22. Project the vector $b = (1, 1)$ onto the lines through $a_1 = (1, 0)$ and $a_2 = (1, 2)$. Draw the projections p_1 and p_2 and add $p_1 + p_2$. The projections do not add to b because the a's are not orthogonal.

23. In Problem 22, the projection of b onto the *plane* of a_1 and a_2 will equal b. Find $P = A(A^{\mathrm{T}}A)^{-1}A^{\mathrm{T}}$ for $A = \begin{bmatrix} a_1 & a_2 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 0 & 2 \end{bmatrix}$.

24. Project $b = (1, 0, 0)$ onto the lines through a_1 and a_2 in Problem 21 and also onto $a_3 = (2, -1, 2)$. Add the three projections $p_1 + p_2 + p_3$.

25. Project $a_1 = (1, 0)$ onto $a_2 = (1, 2)$. Then project the result back onto a_1. Draw these projections and multiply the projection matrices $P_1 P_2$: Is this a projection?

26. Continuing Problems 21, 24 find the projection matrix P_3 onto $a_3 = (2, -1, 2)$. Verify that $P_1 + P_2 + P_3 = I$. The basis a_1, a_2, a_3 is orthogonal!

3.3 PROJECTIONS AND LEAST SQUARES

Up to this point, $Ax = b$ either has a solution or not. If b is not in the column space $C(A)$, the system is inconsistent and Gaussian elimination fails. This failure is almost certain when there are several equations and only one unknown:

$$\textbf{More equations than unknowns—no solution?} \qquad \begin{aligned} 2x &= b_1 \\ 3x &= b_2 \\ 4x &= b_3. \end{aligned}$$

This is solvable when b_1, b_2, b_3 are in the ratio 2:3:4. The solution x will exist only if b is on the same line as the column $a = (2, 3, 4)$.

In spite of their unsolvability, inconsistent equations arise all the time in practice. They have to be solved! One possibility is to determine x from part of the system, and ignore the rest; this is hard to justify if all m equations come from the same source. Rather than expecting no error in some equations and large errors in the others, it is much better *to choose the x that minimizes an average error E in the m equations.*

The most convenient "average" comes from the *sum of squares*:

$$\textbf{Squared error} \qquad E^2 = (2x - b_1)^2 + (3x - b_2)^2 + (4x - b_3)^2.$$

If there is an exact solution, the minimum error is $E = 0$. In the more likely case that b is not proportional to a, the graph of E^2 will be a parabola. The minimum error is at the lowest point, where the derivative is zero:

$$\frac{dE^2}{dx} = 2[(2x - b_1)2 + (3x - b_2)3 + (4x - b_3)4] = 0.$$

Solving for x, the least-squares solution of this model system $ax = b$ is denoted by $\widehat{x}$:

$$\textbf{Least-squares solution} \qquad \widehat{x} = \frac{2b_1 + 3b_2 + 4b_3}{2^2 + 3^2 + 4^2} = \frac{a^{\mathrm{T}}b}{a^{\mathrm{T}}a}.$$

You recognize $a^{\mathrm{T}}b$ in the numerator and $a^{\mathrm{T}}a$ in the denominator.

The general case is the same. We "solve" $ax = b$ by minimizing

$$E^2 = \|ax - b\|^2 = (a_1x - b_1)^2 + \cdots + (a_mx - b_m)^2.$$

The derivative of E^2 is zero at the point $\widehat{x}$, if

$$(a_1\widehat{x} - b_1)a_1 + \cdots + (a_m\widehat{x} - b_m)a_m = 0.$$

We are minimizing the distance from b to the line through a, and calculus gives the same answer, $\widehat{x} = (a_1b_1 + \cdots + a_mb_m)/(a_1^2 + \cdots + a_m^2)$, that geometry did earlier:

3K The least-squares solution to a problem $ax = b$ in one unknown is $\widehat{x} = \dfrac{a^{\mathrm{T}}b}{a^{\mathrm{T}}a}$.

