4.6 CLASSICAL LEAST SQUARES

The following problem arises in almost all areas where mathematics is applied. At discrete points t_i (often points in time), observations b_i of some phenomenon are made, and the results are recorded as a set of ordered pairs

$$\mathcal{D} = \{(t_1, b_1),\quad (t_2, b_2),\ \ldots,\ (t_m, b_m)\}.$$

On the basis of these observations, the problem is to make estimations or predictions at points (times) $\hat{t}$ that are between or beyond the observation points t_i. A standard approach is to find the equation of a curve $y = f(t)$ that closely fits the points in $\mathcal{D}$ so that the phenomenon can be estimated at any nonobservation point $\hat{t}$ with the value $\hat{y} = f(\hat{t})$.

Let's begin by fitting a straight line to the points in $\mathcal{D}$. Once this is understood, it will be relatively easy to see how to fit the data with curved lines.

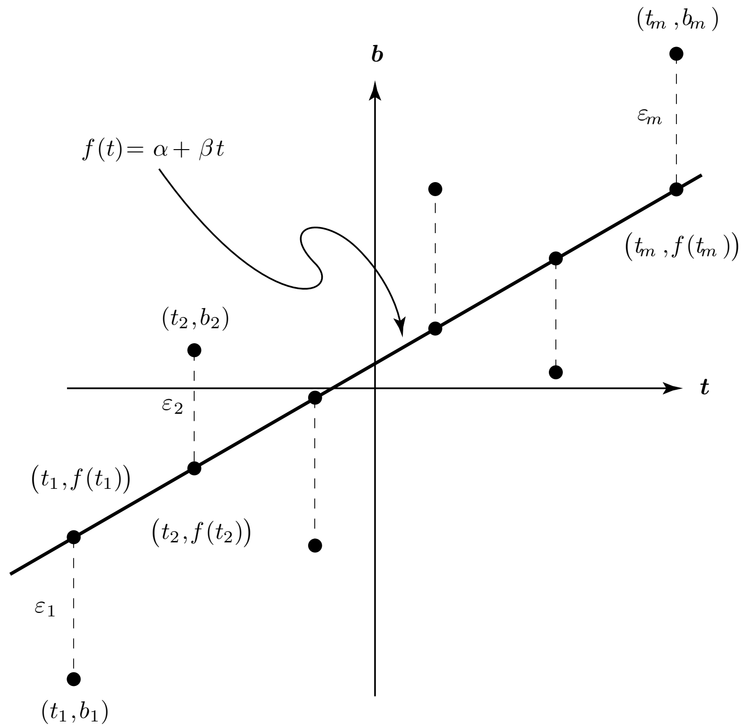

FIGURE 4.6.1

The strategy is to determine the coefficients α and β in the equation of the line $f(t) = \alpha + \beta t$ that best fits the points (t_i, b_i) in the sense that the sum of the squares of the vertical[31] errors $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_m$ indicated in Figure 4.6.1 is

[31] We consider only vertical errors because there is a tacit assumption that only the observations b_i are subject to error or variation. The t_i 's are assumed to be errorless constants—think of them as being exact points in time (as they often are). If the t_i 's are also subject to variation, then horizontal as well as vertical errors have to be considered in Figure 4.6.1, and a more complicated theory known as *total least squares* (not considered in this text) emerges. The least squares line $\mathcal{L}$ obtained by minimizing only vertical deviations will not be the closest line to points in $\mathcal{D}$ in terms of perpendicular distance, but $\mathcal{L}$ is the best line for the purpose of linear estimation—see §5.14 (p. 446).

minimal. The distance from (t_i, b_i) to a line $f(t) = \alpha + \beta t$ is

$$\varepsilon_i = |f(t_i) - b_i| = |\alpha + \beta t_i - b_i|,$$

so that the objective is to find values for α and β such that

$$\sum_{i=1}^{m} \varepsilon_i^2 = \sum_{i=1}^{m} (\alpha + \beta t_i - b_i)^2 \quad \text{is minimal.}$$

Minimization techniques from calculus tell us that the minimum value must occur at a solution to the two equations

$$0 = \frac{\partial \left(\sum_{i=1}^{m} (\alpha + \beta t_i - b_i)^2 \right)}{\partial \alpha} = 2 \sum_{i=1}^{m} (\alpha + \beta t_i - b_i),$$

$$0 = \frac{\partial \left(\sum_{i=1}^{m} (\alpha + \beta t_i - b_i)^2 \right)}{\partial \beta} = 2 \sum_{i=1}^{m} (\alpha + \beta t_i - b_i)\, t_i.$$

Rearranging terms produces two equations in the two unknowns α and β

$$\begin{aligned} \left(\sum_{i=1}^{m} 1 \right) \alpha + \left(\sum_{i=1}^{m} t_i \right) \beta &= \sum_{i=1}^{m} b_i, \\ \left(\sum_{i=1}^{m} t_i \right) \alpha + \left(\sum_{i=1}^{m} t_i^2 \right) \beta &= \sum_{i=1}^{m} t_i b_i. \end{aligned} \tag{4.6.1}$$

By setting

$$\mathbf{A} = \begin{pmatrix} 1 & t_1 \\ 1 & t_2 \\ \vdots & \vdots \\ 1 & t_m \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{pmatrix}, \quad \text{and} \quad \mathbf{x} = \begin{pmatrix} \alpha \\ \beta \end{pmatrix},$$

we see that the two equations (4.6.1) have the matrix form $\mathbf{A}^T\mathbf{A}\mathbf{x} = \mathbf{A}^T\mathbf{b}$. In other words, (4.6.1) is the system of normal equations associated with the system $\mathbf{A}\mathbf{x} = \mathbf{b}$ (see p. 213). The t_i 's are assumed to be distinct numbers, so $rank(\mathbf{A}) = 2$, and (4.5.7) insures that the normal equations have a unique solution given by

$$\begin{aligned} \mathbf{x} &= \left(\mathbf{A}^T\mathbf{A}\right)^{-1}\mathbf{A}^T\mathbf{b} \\ &= \frac{1}{m\sum t_i^2 - \left(\sum t_i\right)^2} \begin{pmatrix} \sum t_i^2 & -\sum t_i \\ -\sum t_i & m \end{pmatrix} \begin{pmatrix} \sum b_i \\ \sum t_i b_i \end{pmatrix} \\ &= \frac{1}{m\sum t_i^2 - \left(\sum t_i\right)^2} \begin{pmatrix} \sum t_i^2 \sum b_i - \sum t_i \sum t_i b_i \\ m \sum t_i b_i - \sum t_i \sum b_i \end{pmatrix} = \begin{pmatrix} \alpha \\ \beta \end{pmatrix}. \end{aligned}$$

Finally, notice that the total sum of squares of the errors is given by

$$\sum_{i=1}^{m} \varepsilon_i^2 = \sum_{i=1}^{m} (\alpha + \beta t_i - b_i)^2 = (\mathbf{A}\mathbf{x} - \mathbf{b})^T (\mathbf{A}\mathbf{x} - \mathbf{b}).$$

Example 4.6.1

Problem: A small company has been in business for four years and has recorded annual sales (in tens of thousands of dollars) as follows.

