Interactive Linear Algebra
Dan Margalit, Joseph Rabinoff

PDF version

6.5 The Method of Least Squares ¶ permalink

Objectives

1. Learn examples of best-fit problems.
2. Learn to turn a best-fit problem into a least-squares problem.
3. *Recipe:* find a least-squares solution (two ways).
4. *Picture:* geometry of a least-squares solution.
5. *Vocabulary words:* ***least-squares solution***.

In this section, we answer the following important question:

> Suppose that $Ax = b$ does not have a solution. What is the best approximate solution?

For our purposes, the best approximate solution is called the *least-squares solution*. We will present two methods for finding least-squares solutions, and we will give several applications to best-fit problems.

Least-Squares Solutions

We begin by clarifying exactly what we will mean by a "best approximate solution" to an inconsistent matrix equation $Ax = b$.

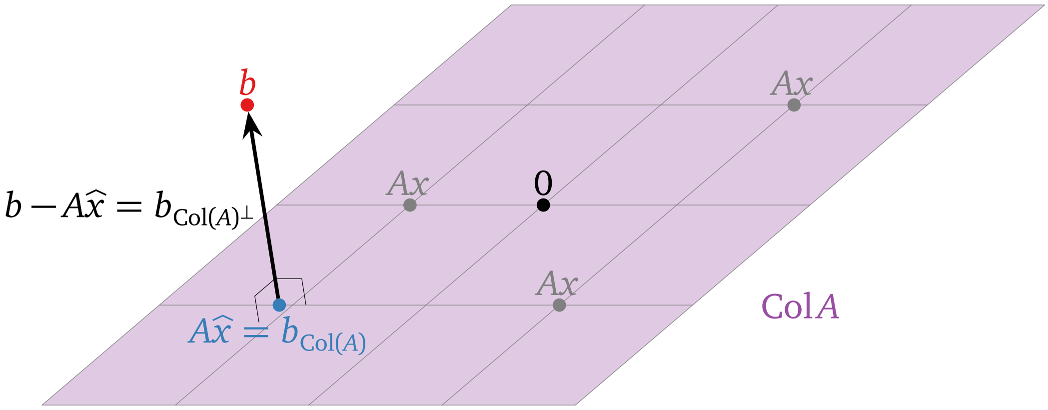
Definition. Let A be an $m \times n$ matrix and let b be a vector in $\mathbf{R}^m$. A ***least-squares solution*** of the matrix equation $Ax = b$ is a vector $\widehat{x}$ in $\mathbf{R}^n$ such that

$$\text{dist}(b, A\widehat{x}) \leq \text{dist}(b, Ax)$$

for all other vectors x in $\mathbf{R}^n$.

Recall that $\text{dist}(v, w) = \|v - w\|$ is the distance between the vectors v and w. The term "least squares" comes from the fact that $\text{dist}(b, Ax) = \|b - A\widehat{x}\|$ is the square root of the sum of the squares of the entries of the vector $b - A\widehat{x}$. So a least-squares solution minimizes the sum of the squares of the differences between the entries of $A\widehat{x}$ and b. In other words, a least-squares solution solves the equation $Ax = b$ as closely as possible, in the sense that the sum of the squares of the difference $b - Ax$ is minimized.

Least Squares: Picture. Suppose that the equation $Ax = b$ is inconsistent. Recall from this note in Section 2.3 that the column space of A is the set of all other vectors c such that $Ax = c$ is consistent. In other words, $\mathrm{Col}(A)$ is the set of all vectors of the form Ax. Hence, the closest vector of the form Ax to b is the orthogonal projection of b onto $\mathrm{Col}(A)$. This is denoted $b_{\mathrm{Col}(A)}$, following this notation in Section 6.3.



