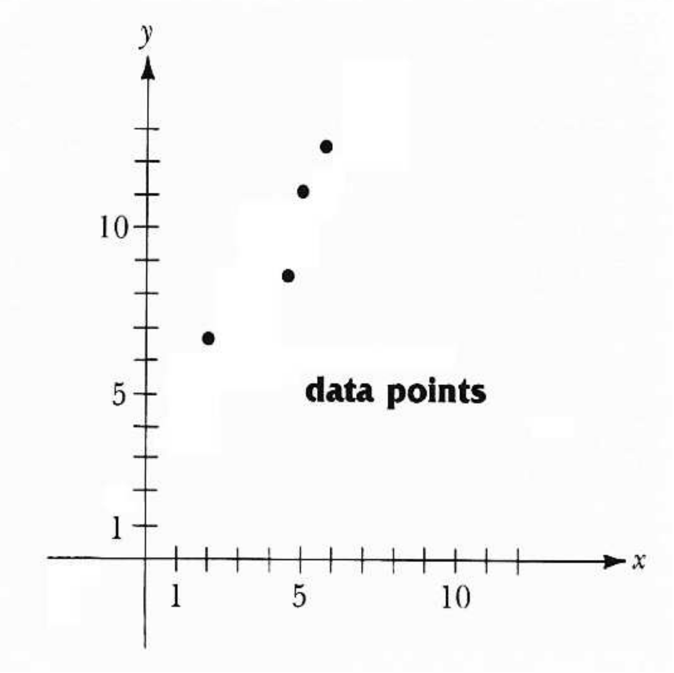
Note. Here, we also deal with linear regression as an optimization problem, but we do so using projections and the following result (the derivation of which appears in Section 6.1, page 328 of the text).

Theorem 6.5.A. Let $\vec{b} \in \mathbb{R}^n$ and let W be a subspace of $\mathbb{R}^n$. Then the projection of $\vec{b}$ onto W, denoted $\vec{b}_W$ (see Theorem 6.1 and Definition 6.2), is the unique vector in W which minimizes the quantity $\|\vec{b} - \vec{w}\|$ where $\vec{w} \in W$.

Note. To further explain the problem, we consider Fraleigh and Beauregard's Problem 1: According to Hooke's Law, the distance that a spring stretches is proportional to the force applied. Suppose that we attach four different weights a_1, a_2, a_3, and a_4 in turn to the bottom of a spring suspended vertically. We measure the four lengths b_1, b_2, b_3, and b_4 of the stretched spring producing the following data:

i	1	2	3	4
a_i (weight in ounces)	2.0	4.0	5.0	6.0
b_i (length in inches)	6.5	8.5	11.0	12.5

If we plot the points (a_i, b_i) in the xy-plane we get



We want to find the "best fit line" of the form $f(x) = r_0 + r_1 x$. From the graph, we see that there is not a line passing through each of the points (as Fraleigh and Beauregard state on page 370: "we expect to have some error in physical measurements"). So if we use $f(x)$ to create four linear equations in the unknowns r_0 and r_1 we get:

i	1	2	3	4
$r_0 + r_1 a_i = b_i$	$r_0 + r_1(2.0) = 6.55$	$r_0 + r_1(4.0) = 8.5$	$r_0 + r_1(5.0) = 11.0$	$r_0 + r_1(6.0) = 12.5$

This yields four equations in two unknowns and we find that the system is "overdetermined" (it has more equations than unknowns) and inconsistent. So we wish to make choices for r_0 and r_1 that are optimal in some sense.