Year	1	2	3	4
Sales	23	27	30	34

When this data is plotted as shown in Figure 4.6.2, we see that although the points do not exactly lie on a straight line, there nevertheless appears to be a linear trend. Predict the sales for any future year if this trend continues.

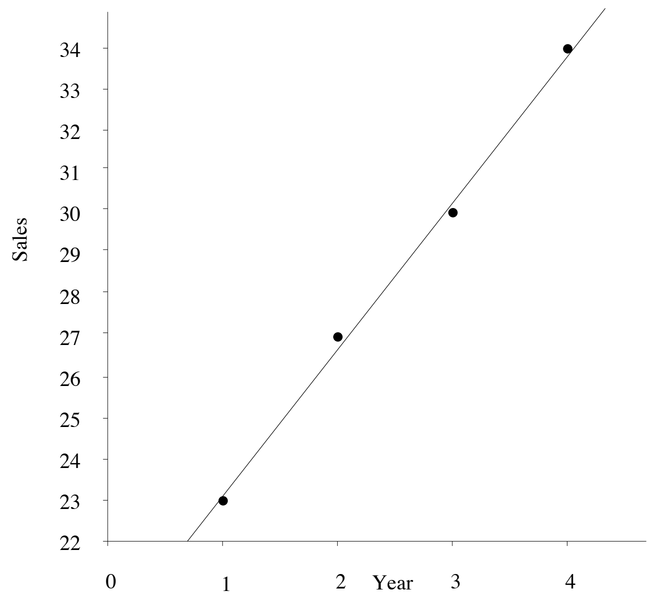

FIGURE 4.6.2

Solution: Determine the line $f(t) = \alpha + \beta t$ that best fits the data in the sense of least squares. If

$$\mathbf{A} = \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} 23 \\ 27 \\ 30 \\ 34 \end{pmatrix}, \quad \text{and} \quad \mathbf{x} = \begin{pmatrix} \alpha \\ \beta \end{pmatrix},$$

then the previous discussion guarantees that $\mathbf{x}$ is the solution of the normal equations $\mathbf{A}^T\mathbf{A}\mathbf{x} = \mathbf{A}^T\mathbf{b}$. That is,

$$\begin{pmatrix} 4 & 10 \\ 10 & 30 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} 114 \\ 303 \end{pmatrix}.$$

The solution is easily found to be $\alpha = 19.5$ and $\beta = 3.6$, so we predict that the sales in year t will be $f(t) = 19.5 + 3.6t$. For example, the estimated sales for year five is \$375,000. To get a feel for how close the least squares line comes to passing through the data points, let $\boldsymbol{\varepsilon} = \mathbf{Ax} - \mathbf{b}$, and compute the sum of the squares of the errors to be

$$\sum_{i=1}^{m} \varepsilon_i^2 = \boldsymbol{\varepsilon}^T \boldsymbol{\varepsilon} = (\mathbf{Ax} - \mathbf{b})^T (\mathbf{Ax} - \mathbf{b}) = .2.$$

General Least Squares Problem

For $\mathbf{A} \in \Re^{m \times n}$ and $\mathbf{b} \in \Re^m$, let $\boldsymbol{\varepsilon} = \boldsymbol{\varepsilon}(\mathbf{x}) = \mathbf{Ax} - \mathbf{b}$. The general least squares problem is to find a vector $\mathbf{x}$ that minimizes the quantity

$$\sum_{i=1}^{m} \varepsilon_i^2 = \boldsymbol{\varepsilon}^T \boldsymbol{\varepsilon} = (\mathbf{Ax} - \mathbf{b})^T (\mathbf{Ax} - \mathbf{b}).$$

Any vector that provides a minimum value for this expression is called a ***least squares solution.***

- The set of all least squares solutions is precisely the set of solutions to the system of normal equations $\mathbf{A}^T\mathbf{Ax} = \mathbf{A}^T\mathbf{b}$.
- There is a unique least squares solution if and only if $rank\,(\mathbf{A}) = n$, in which case it is given by $\mathbf{x} = \left(\mathbf{A}^T\mathbf{A}\right)^{-1}\mathbf{A}^T\mathbf{b}$.
- If $\mathbf{Ax} = \mathbf{b}$ is consistent, then the solution set for $\mathbf{Ax} = \mathbf{b}$ is the same as the set of least squares solutions.

Proof.[32] First prove that if $\mathbf{x}$ minimizes $\boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon}$, then $\mathbf{x}$ must satisfy the normal equations. Begin by using $\mathbf{x}^T\mathbf{A}^T\mathbf{b} = \mathbf{b}^T\mathbf{Ax}$ (scalars are symmetric) to write

$$\sum_{i=1}^{m} \varepsilon_i^2 = \boldsymbol{\varepsilon}^T \boldsymbol{\varepsilon} = (\mathbf{Ax} - \mathbf{b})^T (\mathbf{Ax} - \mathbf{b}) = \mathbf{x}^T\mathbf{A}^T\mathbf{Ax} - 2\mathbf{x}^T\mathbf{A}^T\mathbf{b} + \mathbf{b}^T\mathbf{b}. \tag{4.6.2}$$

To determine vectors $\mathbf{x}$ that minimize the expression (4.6.2), we will again use minimization techniques from calculus and differentiate the function

$$f(x_1, x_2, \ldots, x_n) = \mathbf{x}^T\mathbf{A}^T\mathbf{Ax} - 2\mathbf{x}^T\mathbf{A}^T\mathbf{b} + \mathbf{b}^T\mathbf{b} \tag{4.6.3}$$

with respect to each x_i. Differentiating matrix functions is similar to differentiating scalar functions (see Exercise 3.5.9) in the sense that if $\mathbf{U} = [u_{ij}]$, then

$$\left[\frac{\partial \mathbf{U}}{\partial x}\right]_{ij} = \frac{\partial u_{ij}}{\partial x}, \quad \frac{\partial[\mathbf{U} + \mathbf{V}]}{\partial x} = \frac{\partial \mathbf{U}}{\partial x} + \frac{\partial \mathbf{V}}{\partial x}, \quad \text{and} \quad \frac{\partial[\mathbf{UV}]}{\partial x} = \frac{\partial \mathbf{U}}{\partial x}\mathbf{V} + \mathbf{U}\frac{\partial \mathbf{V}}{\partial x}.$$

32 A more modern development not relying on calculus is given in §5.13 on p. 437, but the more traditional approach is given here because it's worthwhile to view least squares from both perspectives.