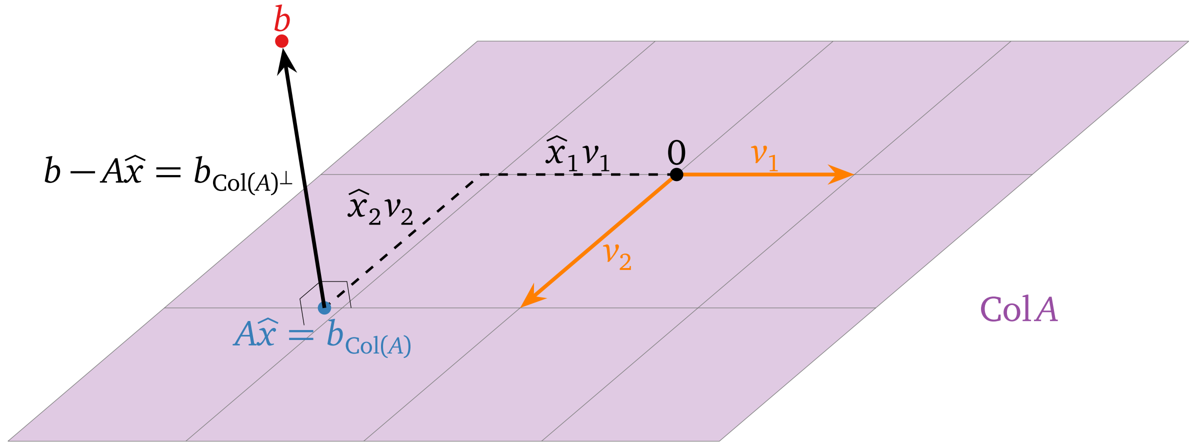
A least-squares solution of $Ax = b$ is a solution $\widehat{x}$ of the consistent equation $Ax = b_{\mathrm{Col}(A)}$

Note. If $Ax = b$ is consistent, then $b_{\mathrm{Col}(A)} = b$, so that a least-squares solution is the same as a usual solution.

Where is $\widehat{x}$ in this picture? If $v_1, v_2, \ldots, v_n$ are the columns of A, then

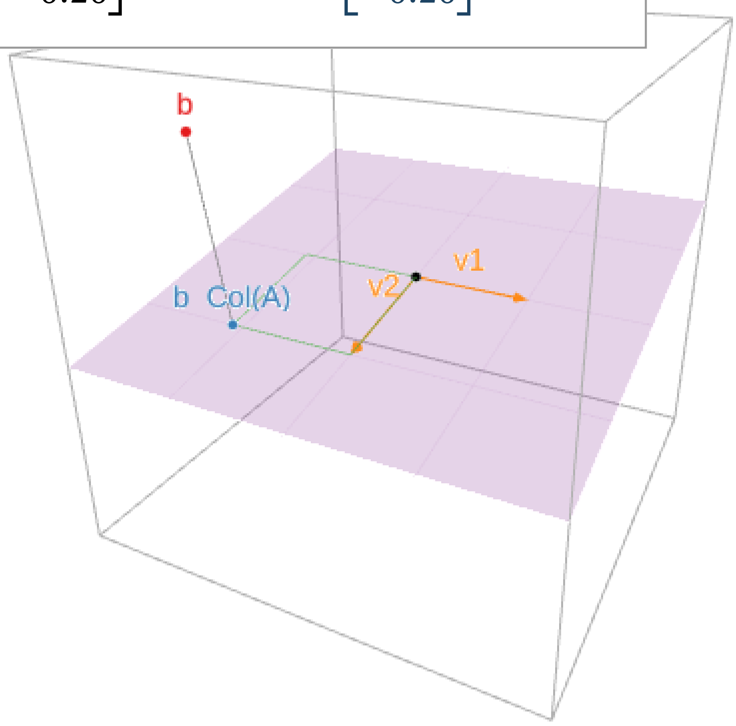
$$A\widehat{x} = A\begin{pmatrix}\widehat{x}_1\\ \widehat{x}_2\\ \vdots\\ \widehat{x}_n\end{pmatrix} = \widehat{x}_1 v_1 + \widehat{x}_2 v_2 + \cdots + \widehat{x}_n v_n.$$

Hence the entries of $\widehat{x}$ are the "coordinates" of $b_{\mathrm{Col}(A)}$ with respect to the spanning set $\{v_1, v_2, \ldots, v_m\}$ of $\mathrm{Col}(A)$. (They are honest $\mathcal{B}$-coordinates if the columns of A are linearly independent.)