You see that we keep coming back to the geometrical interpretation of a least-squares problem—to minimize a distance. By setting the derivative of E^2 to zero, calculus confirms the geometry of the previous section. *The error vector e connecting b to p must be perpendicular to a*:

$$\textbf{Orthogonality of } \boldsymbol{a} \textbf{ and } \boldsymbol{e} \qquad a^{\mathrm{T}}(b - \widehat{x}a) = a^{\mathrm{T}}b - \frac{a^{\mathrm{T}}b}{a^{\mathrm{T}}a}a^{\mathrm{T}}a = 0.$$

As a side remark, notice the degenerate case $a = 0$. All multiples of a are zero, and the line is only a point. Therefore $p = 0$ is the only candidate for the projection. But the formula for $\widehat{x}$ becomes a meaningless $0/0$, and correctly reflects the fact that $\widehat{x}$ is completely undetermined. All values of x give the same error $E = \|0x - b\|$, so E^2 is a horizontal line instead of a parabola. The "pseudoinverse" assigns the definite value $\widehat{x} = 0$, which is a more "symmetric" choice than any other number.

Least-Squares Problems with Several Variables

Now we are ready for the serious step, ***to project b onto a subspace***—rather than just onto a line. This problem arises from $Ax = b$ when A is an m by n matrix. Instead of one column and one unknown x, the matrix now has n columns. The number m of observations is still larger than the number n of unknowns, so it must be expected that $Ax = b$ will be inconsistent. *Probably, there will not exist a choice of x that perfectly fits the data b*. In other words, the vector b probably will not be a combination of the columns of A; it will be outside the column space.

Again the problem is to choose $\widehat{x}$ so as to minimize the error, and again this minimization will be done in the least-squares sense. The error is $E = \|Ax - b\|$, and ***this is exactly the distance from b to the point Ax in the column space***. Searching for the least-squares solution $\widehat{x}$, which minimizes E, is the same as locating the point $p = A\widehat{x}$ that is closer to b than any other point in the column space.

We may use geometry or calculus to determine $\widehat{x}$. In n dimensions, we prefer the appeal of geometry; p must be the "projection of b onto the column space." ***The error vector*** $\boldsymbol{e = b - A\widehat{x}}$ ***must be perpendicular to that space*** (Figure 3.8). Finding $\widehat{x}$ and the projection $p = A\widehat{x}$ is so fundamental that we do it in two ways:

1. All vectors perpendicular to the column space lie in the *left nullspace*. Thus the error vector $e = b - A\widehat{x}$ must be in the nullspace of A^{T}:

$$A^{\mathrm{T}}(b - A\widehat{x}) = 0 \qquad \text{or} \qquad \boldsymbol{A^{\mathrm{T}}A\widehat{x} = A^{\mathrm{T}}b}.$$

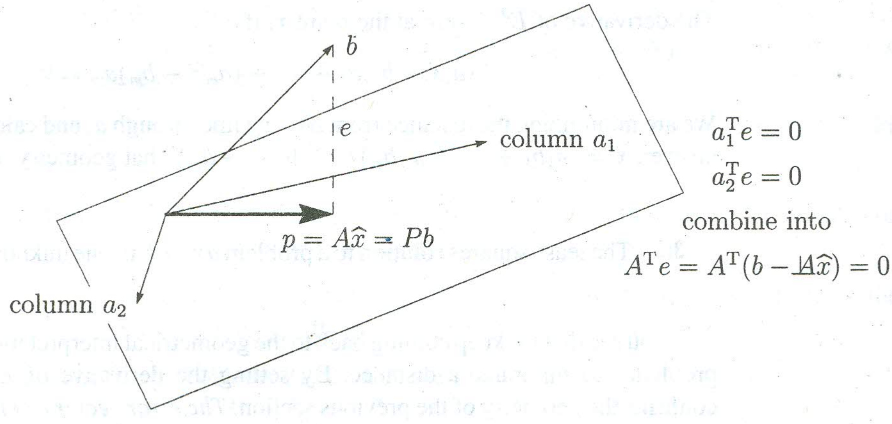

Figure 3.8 Projection onto the column space of a 3 by 2 matrix.