Applying these rules to the function in (4.6.3) produces

$$\frac{\partial f}{\partial x_i} = \frac{\partial \mathbf{x}^T}{\partial x_i}\mathbf{A}^T\mathbf{A}\mathbf{x} + \mathbf{x}^T\mathbf{A}^T\mathbf{A}\frac{\partial \mathbf{x}}{\partial x_i} - 2\frac{\partial \mathbf{x}^T}{\partial x_i}\mathbf{A}^T\mathbf{b}.$$

Since $\partial\mathbf{x}/\partial x_i = \mathbf{e}_i$ (the i^{th} unit vector), we have

$$\frac{\partial f}{\partial x_i} = \mathbf{e}_i^T\mathbf{A}^T\mathbf{A}\mathbf{x} + \mathbf{x}^T\mathbf{A}^T\mathbf{A}\mathbf{e}_i - 2\mathbf{e}_i^T\mathbf{A}^T\mathbf{b} = 2\mathbf{e}_i^T\mathbf{A}^T\mathbf{A}\mathbf{x} - 2\mathbf{e}_i^T\mathbf{A}^T\mathbf{b}.$$

Using $\mathbf{e}_i^T\mathbf{A}^T = \left(\mathbf{A}^T\right)_{i*}$ and setting $\partial f/\partial x_i = 0$ produces the n equations

$$\left(\mathbf{A}^T\right)_{i*}\mathbf{A}\mathbf{x} = \left(\mathbf{A}^T\right)_{i*}\mathbf{b} \quad \text{for } i = 1, 2, \ldots, n,$$

which can be written as the single matrix equation $\mathbf{A}^T\mathbf{A}\mathbf{x} = \mathbf{A}^T\mathbf{b}$. Calculus guarantees that the minimum value of f occurs at *some* solution of this system. But this is not enough—we want to know that *every* solution of $\mathbf{A}^T\mathbf{A}\mathbf{x} = \mathbf{A}^T\mathbf{b}$ is a least squares solution. So we must show that the function f in (4.6.3) attains its minimum value at each solution to $\mathbf{A}^T\mathbf{A}\mathbf{x} = \mathbf{A}^T\mathbf{b}$. Observe that if $\mathbf{z}$ is a solution to the normal equations, then $f(\mathbf{z}) = \mathbf{b}^T\mathbf{b} - \mathbf{z}^T\mathbf{A}^T\mathbf{b}$. For any other $\mathbf{y} \in \Re^{n\times 1}$, let $\mathbf{u} = \mathbf{y} - \mathbf{z}$, so $\mathbf{y} = \mathbf{z} + \mathbf{u}$, and observe that

$$f(\mathbf{y}) = f(\mathbf{z}) + \mathbf{v}^T\mathbf{v}, \quad \text{where} \quad \mathbf{v} = \mathbf{A}\mathbf{u}.$$

Since $\mathbf{v}^T\mathbf{v} = \sum_i \mathbf{v}_i^2 \geq 0$, it follows that $f(\mathbf{z}) \leq f(\mathbf{y})$ for all $\mathbf{y} \in \Re^{n\times 1}$, and thus f attains its minimum value at each solution of the normal equations. The remaining statements in the theorem follow from the properties established on p. 213. ∎

The classical least squares problem discussed at the beginning of this section and illustrated in Example 4.6.1 is part of a broader topic known as *linear regression,* which is the study of situations where attempts are made to express one variable y as a linear combination of other variables $t_1, t_2, \ldots, t_n$. In practice, hypothesizing that y is linearly related to $t_1, t_2, \ldots, t_n$ means that one assumes the existence of a set of constants $\{\alpha_0, \alpha_1, \ldots, \alpha_n\}$ (called *parameters*) such that

$$y = \alpha_0 + \alpha_1 t_1 + \alpha_2 t_2 + \cdots + \alpha_n t_n + \varepsilon,$$

where ε is a "random function" whose values "average out" to zero in some sense. Practical problems almost always involve more variables than we wish to consider, but it is frequently fair to assume that the effect of variables of lesser significance will indeed "average out" to zero. The random function ε accounts for this assumption. In other words, a linear hypothesis is the supposition that the expected (or mean) value of y at each point where the phenomenon can be observed is given by a linear equation

$$E(y) = \alpha_0 + \alpha_1 t_1 + \alpha_2 t_2 + \cdots + \alpha_n t_n.$$

To help seat these ideas, consider the problem of predicting the amount of weight that a pint of ice cream loses when it is stored at very low temperatures. There are many factors that may contribute to weight loss—e.g., storage temperature, storage time, humidity, atmospheric pressure, butterfat content, the amount of corn syrup, the amounts of various gums (guar gum, carob bean gum, locust bean gum, cellulose gum), and the never-ending list of other additives and preservatives. It is reasonable to believe that storage time and temperature are the primary factors, so to predict weight loss we will make a linear hypothesis of the form

$$y = \alpha_0 + \alpha_1 t_1 + \alpha_2 t_2 + \varepsilon,$$

where $y =$ weight loss (grams), $t_1 =$ storage time (weeks), $t_2 =$ storage temperature (^oF), and ε is a random function to account for all other factors. The assumption is that all other factors "average out" to zero, so the expected (or mean) weight loss at each point (t_1, t_2) is

$$E(y) = \alpha_0 + \alpha_1 t_1 + \alpha_2 t_2. \tag{4.6.4}$$

Suppose that we conduct an experiment in which values for weight loss are measured for various values of storage time and temperature as shown below.