Note. With data points $(a_,b_i)$ for $i = 1, 2, \ldots, m$ and the desired function $f(x) = r_0 + r_1 x$, we are looking for r_0 and r_1 such that

$$\begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix} \approx \begin{bmatrix} 1 & a_1 \\ 1 & a_2 \\ \vdots & \vdots \\ 1 & a_m \end{bmatrix} \begin{bmatrix} r_0 \\ r_1 \end{bmatrix},$$

or $\vec{b} \approx A\vec{r}$. We approach this approximation by minimizing the length of the vector $A\vec{r} - \vec{b}$. Defining $d_i = |(r_0 + r_1 x) - b_i|$ we see that d_i represents the vertical distances from the data points to the graph of the line $y = f(x)$ (so we are not minimizing the *distance* from data points to the line but the distances given in Figure 6.13). In fact, to minimize $\|A\vec{r} - \vec{b}\|$ we equivalently minimize $\|A\vec{r} - \vec{b}\|^2$ which is in terms of the d_i's:

$$\|A\vec{r} - \vec{b}\|^2 = d_1^2 + d_2^2 + \cdots + d_m^2.$$

So we are in fact minimizing the sum of the squares of the distances d_i. That is why the technique is called the *method of least squares.*

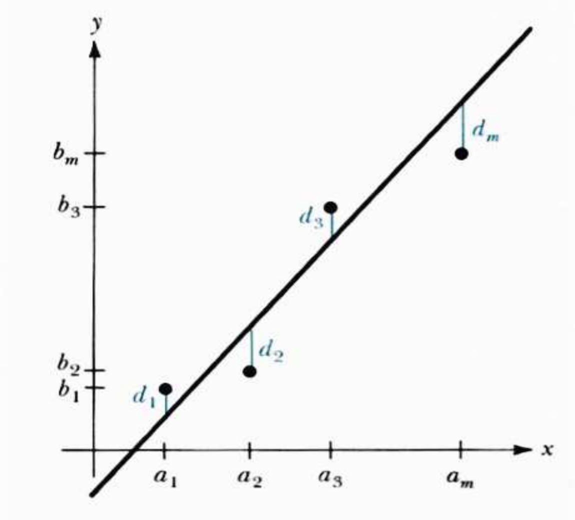


FIGURE 6.13 The distances d_i.

Note. If $\vec{a}_1$ and $\vec{a}_2$ are the columns of A in the system above then vector $A\vec{r} = r_0\vec{a}_1 + r_1\vec{a}_2$ is in the space $W = \text{sp}(\vec{a}_1, \vec{a}_2)$ (the column space of A). So for some given vector $\overline{r} \in \mathbb{R}$ we have geometrically the relationship between $\vec{a}_1$, $\vec{a}_2$, $\vec{b}$, $A\overline{r}$, and $A\overline{r} - \vec{b}$ as given in Figure 6.14 (think of $\overline{r}$ as a variable which minimizes $\|A\overline{r} - \vec{b}\|$. We know by Lemma 6.5.A that $\|A\overline{r} - \vec{b}\| = |\vec{b} - A\overline{r}\|$ is minimized when $A\overline{r} = \vec{b}_W$, where $\vec{b}_W$ is the projection of $\vec{b}$ onto subspace $W = \text{sp}(\vec{a}_1, \vec{a}_2)$.

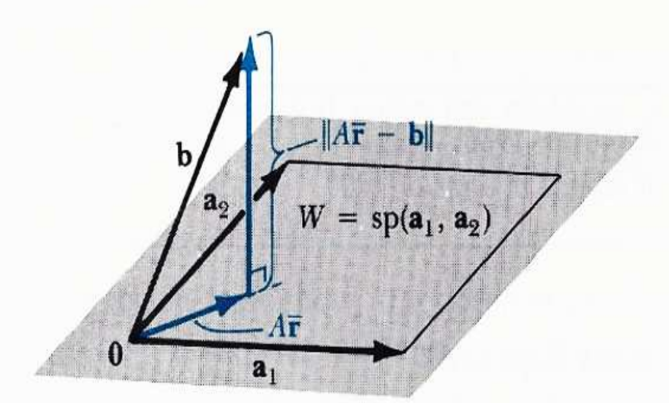

b
$\mathbf{a}_2$
$\|A\bar{\mathbf{r}} - \mathbf{b}\|$
$W = \text{sp}(\mathbf{a}_1, \mathbf{a}_2)$
$A\bar{\mathbf{r}}$
0
$\mathbf{a}_1$

FIGURE 6.14 The length $\|A\bar{\mathbf{r}} - \mathbf{b}\|$.