2. The error vector must be perpendicular to *each column* $a_1, \ldots, a_n$ of A:

$$\begin{matrix} a_1^{\mathrm{T}}(b - A\widehat{x}) = 0 \\ \vdots \\ a_n^{\mathrm{T}}(b - A\widehat{x}) = 0 \end{matrix} \quad \text{or} \quad \begin{bmatrix} a_1^{\mathrm{T}} \\ \vdots \\ a_n^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} b - A\widehat{x} \end{bmatrix} = 0.$$

This is again $A^{\mathrm{T}}(b - A\widehat{x}) = 0$ and $A^{\mathrm{T}}A\widehat{x} = A^{\mathrm{T}}b$. The calculus way is to take partial derivatives of $E^2 = (Ax - b)^{\mathrm{T}}(Ax - b)$. That gives the same $2A^{\mathrm{T}}Ax - 2A^{\mathrm{T}}b = 0$. The fastest way is just ***to multiply the unsolvable equation*** $Ax = b$ ***by*** A^{T}. All these equivalent methods produce a square coefficient matrix $A^{\mathrm{T}}A$. It is symmetric (its transpose is not AA^{T}!) and it is the fundamental matrix of this chapter.

The equations $A^{\mathrm{T}}A\widehat{x} = A^{\mathrm{T}}b$ are known in statistics as the ***normal equations***.

3L When $Ax = b$ is inconsistent, its least-squares solution minimizes $\|Ax - b\|^2$:

$$\textbf{Normal equations} \qquad A^{\mathrm{T}}A\widehat{x} = A^{\mathrm{T}}b. \tag{1}$$

$A^{\mathrm{T}}A$ is invertible exactly when the columns of A are linearly independent! Then,

$$\textbf{Best estimate } \widehat{x} \qquad \widehat{x} = (A^{\mathrm{T}}A)^{-1}A^{\mathrm{T}}b. \tag{2}$$

The projection of b onto the column space is the nearest point $A\widehat{x}$:

$$\textbf{Projection} \qquad p = A\widehat{x} = A(A^{\mathrm{T}}A)^{-1}A^{\mathrm{T}}b. \tag{3}$$

We choose an example in which our intuition is as good as the formulas:

$$A = \begin{bmatrix} 1 & 2 \\ 1 & 3 \\ 0 & 0 \end{bmatrix}, \qquad b = \begin{bmatrix} 4 \\ 5 \\ 6 \end{bmatrix}, \qquad \begin{matrix} Ax = b \text{ has no solution} \\ A^{\mathrm{T}}A\widehat{x} = A^{\mathrm{T}}b \text{ gives the best } x. \end{matrix}$$

Both columns end with a zero, so $C(A)$ is the x-y plane within three-dimensional space. The projection of $b = (4, 5, 6)$ is $p = (4, 5, 0)$—the x and y components stay the same but $z = 6$ will disappear. That is confirmed by solving the normal equations:

$$A^{\mathrm{T}}A = \begin{bmatrix} 1 & 1 & 0 \\ 2 & 3 & 0 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 1 & 3 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 2 & 5 \\ 5 & 13 \end{bmatrix}.$$

$$\widehat{x} = (A^{\mathrm{T}}A)^{-1}A^{\mathrm{T}}b = \begin{bmatrix} 13 & -5 \\ -5 & 2 \end{bmatrix} \begin{bmatrix} 1 & 1 & 0 \\ 2 & 3 & 0 \end{bmatrix} \begin{bmatrix} 4 \\ 5 \\ 6 \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}.$$

Projection $$p = A\widehat{x} = \begin{bmatrix} 1 & 2 \\ 1 & 3 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 2 \\ 1 \end{bmatrix} = \begin{bmatrix} 4 \\ 5 \\ 0 \end{bmatrix}.$$

In this special case, the best we can do is to solve the first two equations of $Ax = b$. Then $\widehat{x}_1 = 2$ and $\widehat{x}_2 = 1$. The error in the equation $0x_1 + 0x_2 = 6$ is sure to be 6.

Remark 1 Suppose b is actually *in* the column space of A—it is a combination $b = Ax$ of the columns. Then the projection of b is still b:

b in column space $$p = A(A^{\mathrm{T}}A)^{-1}A^{\mathrm{T}}Ax = Ax = b.$$

The closest point p is just b itself—which is obvious.

Remark 2 At the other extreme, suppose b is *perpendicular* to every column, so $A^{\mathrm{T}}b = 0$. In this case b projects to the zero vector:

b in left nullspace $$p = A(A^{\mathrm{T}}A)^{-1}A^{\mathrm{T}}b = A(A^{\mathrm{T}}A)^{-1}0 = 0.$$

Remark 3 When A is square and invertible, the column space is the whole space. Every vector projects to itself, p equals b, and $\widehat{x} = x$:

If A is invertible $$p = A(A^{\mathrm{T}}A)^{-1}A^{\mathrm{T}}b = AA^{-1}(A^{\mathrm{T}})^{-1}A^{\mathrm{T}}b = b.$$

This is the only case when we can take apart $(A^{\mathrm{T}}A)^{-1}$, *and write it as* $A^{-1}(A^{\mathrm{T}})^{-1}$. When A is rectangular that is not possible.