Time (weeks)	1	1	1	2	2	2	3	3	3
Temp (^oF)	-10	-5	0	-10	-5	0	-10	-5	0
Loss (grams)	.15	.18	.20	.17	.19	.22	.20	.23	.25

If

$$\mathbf{A} = \begin{pmatrix} 1 & 1 & -10 \\ 1 & 1 & -5 \\ 1 & 1 & 0 \\ 1 & 2 & -10 \\ 1 & 2 & -5 \\ 1 & 2 & 0 \\ 1 & 3 & -10 \\ 1 & 3 & -5 \\ 1 & 3 & 0 \end{pmatrix}, \quad \mathbf{x} = \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \end{pmatrix}, \quad \text{and} \quad \mathbf{b} = \begin{pmatrix} .15 \\ .18 \\ .20 \\ .17 \\ .19 \\ .22 \\ .20 \\ .23 \\ .25 \end{pmatrix},$$

and if we were lucky enough to exactly observe the mean weight loss each time (i.e., if $\mathbf{b}_i = E(y_i)$), then equation (4.6.4) would insure that $\mathbf{Ax} = \mathbf{b}$ is a consistent system, so we could solve for the unknown parameters $\alpha_0, \alpha_1,$ and α_2. However, it is virtually impossible to observe the *exact* value of the mean weight loss for a given storage time and temperature, and almost certainly the system defined by $\mathbf{Ax} = \mathbf{b}$ will be inconsistent—especially when the number of observations greatly exceeds the number of parameters. Since we can't solve $\mathbf{Ax} = \mathbf{b}$ to find exact values for the α_i 's, the best we can hope for is a set of "good estimates" for these parameters.

The famous Gauss–Markov theorem (developed on p. 448) states that under certain reasonable assumptions concerning the random error function ε, the "best" estimates for the α_i's are obtained by minimizing the sum of squares $(\mathbf{Ax}-\mathbf{b})^T(\mathbf{Ax}-\mathbf{b})$. In other words, the least squares estimates are the "best" way to estimate the α_i's.

Returning to our ice cream example, it can be verified that $\mathbf{b} \notin R(\mathbf{A})$, so, as expected, the system $\mathbf{Ax}=\mathbf{b}$ is not consistent, and we cannot determine exact values for α_0, α_1, and α_2. The best we can do is to determine least squares estimates for the α_i's by solving the associated normal equations $\mathbf{A}^T\mathbf{Ax}=\mathbf{A}^T\mathbf{b}$, which in this example are

$$\begin{pmatrix} 9 & 18 & -45 \\ 18 & 42 & -90 \\ -45 & -90 & 375 \end{pmatrix} \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \end{pmatrix} = \begin{pmatrix} 1.79 \\ 3.73 \\ -8.2 \end{pmatrix}.$$

The solution is

$$\begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \end{pmatrix} = \begin{pmatrix} .174 \\ .025 \\ .005 \end{pmatrix},$$

and the estimating equation for mean weight loss becomes

$$\hat{y} = .174 + .025t_1 + .005t_2.$$

For example, the mean weight loss of a pint of ice cream that is stored for nine weeks at a temperature of -35^oF is estimated to be

$$\hat{y} = .174 + .025(9) + .005(-35) = .224 \text{ grams}.$$

Example 4.6.2

Least Squares Curve Fitting Problem: Find a polynomial

$$p(t) = \alpha_0 + \alpha_1 t + \alpha_2 t^2 + \cdots + \alpha_{n-1}t^{n-1}$$

with a specified degree that comes as close as possible in the sense of least squares to passing through a set of data points

$$\mathcal{D} = \{(t_1, b_1),\ \ (t_2, b_2),\ \ \ldots,\ \ (t_m, b_m)\},$$

where the t_i's are distinct numbers, and $n \le m$.

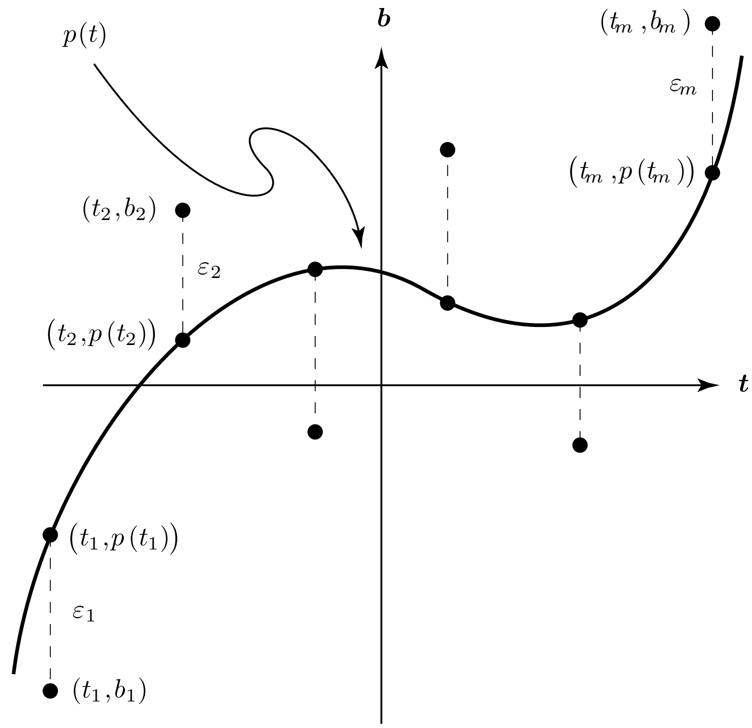

FIGURE 4.6.3

Solution: For the ε_i 's indicated in Figure 4.6.3, the objective is to minimize the sum of squares

$$\sum_{i=1}^{m} \varepsilon_i^2 = \sum_{i=1}^{m} (p(t_i) - b_i)^2 = (\mathbf{Ax} - \mathbf{b})^T(\mathbf{Ax} - \mathbf{b}),$$

where

$$\mathbf{A} = \begin{pmatrix} 1 & t_1 & t_1^2 & \cdots & t_1^{n-1} \\ 1 & t_2 & t_2^2 & \cdots & t_2^{n-1} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & t_m & t_m^2 & \cdots & t_m^{n-1} \end{pmatrix}, \quad \mathbf{x} = \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \vdots \\ \alpha_{n-1} \end{pmatrix}, \quad \text{and} \quad \mathbf{b} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{pmatrix}.$$

In other words, the least squares polynomial of degree $n-1$ is obtained from the least squares solution associated with the system $\mathbf{Ax} = \mathbf{b}$. Furthermore, this least squares polynomial is unique because $\mathbf{A}_{m\times n}$ is the Vandermonde matrix of Example 4.3.4 with $n \le m$, so $rank(\mathbf{A}) = n$, and $\mathbf{Ax} = \mathbf{b}$ has a unique least squares solution given by $\mathbf{x} = \left(\mathbf{A}^T\mathbf{A}\right)^{-1}\mathbf{A}^T\mathbf{b}$.