Note. With our notation, the projection of $\vec{b}$ onto $W = \text{sp}(\vec{a}_1, \vec{a}_2)$ is $\vec{b}_W = A(A^TA)^{-1}\vec{b}$ by "Projection $\vec{b}_W$ of $\vec{b}$ on the Subspace W" on page 362 of the text and on page 4 of the class notes for Section 6.4. So the desired vector $\overline{r}$ satisfies $A\overline{r} = A(A^TA)^{-1}A^T\vec{b}$ and so $\overline{r} = (A^TA)^{-1}A\vec{b}$. We can alternatively solve for $\overline{r}$ by solving the system of equations $(A^TA)\overline{r} = A\vec{b}$.

Example 1. With the data from the Hooke's Law example above, we have

$$\vec{a}_1 = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}, \vec{a}_2 = \begin{bmatrix} 2.0 \\ 4.0 \\ 5.0 \\ 6.0 \end{bmatrix}, \text{ and } \vec{b} = \begin{bmatrix} 6.5 \\ 8.5 \\ 11.0 \\ 12.5 \end{bmatrix}.$$

We find

$$A = \begin{bmatrix} 1 & 2.0 \\ 1 & 4.0 \\ 1 & 5.0 \\ 1 & 6.0 \end{bmatrix}, A^T = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 2.0 & 4.0 & 5.0 & 6.0 \end{bmatrix},$$

$$A^TA = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 2.0 & 4.0 & 5.0 & 6.0 \end{bmatrix} \begin{bmatrix} 1 & 2.0 \\ 1 & 4.0 \\ 1 & 5.0 \\ 1 & 6.0 \end{bmatrix} = \begin{bmatrix} 4 & 17 \\ 17 & 81 \end{bmatrix},$$

and $(A^TA)^{-1} = \dfrac{1}{35}\begin{bmatrix} 81 & -17 \\ -17 & 4 \end{bmatrix}$ (we express A^TA and $(A^TA)^{-1}$ without decimals). So

$$\overline{r} = (A^T A)^{-1}\vec{b} = \frac{1}{35}\begin{bmatrix} 81 & -17 \\ -17 & 4 \end{bmatrix}\begin{bmatrix} 1 & 1 & 1 & 1 \\ 2.0 & 4.0 & 5.0 & 6.0 \end{bmatrix}\begin{bmatrix} 6.5 \\ 8.5 \\ 11.0 \\ 12.5 \end{bmatrix}\frac{1}{35}\begin{bmatrix} 109.5 \\ 53.5 \end{bmatrix}.$$

So we take $r_0 = 109.5/35$ and $r_1 = 53.5/35$. But the measurements of the data are given to one decimal place, so we approximate as $r_0 \approx 3.1$ and $r_1 \approx 1.5$. So we take as the regression line $f(x) = 1.5x + 3.1$. The graph of the line and data together is given in Figure 6.15.

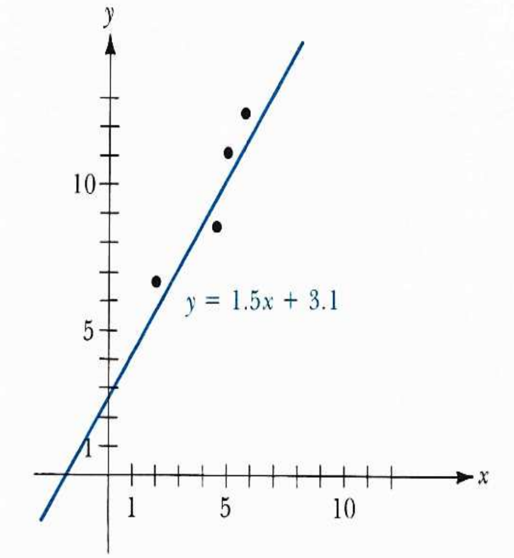


FIGURE 6.15 The least-squares fit of data points.