Remark 4 Suppose A has only one column, containing a. Then the matrix $A^{\mathrm{T}}A$ is the number $a^{\mathrm{T}}a$ and $\widehat{x}$ is $a^{\mathrm{T}}b/a^{\mathrm{T}}a$. We return to the earlier formula.

The Cross-Product Matrix $A^{\mathrm{T}}A$

The matrix $A^{\mathrm{T}}A$ is certainly symmetric. Its transpose is $(A^{\mathrm{T}}A)^{\mathrm{T}} = A^{\mathrm{T}}A^{\mathrm{TT}}$, which is $A^{\mathrm{T}}A$ again. Its i, j entry (and j, i entry) is the inner product of column i of A with column j of A. The key question is the invertibility of $A^{\mathrm{T}}A$, and fortunately

> *$A^{\mathrm{T}}A$ has the same nullspace as A.*

Certainly if $Ax = 0$ then $A^{\mathrm{T}}Ax = 0$. Vectors x in the nullspace of A are also in the nullspace of $A^{\mathrm{T}}A$. To go in the other direction, start by supposing that $A^{\mathrm{T}}Ax = 0$, and take the inner product with x to show that $Ax = 0$:

$$x^T A^T Ax = 0, \quad \text{or} \quad \|Ax\|^2 = 0, \quad \text{or} \quad Ax = 0.$$

The two nullspaces are identical. In particular, if A has independent columns (and only $x = 0$ is in its nullspace), then the same is true for $A^T A$:

3M If A has independent columns, then $A^T A$ is *square, symmetric*, and *invertible*.

We show later that $A^T A$ is also positive definite (all pivots and eigenvalues are positive).

This case is by far the most common and most important. Independence is not so hard in m-dimensional space if $m > n$. We assume it in what follows.

Projection Matrices

We have shown that the closest point to b is $p = A(A^T A)^{-1} A^T b$. *This formula expresses in matrix terms the construction of a perpendicular line from b to the column space of* A. The matrix that gives p is a projection matrix, denoted by P:

$$\textbf{Projection matrix} \qquad P = A(A^T A)^{-1} A^T. \tag{4}$$

This matrix projects any vector b onto the column space of A.* In other words, $p = Pb$ is the component of b in the column space, and the error $e = b - Pb$ is the component in the orthogonal complement. ($I - P$ is also a projection matrix! It projects b onto the orthogonal complement, and the projection is $b - Pb$.)

In short, we have a matrix formula for splitting any b into two perpendicular components. Pb is in the column space $\boldsymbol{C}(A)$, and the other component $(I - P)b$ is in the left nullspace $\boldsymbol{N}(A^T)$—which is orthogonal to the column space.

These projection matrices can be understood geometrically and algebraically.

3N The projection matrix $P = A(A^T A)^{-1} A^T$ has two basic properties:

(i) It equals its square: $P^2 = P$.

(ii) It equals its transpose: $P^T = P$.

Conversely, any symmetric matrix with $P^2 = P$ represents a projection.

Proof It is easy to see why $P^2 = P$. If we start with any b, then Pb lies in the subspace we are projecting onto. ***When we project again nothing is changed***. The vector Pb is already in the subspace, and $P(Pb)$ is still Pb. In other words $P^2 = P$. Two or three or fifty projections give the same point p as the first projection:

$$P^2 = A(A^T A)^{-1} A^T A(A^T A)^{-1} A^T = A(A^T A)^{-1} A^T = P.$$

* There may be a risk of confusion with permutation matrices, also denoted by P, but the risk should be small, and we try never to let both appear on the same page.