Note: We know from Example 4.3.5 on p. 186 that the Lagrange interpolation polynomial $\ell(t)$ of degree $m-1$ will *exactly* fit the data—i.e., it passes through each point in $\mathcal{D}$. So why would one want to settle for a least squares fit when an exact fit is possible? One answer stems from the fact that in practical work the observations b_i are rarely exact due to small errors arising from imprecise measurements or from simplifying assumptions. For this reason, it is the *trend* of the observations that needs to be fitted and not the observations themselves. To hit the data points, the interpolation polynomial $\ell(t)$ is usually forced to oscillate between or beyond the data points, and as m becomes larger the oscillations can become more pronounced. Consequently, $\ell(t)$ is generally not useful in making estimations concerning the trend of the observations—Example 4.6.3 drives this point home. In addition to exactly hitting a prescribed set of data points, an interpolation polynomial called the *Hermite polynomial* (p. 607) can be constructed to have specified derivatives at each data point. While this helps, it still is not as good as least squares for making estimations on the basis of observations.

Example 4.6.3

A missile is fired from enemy territory, and its position in flight is observed by radar tracking devices at the following positions.

Position down range (miles)	0	250	500	750	1000
Height (miles)	0	8	15	19	20

Suppose our intelligence sources indicate that enemy missiles are programmed to follow a parabolic flight path—a fact that seems to be consistent with the diagram obtained by plotting the observations on the coordinate system shown in Figure 4.6.4.

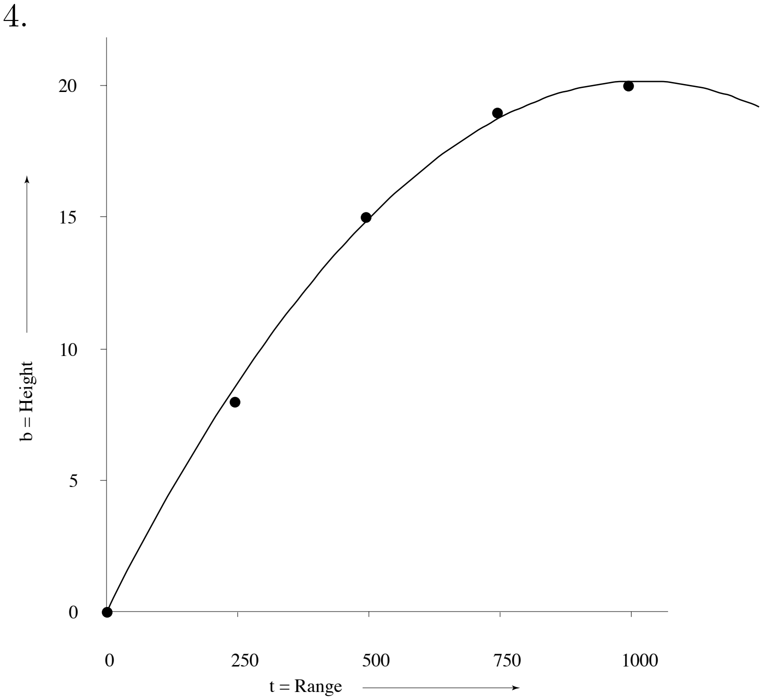

FIGURE 4.6.4

Problem: Predict how far down range the missile will land.

Solution: Determine the parabola $f(t) = \alpha_0 + \alpha_1 t + \alpha_2 t^2$ that best fits the observed data in the least squares sense. Then estimate where the missile will land by determining the roots of f (i.e., determine where the parabola crosses the horizontal axis). As it stands, the problem will involve numbers having relatively large magnitudes in conjunction with relatively small ones. Consequently, it is better to first scale the data by considering one unit to be 1000 miles. If

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & .25 & .0625 \\ 1 & .5 & .25 \\ 1 & .75 & .5625 \\ 1 & 1 & 1 \end{pmatrix}, \quad \mathbf{x} = \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \end{pmatrix}, \quad \text{and} \quad \mathbf{b} = \begin{pmatrix} 0 \\ .008 \\ .015 \\ .019 \\ .02 \end{pmatrix},$$

and if $\boldsymbol{\varepsilon} = \mathbf{Ax} - \mathbf{b}$, then the object is to find a least squares solution $\mathbf{x}$ that minimizes

$$\sum_{i=1}^{5} \varepsilon_i^2 = \boldsymbol{\varepsilon}^T \boldsymbol{\varepsilon} = (\mathbf{Ax} - \mathbf{b})^T (\mathbf{Ax} - \mathbf{b}).$$

We know that such a least squares solution is given by the solution to the system of normal equations $\mathbf{A}^T\mathbf{Ax} = \mathbf{A}^T\mathbf{b}$, which in this case is

$$\begin{pmatrix} 5 & 2.5 & 1.875 \\ 2.5 & 1.875 & 1.5625 \\ 1.875 & 1.5625 & 1.3828125 \end{pmatrix} \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \end{pmatrix} = \begin{pmatrix} .062 \\ .04375 \\ .0349375 \end{pmatrix}.$$

The solution (rounded to four significant digits) is

$$\mathbf{x} = \begin{pmatrix} -2.286 \times 10^{-4} \\ 3.983 \times 10^{-2} \\ -1.943 \times 10^{-2} \end{pmatrix},$$

and the least squares parabola is

$$f(t) = -.0002286 + .03983t - .01943t^2.$$

To estimate where the missile will land, determine where this parabola crosses the horizontal axis by applying the quadratic formula to find the roots of $f(t)$ to be $t = .005755$ and $t = 2.044$. Therefore, we estimate that the missile will land 2044 miles down range. The sum of the squares of the errors associated with the least squares solution is

$$\sum_{i=1}^{5} \varepsilon_i^2 = \boldsymbol{\varepsilon}^T \boldsymbol{\varepsilon} = (\mathbf{Ax} - \mathbf{b})^T (\mathbf{Ax} - \mathbf{b}) = 4.571 \times 10^{-7}.$$

Least Squares vs. Lagrange Interpolation. Instead of using least squares, fit the observations exactly with the fourth-degree Lagrange interpolation polynomial

$$\ell(t) = \frac{11}{375}t + \frac{17}{750000}t^2 - \frac{1}{18750000}t^3 + \frac{1}{46875000000}t^4$$

described in Example 4.3.5 on p. 186 (you can verify that $\ell(t_i) = b_i$ for each observation). As the graph in Figure 4.6.5 indicates, $\ell(t)$ has only one real nonnegative root, so it is worthless for predicting where the missile will land. This is characteristic of Lagrange interpolation.

