

# L22LeastSquares

December 1, 2015

## 1 Least Squares



Let's go back to week 1. A long time ago!

Recall Gauss's remarkable accomplishment in his early 20s. He took the set of measurements made by Piazzi of the dwarf planet Ceres and predicted where Ceres subsequently would appear in the sky (after it was lost behind the sun). This told Olbers exactly where to look, and lo and behold . . .

We can understand now a little better what Gauss had to do.

Kepler had discovered, and Newton had explained, that each planet orbits the sun following the path of an ellipse.

To describe the orbit of Ceres, Gauss had to construct the equation for its ellipse:

$$a_1x_1^2 + a_2x_2^2 + a_3x_1x_2 + a_4x_1 + a_5x_2 + a_6 = 0.$$

He had many measurements of  $(x_1, x_2)$  pairs and had to find the  $a_1, \dots, a_6$ .

This is actually a linear system:

$$\begin{bmatrix} x_{11}^2 & x_{21}^2 & x_{11}x_{21} & x_{11} & x_{21} & 1 \\ x_{12}^2 & x_{22}^2 & x_{12}x_{22} & x_{12} & x_{22} & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{1n}^2 & x_{2n}^2 & x_{1n}x_{2n} & x_{1n} & x_{2n} & 1 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \\ a_5 \\ a_6 \end{bmatrix} = \mathbf{0}$$

Now, according to Newton, this is a consistent linear system. The equation for the ellipse is exactly correct and all we need is six  $(x_1, x_2)$  sets of measurements to know the orbit of Ceres exactly.

What could go wrong? :)

Obviously, there are going to be measurement errors in Piazzi's observations. If we just solve the system using six measurements, we will probably get incorrect values for the coefficients  $a_1, \dots, a_6$ .

A better idea is to use all of the  $n$  measurement data available, and try to find a way to cancel out errors. So, using all the  $n$  data measurements available, we construct a linear system:

$$X\mathbf{a} = \mathbf{b}$$

where  $X$  is  $n \times 6$  and  $\mathbf{b} \in \mathbb{R}^n$ .

But now, due to measurement errors, we can't expect  $\mathbf{b}$  will lie in the column space of  $X$ . We have an inconsistent system.

This system has **no solutions!**

What can we do if  $A\mathbf{x} = \mathbf{b}$  has **no solutions?**

We now understand if  $A$  is  $m \times n$  and  $A\mathbf{x} = \mathbf{b}$  has no solutions, that is because the columns of  $A$  do not span  $\mathbb{R}^m$ , and  $\mathbf{b}$  is not in the column space of  $A$ .

In many cases we will be quite satisfied to find an  $\mathbf{x}$  that makes  $A\mathbf{x}$  as close as possible to  $\mathbf{b}$ .

In other words, we are looking for an  $\mathbf{x}$  such that  $A\mathbf{x}$  makes a good **approximation** to  $\mathbf{b}$ .

We can think of the quality of the approximation of  $A\mathbf{x}$  to  $\mathbf{b}$  as the distance from  $A\mathbf{x}$  to  $\mathbf{b}$ , which is

$$\|A\mathbf{x} - \mathbf{b}\|.$$

The **general least-squares problem** is to find an  $\mathbf{x}$  that makes  $\|A\mathbf{x} - \mathbf{b}\|$  as small as possible.

This is called "least squares" because it is equivalent to minimizing  $\|A\mathbf{x} - \mathbf{b}\|^2$ , which is the sum of squared differences.

Just to make this explicit: say that we denote  $A\mathbf{x}$  by  $\mathbf{y}$ . Then

$$\|A\mathbf{x} - \mathbf{b}\|^2 = \sum_i (y_i - b_i)^2$$

Where we interpret  $y_i$  as the *estimated value* and  $b_i$  as the *measured value*.

So this expression is the **sum of squared error**. This is the most common measure of error used in statistics.

This is a key principle!

**Minimizing the length of  $A\mathbf{x} - \mathbf{b}$  is the same as minimizing the sum of squared error.**

**Definition.** If  $A$  is  $m \times n$  and  $\mathbf{b}$  is in  $\mathbb{R}^m$ , a **least squares solution** of  $A\mathbf{x} = \mathbf{b}$  is an  $\hat{\mathbf{x}}$  in  $\mathbb{R}^n$  such that