To prove that P is also symmetric, take its transpose. Multiply the transposes in reverse order, and use symmetry of $(A^{\mathrm{T}}A)^{-1}$, to come back to P:

$$P^{\mathrm{T}} = (A^{\mathrm{T}})^{\mathrm{T}}((A^{\mathrm{T}}A)^{-1})^{\mathrm{T}}A^{\mathrm{T}} = A((A^{\mathrm{T}}A)^{\mathrm{T}})^{-1}A^{\mathrm{T}} = A(A^{\mathrm{T}}A)^{-1}A^{\mathrm{T}} = P.$$

For the converse, we have to deduce from $P^2 = P$ and $P^{\mathrm{T}} = P$ that **Pb is the projection of b onto the column space of P**. The error vector $b - Pb$ is *orthogonal to the space*. For any vector Pc in the space, the inner product is zero:

$$(b - Pb)^{\mathrm{T}}Pc = b^{\mathrm{T}}(I - P)^{\mathrm{T}}Pc = b^{\mathrm{T}}(P - P^2)c = 0.$$

Thus $b - Pb$ is orthogonal to the space, and Pb is the projection onto the column space. ■

Example 1 Suppose A is actually invertible. If it is 4 by 4, then its four columns are independent and its column space is all of $\mathbf{R}^4$. What is the projection *onto the whole space*? It is the identity matrix.

$$P = A(A^{\mathrm{T}}A)^{-1}A^{\mathrm{T}} = AA^{-1}(A^{\mathrm{T}})^{-1}A^{\mathrm{T}} = I. \tag{5}$$

The identity matrix is symmetric, $I^2 = I$, and the error $b - Ib$ is zero.

The point of all other examples is that what happened in equation (5) is *not allowed*. To repeat: We cannot invert the separate parts A^{T} and A when those matrices are rectangular. It is the square matrix $A^{\mathrm{T}}A$ that is invertible.

Least-Squares Fitting of Data

Suppose we do a series of experiments, and expect the output b to be a linear function of the input t. We look for a ***straight line*** $b = C + Dt$. For example:

1. At different times we measure the distance to a satellite on its way to Mars. In this case t is the time and b is the distance. Unless the motor was left on or gravity is strong, the satellite should move with nearly constant velocity v: $b = b_0 + vt$.
2. We vary the load on a structure, and measure the movement it produces. In this experiment t is the load and b is the reading from the strain gauge. Unless the load is so great that the material becomes plastic, a linear relation $b = C + Dt$ is normal in the theory of elasticity.
3. The cost of producing t books like this one is nearly linear, $b = C + Dt$, with editing and typesetting in C and then printing and binding in D. C is the set-up cost and D is the cost for each additional book.

How to compute C and D? If there is no experimental error, then two measurements of b will determine the line $b = C + Dt$. But if there is error, we must be prepared to "average" the experiments and find an optimal line. That line is not to be confused with the line through a on which b was projected in the previous section! In fact, since there are two unknowns C and D to be determined, we now project onto a *two-dimensional* subspace. A perfect experiment would give a perfect C and D:

$$\begin{aligned} C + Dt_1 &= b_1 \\ C + Dt_2 &= b_2 \\ &\vdots \\ C + Dt_m &= b_m. \end{aligned} \tag{6}$$

This is an *overdetermined* system, with m equations and only two unknowns. If errors are present, it will have no solution. A has two columns, and $x = (C, D)$:

$$\begin{bmatrix} 1 & t_1 \\ 1 & t_2 \\ \vdots & \vdots \\ 1 & t_m \end{bmatrix} \begin{bmatrix} C \\ D \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix}, \qquad \text{or} \qquad Ax = b. \tag{7}$$

The best solution $(\widehat{C}, \widehat{D})$ is the $\widehat{x}$ that minimizes the squared error E^2:

$$\textbf{Minimize} \qquad E^2 = \|b - Ax\|^2 = (b_1 - C - Dt_1)^2 + \cdots + (b_m - C - Dt_m)^2.$$

The vector $p = A\widehat{x}$ is as close as possible to b. Of all straight lines $b = C + Dt$, we are choosing the one that best fits the data (Figure 3.9). On the graph, the errors are the ***vertical distances*** $b - C - Dt$ to the straight line (not perpendicular distances!). It is the vertical distances that are squared, summed, and minimized.