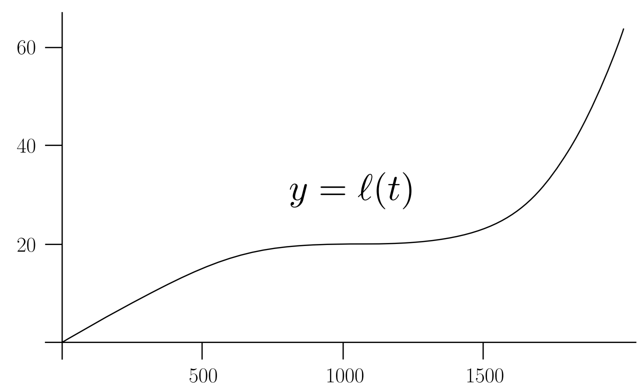

FIGURE 4.6.5

Computational Note: Theoretically, the least squares solutions of $\mathbf{Ax} = \mathbf{b}$ are exactly the solutions of the normal equations $\mathbf{A}^T\mathbf{Ax} = \mathbf{A}^T\mathbf{b}$, but forming and solving the normal equations to compute least squares solutions with floating-point arithmetic is not recommended. As pointed out in Example 4.5.1 on p. 214, any sensitivities to small perturbations that are present in the underlying problem are magnified by forming the normal equations. In other words, if the underlying problem is somewhat ill-conditioned, then the system of normal equations will be ill-conditioned to an even greater extent. Numerically stable techniques that avoid the normal equations are presented in Example 5.5.3 on p. 313 and Example 5.7.3 on p. 346.

Epilogue

While viewing a region in the Taurus constellation on January 1, 1801, Giuseppe Piazzi, an astronomer and director of the Palermo observatory, observed a small "star" that he had never seen before. As Piazzi and others continued to watch this new "star"—which was really an asteroid—they noticed that it was in fact moving, and they concluded that a new "planet" had been discovered. However, their new "planet" completely disappeared in the autumn of 1801. Well-known astronomers of the time joined the search to relocate the lost "planet," but all efforts were in vain.

In September of 1801 Carl F. Gauss decided to take up the challenge of finding this lost "planet." Gauss allowed for the possibility of an elliptical orbit rather than constraining it to be circular—which was an assumption of the others—and he proceeded to develop the method of least squares. By December the task was completed, and Gauss informed the scientific community not only where the lost "planet" was located, but he also predicted its position at future times. They looked, and it was exactly where Gauss had predicted it would be! The asteroid was named *Ceres*, and Gauss's contribution was recognized by naming another minor asteroid *Gaussia*.

This extraordinary feat of locating a tiny and distant heavenly body from apparently insufficient data astounded the scientific community. Furthermore, Gauss refused to reveal his methods, and there were those who even accused him of sorcery. These events led directly to Gauss's fame throughout the entire European community, and they helped to establish his reputation as a mathematical and scientific genius of the highest order.

Gauss waited until 1809, when he published his *Theoria Motus Corporum Coelestium In Sectionibus Conicis Solem Ambientium,* to systematically develop the theory of least squares and his methods of orbit calculation. This was in keeping with Gauss's philosophy to publish nothing but well-polished work of lasting significance. When criticized for not revealing more motivational aspects in his writings, Gauss remarked that architects of great cathedrals do not obscure the beauty of their work by leaving the scaffolds in place after the construction has been completed. Gauss's theory of least squares approximation has indeed proven to be a great mathematical cathedral of lasting beauty and significance.

Exercises for section 4.6

4.6.1. Hooke's law says that the displacement y of an ideal spring is proportional to the force x that is applied—i.e., $y = kx$ for some constant k. Consider a spring in which k is unknown. Various masses are attached, and the resulting displacements shown in Figure 4.6.6 are observed. Using these observations, determine the least squares estimate for k.

x (lb)	y (in)
5	11.1
7	15.4
8	17.5
10	22.0
12	26.3

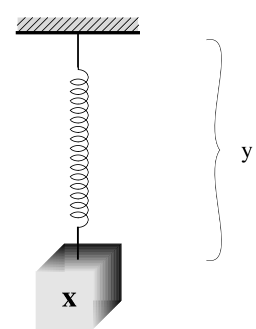

FIGURE 4.6.6

4.6.2. Show that the slope of the line that passes through the origin in $\Re^2$ and comes closest in the least squares sense to passing through the points $\{(x_1, y_1),\ (x_2, y_2), \ldots, (x_n, y_n)\}$ is given by $m = \sum_i x_i y_i / \sum_i x_i^2$.

4.6.3. A small company has been in business for three years and has recorded annual profits (in thousands of dollars) as follows.

Year	1	2	3
Sales	7	4	3

Assuming that there is a linear trend in the declining profits, predict the year and the month in which the company begins to lose money.

4.6.4. An economist hypothesizes that the change (in dollars) in the price of a loaf of bread is primarily a linear combination of the change in the price of a bushel of wheat and the change in the minimum wage. That is, if B is the change in bread prices, W is the change in wheat prices, and M is the change in the minimum wage, then $B = \alpha W + \beta M$. Suppose that for three consecutive years the change in bread prices, wheat prices, and the minimum wage are as shown below.

	Year 1	Year 2	Year 3
B	+\$1	+\$1	+\$1
W	+\$1	+\$2	0\$
M	+\$1	0\$	−\$1

Use the theory of least squares to estimate the change in the price of bread in Year 4 if wheat prices and the minimum wage each fall by \$1.

4.6.5. Suppose that a researcher hypothesizes that the weight loss of a pint of ice cream during storage is primarily a linear function of time. That is,

$$y = \alpha_0 + \alpha_1 t + \varepsilon,$$

where $y =$ the weight loss in grams, $t =$ the storage time in weeks, and ε is a random error function whose mean value is 0. Suppose that an experiment is conducted, and the following data is obtained.

Time (t)	1	2	3	4	5	6	7	8
Loss (y)	.15	.21	.30	.41	.49	.59	.72	.83

(a) Determine the least squares estimates for the parameters α_0 and α_1.

(b) Predict the mean weight loss for a pint of ice cream that is stored for 20 weeks.