$-1.10\, v_1 + 1.00\, v_2 = b_{\text{Col}(A)}$

$$A\hat{x} = \begin{bmatrix} 0.00 & 1.10 \\ 1.00 & 0.00 \\ 0.00 & -0.20 \end{bmatrix} \begin{bmatrix} -1.10 \\ 1.00 \end{bmatrix} = \begin{bmatrix} 1.10 \\ -1.10 \\ -0.20 \end{bmatrix} = b_{\text{Col}(A)}$$



The violet plane is Col(A). *The closest that Ax can get to b is the closest vector on* Col(A) *to b, which is the orthogonal projection $b_{\text{Col}(A)}$ (in blue). The vectors v_1, v_2 are the columns of A, and the coefficients of $\widehat{x}$ are the lengths of the green lines. Click and drag b to move it.*

Note. If $Ax = b$ is consistent, then $b_{\text{Col}(A)} = b$, so that a least-squares solution is the same as a usual solution.

We learned to solve this kind of orthogonal projection problem in Section 6.3.

Theorem. *Let A be an $m \times n$ matrix and let b be a vector in $\mathbf{R}^m$. The least-squares solutions of $Ax = b$ are the solutions of the matrix equation*

$$A^T Ax = A^T b$$

Proof.

In particular, finding a least-squares solution means solving a consistent system of linear equations. We can translate the above theorem into a recipe:

Recipe 1: Compute a least-squares solution.

Let A be an $m \times n$ matrix and let b be a vector in $\mathbf{R}^n$. Here is a method for computing a least-squares solution of $Ax = b$:

1. Compute the matrix $A^T A$ and the vector $A^T b$.
2. Form the augmented matrix for the matrix equation $A^T Ax = A^T b$, and row reduce.
3. This equation is always consistent, and any solution $\hat{x}$ is a least-squares solution.

To reiterate: once you have found a least-squares solution $\hat{x}$ of $Ax = b$, then $b_{\mathrm{Col}(A)}$ is equal to $A\hat{x}$.

Example.

Example.

The reader may have noticed that we have been careful to say "the least-squares solutions" in the plural, and "a least-squares solution" using the indefinite article. This is because a least-squares solution need not be unique: indeed, if the columns of A are linearly dependent, then $Ax = b_{\mathrm{Col}(A)}$ has infinitely many solutions. The following theorem, which gives equivalent criteria for uniqueness, is an analogue of this corollary in Section 6.3.

Theorem. *Let A be an $m \times n$ matrix and let b be a vector in $\mathbf{R}^m$. The following are equivalent:*

1. *$Ax = b$ has a unique least-squares solution.*
2. *The columns of A are linearly independent.*
3. *$A^T A$ is invertible.*

In this case, the least-squares solution is

$$\hat{x} = (A^T A)^{-1} A^T b.$$

Proof.

Example (Infinitely many least-squares solutions).

As usual, calculations involving projections become easier in the presence of an orthogonal set. Indeed, if A is an $m \times n$ matrix with *orthogonal* columns $u_1, u_2, \ldots, u_m$, then we can use the projection formula in Section 6.4 to write

$$b_{\text{Col}(A)} = \frac{b \cdot u_1}{u_1 \cdot u_1} u_1 + \frac{b \cdot u_2}{u_2 \cdot u_2} u_2 + \cdots + \frac{b \cdot u_m}{u_m \cdot u_m} u_m = A \begin{pmatrix} (b \cdot u_1)/(u_1 \cdot u_1) \\ (b \cdot u_2)/(u_2 \cdot u_2) \\ \vdots \\ (b \cdot u_m)/(u_m \cdot u_m) \end{pmatrix}.$$

Note that the least-squares solution is unique in this case, since an orthogonal set is linearly independent.

Recipe 2: Compute a least-squares solution.

Let A be an $m \times n$ matrix with *orthogonal* columns $u_1, u_2, \ldots, u_m$, and let b be a vector in $\mathbf{R}^n$. Then the least-squares solution of $Ax = b$ is the vector

$$\widehat{x} = \left(\frac{b \cdot u_1}{u_1 \cdot u_1}, \frac{b \cdot u_2}{u_2 \cdot u_2}, \ldots, \frac{b \cdot u_m}{u_m \cdot u_m} \right).$$

This formula is particularly useful in the sciences, as matrices with orthogonal columns often arise in nature.

Example.