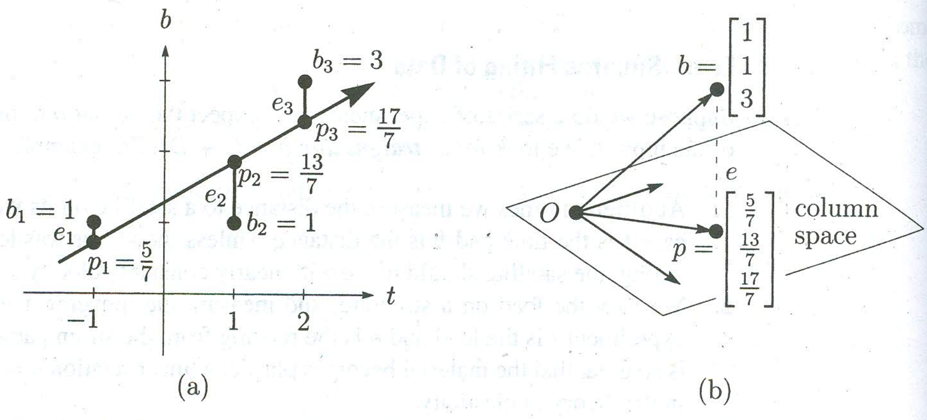

Figure 3.9 Straight-line approximation matches the projection p of b.

Example 2 Three measurements b_1, b_2, b_3 are marked on Figure 3.9a:

$$b = 1 \quad \text{at} \quad t = -1, \qquad b = 1 \quad \text{at} \quad t = 1, \qquad b = 3 \quad \text{at} \quad t = 2.$$

Note that the values $t = -1, 1, 2$ are not required to be equally spaced. The first step is *to write the equations that **would** hold if a line could go through all three points.*

Then every $C + Dt$ would agree exactly with b:

$$Ax = b \quad \text{is} \quad \begin{aligned} C - D &= 1 \\ C + D &= 1 \\ C + 2D &= 3 \end{aligned} \quad \text{or} \quad \begin{bmatrix} 1 & -1 \\ 1 & 1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} C \\ D \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 3 \end{bmatrix}.$$

If those equations $Ax = b$ could be solved, there would be no errors. They can't be solved because the points are not on a line. Therefore they are solved by least squares:

$$A^{\mathrm{T}}A\widehat{x} = A^{\mathrm{T}}b \quad \text{is} \quad \begin{bmatrix} 3 & 2 \\ 2 & 6 \end{bmatrix} \begin{bmatrix} \widehat{C} \\ \widehat{D} \end{bmatrix} = \begin{bmatrix} 5 \\ 6 \end{bmatrix}.$$

The best solution is $\widehat{C} = \frac{9}{7}$, $\widehat{D} = \frac{4}{7}$ and the best line is $\frac{9}{7} + \frac{4}{7}t$.

Note the beautiful connections between the two figures. The problem is the same but the art shows it differently. In Figure 3.9b, b is not a combination of the columns $(1, 1, 1)$ and $(-1, 1, 2)$. In Figure 3.9, the three points are not on a line. Least squares replaces points b that are not on a line by points p that are! Unable to solve $Ax = b$, we solve $A\widehat{x} = p$.

The line $\frac{9}{7} + \frac{4}{7}t$ has heights $\frac{5}{7}, \frac{13}{7}, \frac{17}{7}$ at the measurement times $-1, 1, 2$. ***Those points do lie on a line.*** Therefore the vector $p = \left(\frac{5}{7}, \frac{13}{7}, \frac{17}{7}\right)$ is in the column space. *This vector is the projection.* Figure 3.9b is in three dimensions (or m dimensions if there are m points) and Figure 3.9a is in two dimensions (or n dimensions if there are n parameters).

Subtracting p from b, the errors are $e = \left(\frac{2}{7}, -\frac{6}{7}, \frac{4}{7}\right)$. Those are the vertical errors in Figure 3.9a, and they are the components of the dashed vector in Figure 3.9b. This error vector is orthogonal to the first column $(1, 1, 1)$, since $\frac{2}{7} - \frac{6}{7} + \frac{4}{7} = 0$. It is orthogonal to the second column $(-1, 1, 2)$, because $-\frac{2}{7} - \frac{6}{7} + \frac{8}{7} = 0$. It is orthogonal to the column space, and it is in the left nullspace.

Question: If the measurements $b = \left(\frac{2}{7}, -\frac{6}{7}, \frac{4}{7}\right)$ were those errors, what would be the best line and the best $\widehat{x}$? Answer: The zero line—which is the horizontal axis—and $\widehat{x} = 0$. Projection to zero.