4.6.6. After studying a certain type of cancer, a researcher hypothesizes that in the short run the number (y) of malignant cells in a particular tissue grows exponentially with time (t). That is, $y = \alpha_0 e^{\alpha_1 t}$. Determine least squares estimates for the parameters α_0 and α_1 from the researcher's observed data given below.

t (days)	1	2	3	4	5
y (cells)	16	27	45	74	122

Hint: What common transformation converts an exponential function into a linear function?

4.6.7. Using least squares techniques, fit the following data

x	-5	-4	-3	-2	-1	0	1	2	3	4	5
y	2	7	9	12	13	14	14	13	10	8	4

with a line $y = \alpha_0 + \alpha_1 x$ and then fit the data with a quadratic $y = \alpha_0 + \alpha_1 x + \alpha_2 x^2$. Determine which of these two curves best fits the data by computing the sum of the squares of the errors in each case.

4.6.8. Consider the time (T) it takes for a runner to complete a marathon (26 miles and 385 yards). Many factors such as height, weight, age, previous training, etc. can influence an athlete's performance, but experience has shown that the following three factors are particularly important:

$$x_1 = \text{Ponderal index} = \frac{\text{height (in.)}}{[\text{weight (lbs.)}]^{\frac{1}{3}}},$$

$$x_2 = \text{Miles run the previous 8 weeks},$$

$$x_3 = \text{Age (years)}.$$

A linear model hypothesizes that the time T (in minutes) is given by $T = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 + \varepsilon$, where ε is a random function accounting for all other factors and whose mean value is assumed to be zero. On the basis of the five observations given below, estimate the expected marathon time for a 43-year-old runner of height 74 in., weight 180 lbs., who has run 450 miles during the previous eight weeks.

T	x_1	x_2	x_3
181	13.1	619	23
193	13.5	803	42
212	13.8	207	31
221	13.1	409	38
248	12.5	482	45

What is your personal predicted mean marathon time?

4.6.9. For $\mathbf{A} \in \Re^{m\times n}$ and $\mathbf{b} \in \Re^m$, prove that $\mathbf{x}_2$ is a least squares solution for $\mathbf{Ax} = \mathbf{b}$ if and only if $\mathbf{x}_2$ is part of a solution to the larger system

$$\begin{pmatrix} \mathbf{I}_{m\times m} & \mathbf{A} \\ \mathbf{A}^T & \mathbf{0}_{n\times n} \end{pmatrix} \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{b} \\ \mathbf{0} \end{pmatrix}. \tag{4.6.5}$$

Note: It is not uncommon to encounter least squares problems in which $\mathbf{A}$ is extremely large but very sparse (mostly zero entries). For these situations, the system (4.6.5) will usually contain significantly fewer nonzero entries than the system of normal equations, thereby helping to overcome the memory requirements that plague these problems. Using (4.6.5) also eliminates the undesirable need to explicitly form the product $\mathbf{A}^T\mathbf{A}$—recall from Example 4.5.1 that forming $\mathbf{A}^T\mathbf{A}$ can cause loss of significant information.

4.6.10. In many least squares applications, the underlying data matrix $\mathbf{A}_{m\times n}$ does not have independent columns—i.e., $rank(\mathbf{A}) < n$—so the corresponding system of normal equations $\mathbf{A}^T\mathbf{Ax} = \mathbf{A}^T\mathbf{b}$ will fail to have a unique solution. This means that in an associated linear estimation problem of the form

$$y = \alpha_1 t_1 + \alpha_2 t_2 + \cdots + \alpha_n t_n + \varepsilon$$

there will be infinitely many least squares estimates for the parameters α_i, and hence there will be infinitely many estimates for the mean value of y at any given point $(t_1, t_2, \ldots, t_n)$—which is clearly an undesirable situation. In order to remedy this problem, we restrict ourselves to making estimates only at those points $(t_1, t_2, \ldots, t_n)$ that are in the row space of $\mathbf{A}$. If

$$\mathbf{t} = \begin{pmatrix} t_1 \\ t_2 \\ \vdots \\ t_n \end{pmatrix} \in R\left(\mathbf{A}^T\right), \quad \text{and if} \quad \mathbf{x} = \begin{pmatrix} \hat{\alpha}_1 \\ \hat{\alpha}_2 \\ \vdots \\ \hat{\alpha}_n \end{pmatrix}$$

is any least squares solution (i.e., $\mathbf{A}^T\mathbf{Ax} = \mathbf{A}^T\mathbf{b}$), prove that the estimate defined by

$$\hat{y} = \mathbf{t}^T\mathbf{x} = \sum_{i=1}^{n} t_i \hat{\alpha}_i$$

is unique in the sense that $\hat{y}$ is independent of which least squares solution $\mathbf{x}$ is used.

Solutions for exercises in section 4. 6

4.6.1. $\mathbf{A} = \begin{pmatrix} 5 \\ 7 \\ 8 \\ 10 \\ 12 \end{pmatrix}$ and $\mathbf{b} = \begin{pmatrix} 11.1 \\ 15.4 \\ 17.5 \\ 22.0 \\ 26.3 \end{pmatrix}$, so $\mathbf{A}^T\mathbf{A} = 382$ and $\mathbf{A}^T\mathbf{b} = 838.9$. Thus the least squares estimate for k is $838.9/382 = 2.196$.

4.6.2. This is essentially the same problem as Exercise 4.6.1. Because it must pass through the origin, the equation of the least squares line is $y = mx$, and hence $\mathbf{A} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$ and $\mathbf{b} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$, so $\mathbf{A}^T\mathbf{A} = \sum_i x_i^2$ and $\mathbf{A}^T\mathbf{b} = \sum_i x_i y_i$.

4.6.3. Look for the line $p = \alpha + \beta t$ that comes closest to the data in the least squares sense. That is, find the least squares solution for the system $\mathbf{Ax} = \mathbf{b}$, where

$$\mathbf{A} = \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{pmatrix}, \quad \mathbf{x} = \begin{pmatrix} \alpha \\ \beta \end{pmatrix}, \quad \text{and} \quad \mathbf{b} = \begin{pmatrix} 7 \\ 4 \\ 3 \end{pmatrix}.$$

Set up normal equations $\mathbf{A}^T\mathbf{Ax} = \mathbf{A}^T\mathbf{b}$ to get

$$\begin{pmatrix} 3 & 6 \\ 6 & 14 \end{pmatrix}\begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} 14 \\ 24 \end{pmatrix} \implies \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} 26/3 \\ -2 \end{pmatrix} \implies p = (26/3) - 2t.$$

Setting $p = 0$ gives $t = 13/3$. In other words, we expect the company to begin losing money on May 1 of year five.