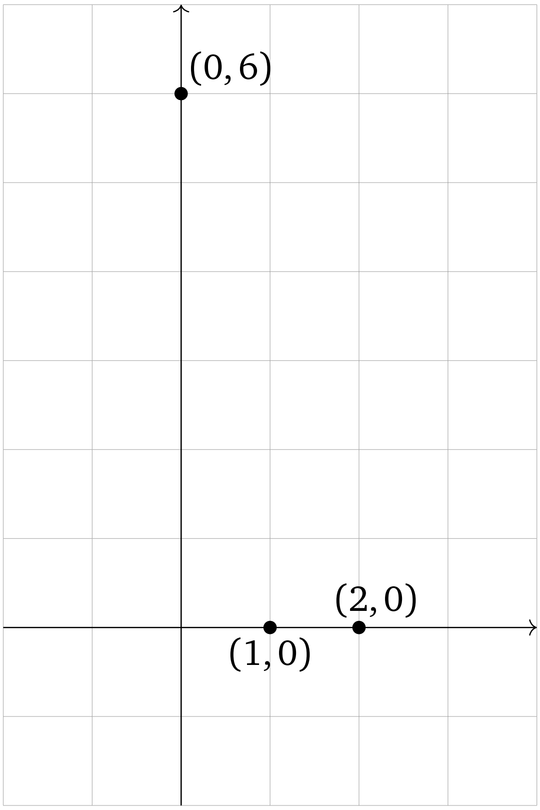
Best-Fit Problems

In this subsection we give an application of the method of least squares to data modeling. We begin with a basic example.

Example (Best-fit line). Suppose that we have measured three data points

$$(0, 6), \quad (1, 0), \quad (2, 0),$$

and that our model for these data asserts that the points should lie on a line. Of course, these three points do not actually lie on a single line, but this could be due to errors in our measurement. How do we predict which line they are supposed to lie on?



The general equation for a (non-vertical) line is

$$y = Mx + B.$$

If our three data points were to lie on this line, then the following equations would be satisfied:

$$\begin{aligned} 6 &= M \cdot 0 + B \\ 0 &= M \cdot 1 + B \\ 0 &= M \cdot 2 + B. \end{aligned} \tag{6.5.1}$$

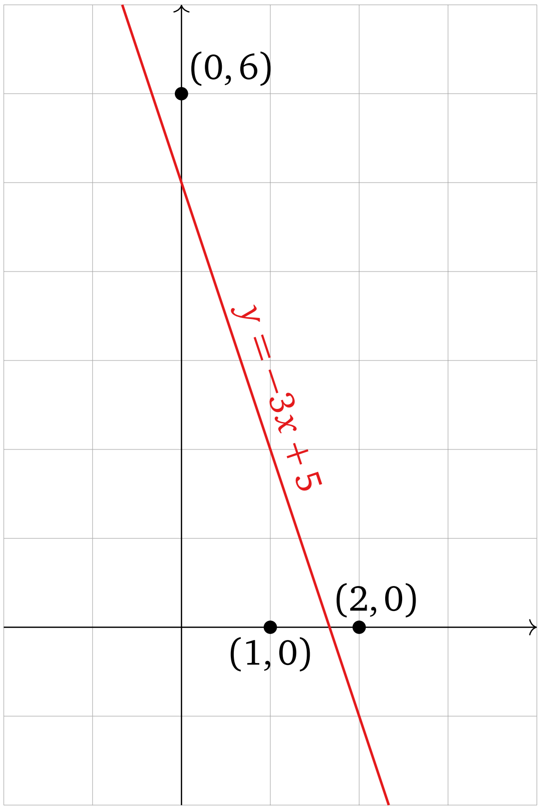
In order to find the best-fit line, we try to solve the above equations in the unknowns M and B. As the three points do not actually lie on a line, there is no actual solution, so instead we compute a least-squares solution.