Example 2. In Problem 3 of this section, the size of a population of rabbits is given for four consecutive years as

a_i (year of observation)	1	2	3	4
b_i (# rabbits in 1,000s)	3	4.5	8	17

As you see in Calculus 2, populations are expected to follow an exponential growth function (see my online notes for "Exponential Change and Separable Differential Equations" at `http://faculty.etsu.edu/gardnerr/1920/12/c7s2.pdf`). So we look for a function of the form $y = f(x) = re^{sx}$ for some r and s. This function is equivalent to

$$\ln y = \ln(re^{sx}) = \ln r + \ln(e^{sx}) = \ln r + sx.$$

So we can treat the equation $\ln y = \ln r + sx$ as a linear relation between x and $\ln y$. So we transform the data from points (a_i, b_i) to points $(a_i, \ln b_i)$:

a_i	1	2	3	4
b_i	3	4.5	8	17
b_i	1.10	1.50	2.08	2.83

where we approximate the logarithm values to two decimal places. Now we take

$$A = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{bmatrix} \text{ (a before) and } \vec{b} = \begin{bmatrix} 1.10 \\ 1.50 \\ 2.08 \\ 2.83 \end{bmatrix}.$$

We find (see page 376 for some of the computations) that

$$\overline{r} = (A^T A)^{-1} A^T \vec{b} \approx \begin{bmatrix} 0.435 \\ 0.577 \end{bmatrix}$$

(where we approximate results to three decimal places). Now this gives $\ln r \approx 0.435$ and $s \approx 0.577$. So $\approx e^{0.435} \approx 1.54$. Therefore we have used linear regression to get the regression exponential function $f(x) = 1.54e^{0.577x}$. See Figure 6.16 for a graph of the data and $f(x)$.

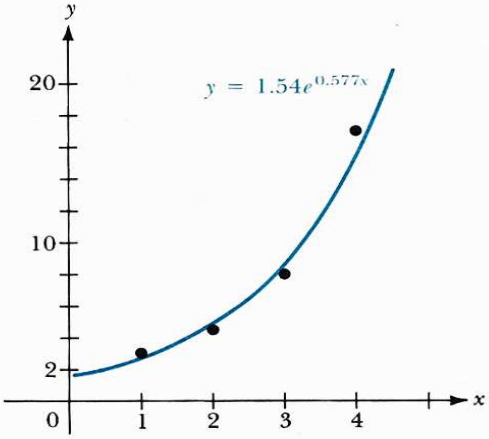


FIGURE 6.16 **Data points and the exponential fit.**

Note. There is nothing special about the exponential function in the previous example. As long as the desired function can be translated into a linear combination, as in the previous example; we could use a function $f(x) = r + sx^x$ for example.

Note. Suppose we perform an experiment where we measure an output value b_i in terms of several input values $a_{i1}, a_{i2}, \ldots, a_{in}$. For example, we might survey people on how comfortable they currently consider the weather and measure the temperature, humidity, wind speed, and barometric pressure at the time of their response. We could then perform a least squares linear regression, as before, in search of a function of the form

$$y = f(x) = r_0 + r_1 x_1 + r_2 x_2 + \cdots + r_n x_n$$

where we have the data points $(a_{i1}, 1_{i2}, \ldots, a_{in}, b_i)$ for $1 \leq i \leq m$. As above, define

$$\vec{b} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix}, A = \begin{bmatrix} 1 & a_{11} & a_{12} & \cdots & a_{1n} \\ 1 & a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}, \text{ and } \vec{r} = \begin{bmatrix} r_0 \\ r_1 \\ \vdots \\ r_n \end{bmatrix}.$$

If $m > n + 1$ (so there are more data points than unknowns) then we have an overdetermined linear systems as before. So by the same geometric argument as above (just in higher dimensions so that Figure 6.14 no longer holds, strictly speaking). So again we choose $\overline{r} = (A^T A)^{-1} A^T \vec{b}$ (or $(A^T A)\overline{r} = A^T \vec{b}$). This is often called "multiple linear regression." Of course we don't want to do these computations by hand!

Note. As in the population growth example, we can use multiple linear regression to find a least squares nonlinear function.