$$\|A\hat{\mathbf{x}} - \mathbf{b}\| \leq \|A\mathbf{x} - \mathbf{b}\|$$

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

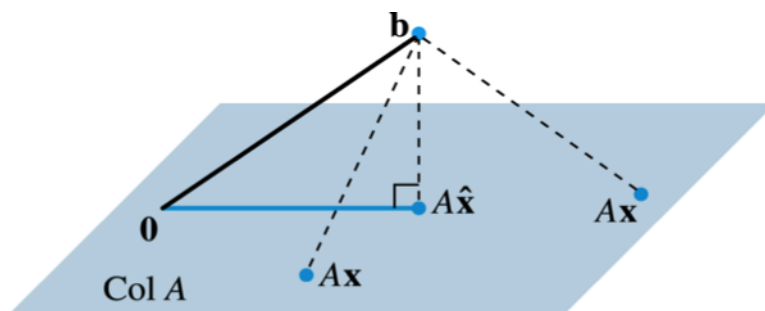
An equivalent (and more common) way to express this is:

$$\hat{\mathbf{x}} = \arg \min_{\mathbf{x}} \|A\mathbf{x} - \mathbf{b}\|.$$

which emphasizes that this is a minimization problem, also called an *optimization* problem.

**Interpretation of the Least Squares Problem.** The point to remember is that no matter what  $\mathbf{x}$  is,  $A\mathbf{x}$  will be in the column space of  $A$ , Col  $A$ .

So  $\mathbf{b}$  is outside Col  $A$ , and we are looking for  $\mathbf{x}$  that specifies the closest point in Col  $A$  to  $\mathbf{b}$ .



The vector  $\mathbf{b}$  is closer to  $A\hat{\mathbf{x}}$  than to  $A\mathbf{x}$  for other  $\mathbf{x}$ .

### Solving the General Least Squares Problem.

The last lecture developed methods for finding the point in a 1D subspace that is closest to a given point.

We just need to generalize what we learned to the case of an arbitrary subspace. This leads to two theorems: the **Orthogonal Decomposition Theorem** and the **Best Approximation Theorem**.

**The Orthogonal Decomposition Theorem.** Let  $W$  be a subspace of  $\mathbb{R}^n$ . Then each  $\mathbf{y}$  in  $\mathbb{R}^n$  can be written uniquely in the form

$$\mathbf{y} = \hat{\mathbf{y}} + \mathbf{z}$$

where  $\hat{\mathbf{y}}$  is in  $W$  and  $\mathbf{z}$  is orthogonal to every vector in  $W$ .

**Proof.** In the book (straightforward).

Just as in the case of a 1D subspace (in the last lecture), we say that  $\hat{\mathbf{y}}$  is the **orthogonal projection of  $\mathbf{y}$  onto  $W$**  and write  $\hat{\mathbf{y}} = \text{proj}_W \mathbf{y}$ .

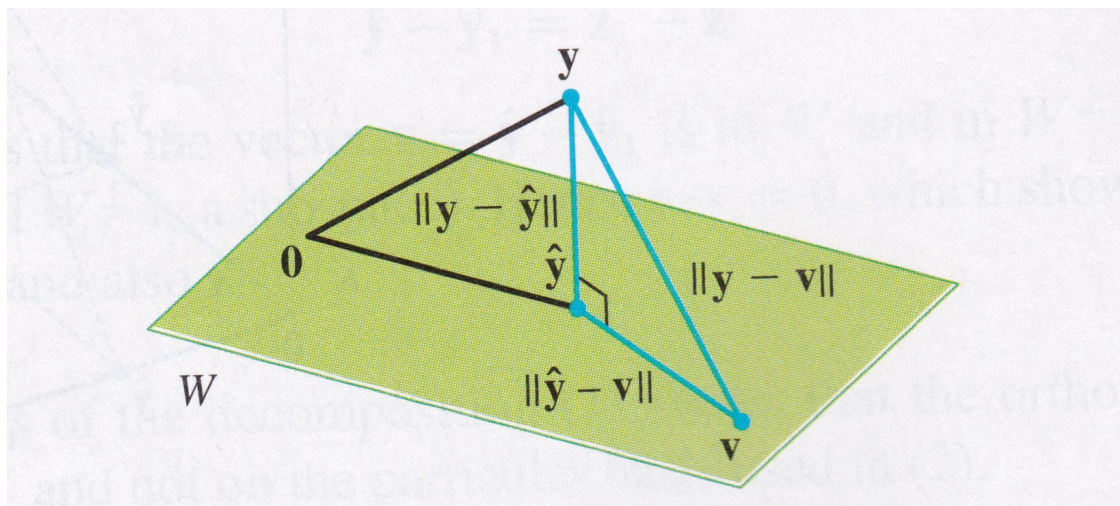
**The Best Approximation Theorem.** Let  $W$  be a subspace of  $\mathbb{R}^n$ , let  $\mathbf{y}$  be any vector in  $\mathbb{R}^n$ , and let  $\hat{\mathbf{y}}$  be the orthogonal projection of  $\mathbf{y}$  onto  $W$ . Then  $\hat{\mathbf{y}}$  is the closest point in  $W$  to  $\mathbf{y}$ , in the sense that