We can quickly summarize the equations for fitting by a straight line. The first column of A contains 1s, and the second column contains the times t_i. Therefore $A^{\mathrm{T}}A$ contains the sum of the 1s and the t_i and the t_i^2:

30 The measurements $b_1, \ldots, b_m$ are given at distinct points $t_1, \ldots, t_m$. Then the straight line $\widehat{C} + \widehat{D}t$ which minimizes E^2 comes from least squares:

$$A^{\mathrm{T}}A \begin{bmatrix} \widehat{C} \\ \widehat{D} \end{bmatrix} = A^{\mathrm{T}}b \quad \text{or} \quad \begin{bmatrix} m & \sum t_i \\ \sum t_i & \sum t_i^2 \end{bmatrix} \begin{bmatrix} \widehat{C} \\ \widehat{D} \end{bmatrix} = \begin{bmatrix} \sum b_i \\ \sum t_i b_i \end{bmatrix}.$$

Remark The mathematics of least squares is not limited to fitting the data by straight lines. In many experiments there is no reason to expect a linear relationship, and it would be crazy to look for one. Suppose we are handed some radioactive material. The output b will be the reading on a Geiger counter at various times t. We may know that we are holding a mixture of two chemicals, and we may know their half-lives (or rates of decay), but we do not know how much of each is in our hands. If these two unknown amounts are C and D, then the Geiger counter readings would behave like the sum of two exponentials (and not like a straight line):

$$b = Ce^{-\lambda t} + De^{-\mu t}. \tag{8}$$

In practice, the Geiger counter is not exact. Instead, we make readings $b_1, \ldots, b_m$ at times $t_1, \ldots, t_m$, and equation (8) is approximately satisfied:

$$Ax = b \quad \text{is} \quad \begin{array}{c} Ce^{-\lambda t_1} + De^{-\mu t_1} \approx b_1 \\ \vdots \\ Ce^{-\lambda t_m} + De^{-\mu t_m} \approx b_m. \end{array}$$

If there are more than two readings, $m > 2$, then in all likelihood we cannot solve for C and D. But the least-squares principle will give optimal values $\widehat{C}$ and $\widehat{D}$.

The situation would be completely different if we knew the amounts C and D, and were trying to discover the decay rates λ and μ. This is a problem in *nonlinear least squares*, and it is harder. We would still form E^2, the sum of the squares of the errors, and minimize it. But setting its derivatives to zero will not give linear equations for the optimal λ and μ. In the exercises, we stay with linear least squares.

Weighted Least Squares

A simple least-squares problem is the estimate $\widehat{x}$ of a patient's weight from two observations $x = b_1$ and $x = b_2$. Unless $b_1 = b_2$, we are faced with an inconsistent system of two equations in one unknown:

$$\begin{bmatrix} 1 \\ 1 \end{bmatrix} [x] = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}.$$

Up to now, we accepted b_1 and b_2 as equally reliable. We looked for the value $\widehat{x}$ that minimized $E^2 = (x - b_1)^2 + (x - b_2)^2$:

$$\frac{dE^2}{dx} = 0 \quad \text{at} \quad \widehat{x} = \frac{b_1 + b_2}{2}.$$

The optimal $\widehat{x}$ is the average. The same conclusion comes from $A^{\mathrm{T}}A\widehat{x} = A^{\mathrm{T}}b$. In fact $A^{\mathrm{T}}A$ is a 1 by 1 matrix, and the normal equation is $2\widehat{x} = b_1 + b_2$.

Now suppose the two observations are not trusted to the same degree. The value $x = b_1$ may be obtained from a more accurate scale—or, in a statistical problem, from a larger sample—than $x = b_2$. Nevertheless, if b_2 contains some information, we are not willing to rely totally on b_1. The simplest compromise is to attach different weights w_1^2 and w_2^2, and choose the $\widehat{x}_W$ that minimizes the *weighted sum of squares*:

$$\textbf{Weighted error} \qquad E^2 = w_1^2(x - b_1)^2 + w_2^2(x - b_2)^2.$$

If $w_1 > w_2$, more importance is attached to b_1. The minimizing process (derivative $= 0$) tries harder to make $(x - b_1)^2$ small:

$$\frac{dE^2}{dx} = 2\left[w_1^2(x - b_1) + w_2^2(x - b_2)\right] = 0 \quad \text{at} \quad \widehat{x}_W = \frac{w_1^2 b_1 + w_2^2 b_2}{w_1^2 + w_2^2}. \tag{9}$$