Example 3. In Problem 2 on page 371 we are given the data (a_i, b_i) (representing the weight in tons of a boat, a_i, and the price of the boat in units of \$10,000, b_i):

a_i	2	4	5	8
b_i	1	3	5	12

Eyeing the data, it seems to fall along a parabola. So we find a quadratic function of the form $y = f(x) = r_0 + r_1 x + r_2 x^2$. We now consider the transformed data:

a_i	2	4	5	8
a_i^2	4	16	25	64
b_i	1	3	5	12

so that

$$A = \begin{bmatrix} 1 & 2 & 4 \\ 1 & 4 & 16 \\ 1 & 5 & 25 \\ 1 & 8 & 64 \end{bmatrix} \text{ and } \vec{b} = \begin{bmatrix} 1 \\ 3 \\ 5 \\ 12 \end{bmatrix}.$$

After computations we find (to three decimal places) $\overline{r} = (A^T A)^{-1} A^T \vec{b} \approx \begin{bmatrix} 0.207 \\ 0.010 \\ 0.183 \end{bmatrix}$,

so that the regression function is $f(x) = 0.207 + 0.10x + 0.183x^2$. See Figure 6.17.

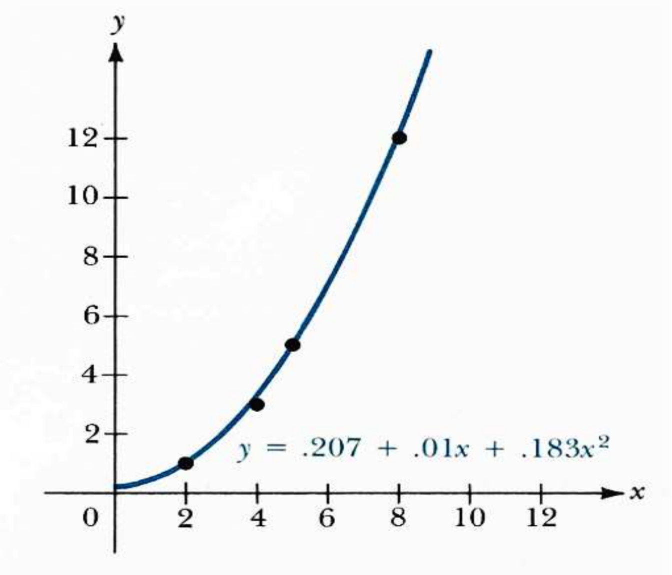


FIGURE 6.17 **The graph and data points for Example 3.**

Note. We have used the method of least squares to estimate a solution $\overline{r}$ to the system of equations $A\vec{r} = \vec{b}$. In fact, we could do this for any overdetermined system $A\vec{x} = \vec{b}$, in which for any overdetermined system $A\vec{x} = \vec{b}$, in which case we take $\overline{x} = (A^T A)^{-1} A^T \vec{b}$ as the *least squares solution* of the overdetermined solution.

Note. If you have seem simple linear regression before, then you may recall that there are easy formulas for r_0 and r_1 in terms of the data. Exercise 6.5.14 gives a special case of this:

> Let $(a_1, b_1), (a_2, b_2), \ldots, (a_m, b_m)$ be data points. If $\sum_{i=1}^m a_i = 0$, show that the line that best fits the data in the least squares sense is given by $r_0 + r_1 x$ where $r_0 = \left(\sum_{i=1}^m b_i\right)/m$ and $r_1 = (\sum_{i=1}^m a_i b_i)/\left(\sum_{i=1}^m a_i^2\right)$.

This exercise depends on $\overline{a} = (\sum_{i=1}^m a_i)/\, m = 0$. Any data set can be translated to satisfy this condition (by replacing a_i with $a_i - \overline{a}$). This observation allows us to derive the general formula for r_0 and r_1 without the $\overline{a} = 0$ restriction:

$$r_1 = \frac{\sum_{i=1}^m (a_i - \overline{a})(b_i - \overline{b})}{\sum_{i=1}^m (a_i - \overline{a})^2} \text{ and } r_0 = \overline{b} - r_1 \overline{a}$$

where $\overline{a} = (\sum_{i=1}^m a_i)/\, m$ and $\overline{b} = (\sum_{i=1}^m b_i)/\, m$. An important topic we have not touched on here is *correlation* and the *correlation coefficient*.

Revised: 5/11/2018