4.6.4. The associated linear system $\mathbf{Ax} = \mathbf{b}$ is

$$\begin{array}{ll} \text{Year 1:} & \alpha + \beta = 1 \\ \text{Year 2:} & 2\alpha = 1 \\ \text{Year 3:} & -\beta = 1 \end{array} \quad \text{or} \quad \begin{pmatrix} 1 & 1 \\ 2 & 0 \\ 0 & -1 \end{pmatrix}\begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}.$$

The least squares solution to this inconsistent system is obtained from the system of normal equations $\mathbf{A}^T\mathbf{Ax} = \mathbf{A}^T\mathbf{b}$ that is $\begin{pmatrix} 5 & 1 \\ 1 & 2 \end{pmatrix}\begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} 3 \\ 0 \end{pmatrix}$. The unique solution is $\begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} 2/3 \\ -1/3 \end{pmatrix}$, so the least squares estimate for the increase in bread prices is

$$B = \frac{2}{3}W - \frac{1}{3}M.$$

When $W = -1$ and $M = -1$, we estimate that $B = -1/3$.

4.6.5. (a) $\alpha_0 = .02$ and $\alpha_1 = .0983$. (b) 1.986 grams.

4.6.6. Use $\ln y = \ln \alpha_0 + \alpha_1 t$ to obtain the least squares estimates $\alpha_0 = 9.73$ and $\alpha_1 = .507$.

4.6.7. The least squares line is $y = 9.64 + .182x$ and for $\varepsilon_i = 9.64 + .182x_i - y_i$, the sum of the squares of these errors is $\sum_i \varepsilon_i^2 = 162.9$. The least squares quadratic is $y = 13.97 + .1818x - .4336x^2$, and the corresponding sum of squares of the errors is $\sum \varepsilon_i^2 = 1.622$. Therefore, we conclude that the quadratic provides a much better fit.

4.6.8. 230.7 min. $(\alpha_0 = 492.04, \alpha_1 = -23.435, \alpha_2 = -.076134, \alpha_3 = 1.8624)$

4.6.9. $\mathbf{x}_2$ is a least squares solution $\implies$ $\mathbf{A}^T\mathbf{A}\mathbf{x}_2 = \mathbf{A}^T\mathbf{b}$ $\implies$ $\mathbf{0} = \mathbf{A}^T(\mathbf{b} - \mathbf{A}\mathbf{x}_2)$. If we set $\mathbf{x}_1 = \mathbf{b} - \mathbf{A}\mathbf{x}_2$, then

$$\begin{pmatrix} \mathbf{I}_{m\times m} & \mathbf{A} \\ \mathbf{A}^T & \mathbf{0}_{n\times n} \end{pmatrix} \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{I}_{m\times m} & \mathbf{A} \\ \mathbf{A}^T & \mathbf{0}_{n\times n} \end{pmatrix} \begin{pmatrix} \mathbf{b} - \mathbf{A}\mathbf{x}_2 \\ \mathbf{x}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{b} \\ \mathbf{0} \end{pmatrix}.$$

The converse is true because

$$\begin{aligned} \begin{pmatrix} \mathbf{I}_{m\times m} & \mathbf{A} \\ \mathbf{A}^T & \mathbf{0}_{n\times n} \end{pmatrix} \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{b} \\ \mathbf{0} \end{pmatrix} &\implies \mathbf{A}\mathbf{x}_2 = \mathbf{b} - \mathbf{x}_1 \quad \text{and} \quad \mathbf{A}^T\mathbf{x}_1 = \mathbf{0} \\ &\implies \mathbf{A}^T\mathbf{A}\mathbf{x}_2 = \mathbf{A}^T\mathbf{b} - \mathbf{A}^T\mathbf{x}_1 = \mathbf{A}^T\mathbf{b}. \end{aligned}$$

4.6.10. $\mathbf{t} \in R\left(\mathbf{A}^T\right) = R\left(\mathbf{A}^T\mathbf{A}\right) \implies \mathbf{t}^T = \mathbf{z}^T\mathbf{A}^T\mathbf{A}$ for some $\mathbf{z}$. For each $\mathbf{x}$ satisfying $\mathbf{A}^T\mathbf{A}\mathbf{x} = \mathbf{A}^T\mathbf{b}$, write

$$\hat{y} = \mathbf{t}^T\mathbf{x} = \mathbf{z}^T\mathbf{A}^T\mathbf{A}\mathbf{x} = \mathbf{z}^T\mathbf{A}^T\mathbf{b},$$

and notice that $\mathbf{z}^T\mathbf{A}^T\mathbf{b}$ is independent of $\mathbf{x}$.

Solutions for exercises in section 4. 7

4.7.1. (b) and (f)

4.7.2. (a), (c), and (d)

4.7.3. Use any $\mathbf{x}$ to write $\mathbf{T}(\mathbf{0}) = \mathbf{T}(\mathbf{x} - \mathbf{x}) = \mathbf{T}(\mathbf{x}) - \mathbf{T}(\mathbf{x}) = \mathbf{0}$.

4.7.4. (a)

4.7.5. (a) No (b) Yes

4.7.6. $\mathbf{T}(\mathbf{u}_1) = (2,2) = 2\mathbf{u}_1 + 0\mathbf{u}_2$ and $\mathbf{T}(\mathbf{u}_2) = (3,6) = 0\mathbf{u}_1 + 3\mathbf{u}_2$ so that $[\mathbf{T}]_{\mathcal{B}} = \begin{pmatrix} 2 & 0 \\ 0 & 3 \end{pmatrix}$.

4.7.7. (a) $[\mathbf{T}]_{\mathcal{SS}'} = \begin{pmatrix} 1 & 3 \\ 0 & 0 \\ 2 & -4 \end{pmatrix}$ (b) $[\mathbf{T}]_{\mathcal{SS}''} = \begin{pmatrix} 2 & -4 \\ 0 & 0 \\ 1 & 3 \end{pmatrix}$

4.7.8. $[\mathbf{T}]_{\mathcal{B}} = \begin{pmatrix} 1 & -3/2 & 1/2 \\ -1 & 1/2 & 1/2 \\ 0 & 1/2 & -1/2 \end{pmatrix}$ and $[\mathbf{v}]_{\mathcal{B}} = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}$.