Putting our linear equations into matrix form, we are trying to solve $Ax = b$ for

$$A = \begin{pmatrix} 0 & 1 \\ 1 & 1 \\ 2 & 1 \end{pmatrix} \qquad x = \begin{pmatrix} M \\ B \end{pmatrix} \qquad b = \begin{pmatrix} 6 \\ 0 \\ 0 \end{pmatrix}.$$

We solved this least-squares problem in this example: the only least-squares solution to $Ax = b$ is $\widehat{x} = \binom{M}{B} = \binom{-3}{5}$, so the best-fit line is

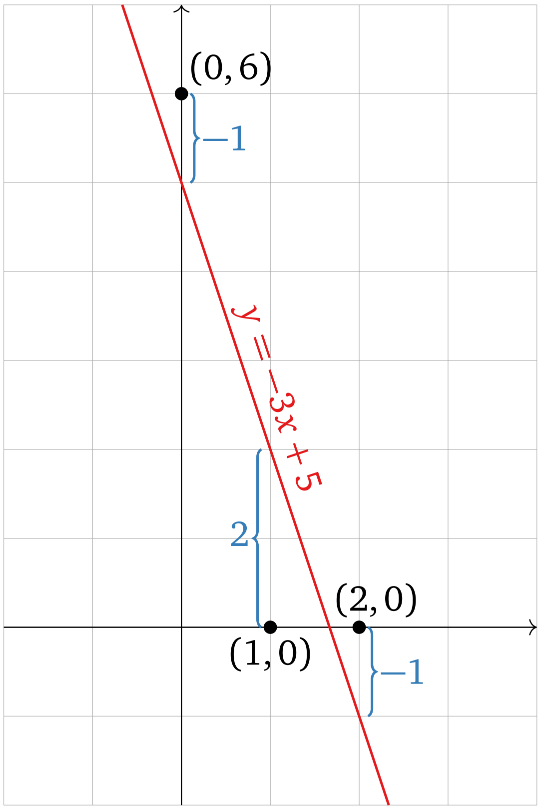
$$y = -3x + 5.$$



What exactly is the line $y = f(x) = -3x + 5$ minimizing? The least-squares solution $\widehat{x}$ minimizes the sum of the squares of the entries of the vector $b - A\widehat{x}$. The vector b is the left-hand side of (6.5.1), and

$$A\begin{pmatrix}-3\\5\end{pmatrix} = \begin{pmatrix}-3(0)+5\\-3(1)+5\\-3(2)+5\end{pmatrix} = \begin{pmatrix}f(0)\\f(1)\\f(2)\end{pmatrix}.$$

In other words, $A\widehat{x}$ is the vector whose entries are the y-coordinates of the graph of the line at the values of x we specified in our data points, and b is the vector whose entries are the y-coordinates of those data points. The difference $b - A\widehat{x}$ is the vertical distance of the graph from the data points:



$$b - A\hat{x} = \begin{pmatrix} 6 \\ 0 \\ 0 \end{pmatrix} - A\begin{pmatrix} -3 \\ 5 \end{pmatrix} = \begin{pmatrix} -1 \\ 2 \\ -1 \end{pmatrix}$$

The best-fit line minimizes the sum of the squares of these vertical distances.

Interactive: Best-fit line.

Example (Best-fit parabola).

Example (Best-fit linear function).

All of the above examples have the following form: some number of data points (x, y) are specified, and we want to find a function

$$y = B_1 g_1(x) + B_2 g_2(x) + \cdots + B_m g_m(x)$$

that best approximates these points, where $g_1, g_2, \ldots, g_m$ are fixed functions of x. Indeed, in the best-fit line example we had $g_1(x) = x$ and $g_2(x) = 1$; in the best-fit parabola example we had $g_1(x) = x^2$, $g_2(x) = x$, and $g_3(x) = 1$; and in the best-fit linear function example we had $g_1(x_1, x_2) = x_1$, $g_2(x_1, x_2) = x_2$, and $g_3(x_1, x_2) = 1$ (in this example we take x to be a vector with two entries). We evaluate the above equation on the given data points to obtain a system of linear equations in the unknowns $B_1, B_2, \ldots, B_m$—once we evaluate the g_i, they just become numbers, so it does not matter what they are—and we find the least-squares solution. The resulting best-fit function minimizes the sum of the squares of the vertical distances from the graph of $y = f(x)$ to our original data points.

To emphasize that the nature of the functions g_i really is irrelevant, consider the following example.

Example (Best-fit trigonometric function).

The next example has a somewhat different flavor from the previous ones.

Example (Best-fit ellipse).

> **Note.** Gauss invented the method of least squares to find a best-fit ellipse: he correctly predicted the (elliptical) orbit of the asteroid Ceres as it passed behind the sun in 1801.

Comments, corrections or suggestions?
(Free GitHub account required)