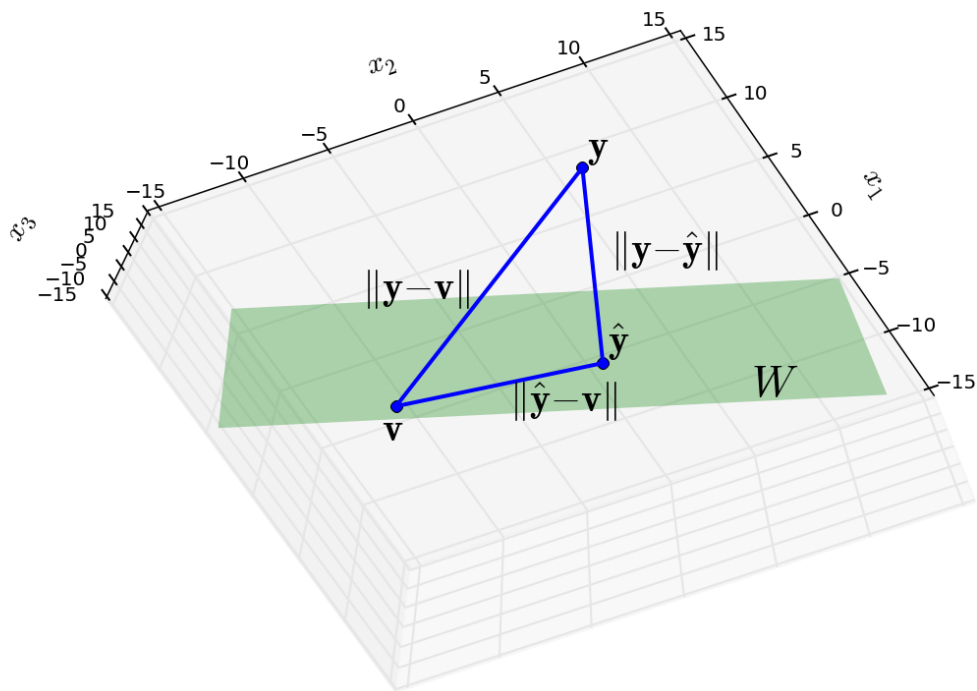
$$\|\mathbf{y} - \hat{\mathbf{y}}\| < \|\mathbf{y} - \mathbf{v}\|$$

for all  $\mathbf{v}$  distinct from  $\hat{\mathbf{y}}$ .

**Proof.**

Take  $\mathbf{v}$  in  $W$  distinct from  $\hat{\mathbf{y}}$ . Here is what the setup looks like:





Both  $\hat{\mathbf{y}}$  and  $\mathbf{v}$  are in  $W$ , so  $\hat{\mathbf{y}} - \mathbf{v}$  is in  $W$ .

By the orthogonal decomposition theorem,  $\mathbf{y} - \hat{\mathbf{y}}$  is orthogonal to every vector in  $W$ , so it is orthogonal to  $\hat{\mathbf{y}} - \mathbf{v}$ .

Now, these three points form a (right) triangle because

$$\mathbf{y} - \mathbf{v} = (\mathbf{y} - \hat{\mathbf{y}}) + (\hat{\mathbf{y}} - \mathbf{v}).$$

So the Pythagorean Theorem tells us that

$$\|\mathbf{y} - \mathbf{v}\|^2 = \|\mathbf{y} - \hat{\mathbf{y}}\|^2 + \|\hat{\mathbf{y}} - \mathbf{v}\|^2.$$

Now  $\hat{\mathbf{y}} - \mathbf{v} \neq \mathbf{0}$  because  $\mathbf{y}$  is distinct from  $\mathbf{v}$ .

So

$$\|\hat{\mathbf{y}} - \mathbf{v}\| > 0.$$

So

$$\|\mathbf{y} - \mathbf{v}\|^2 > \|\mathbf{y} - \hat{\mathbf{y}}\|^2.$$

Let's apply these ideas to solving the least squares problem:

$$\hat{\mathbf{x}} = \arg \min_{\mathbf{x}} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|.$$

We know that the closest point to  $\mathbf{b}$  in a subspace  $W$  is the **projection** of  $\mathbf{b}$  onto  $W$ .

So the point we are looking for, which we'll call  $\hat{\mathbf{b}}$ , is:

$$\hat{\mathbf{b}} = \text{proj}_{\text{Col } A} \mathbf{b}$$

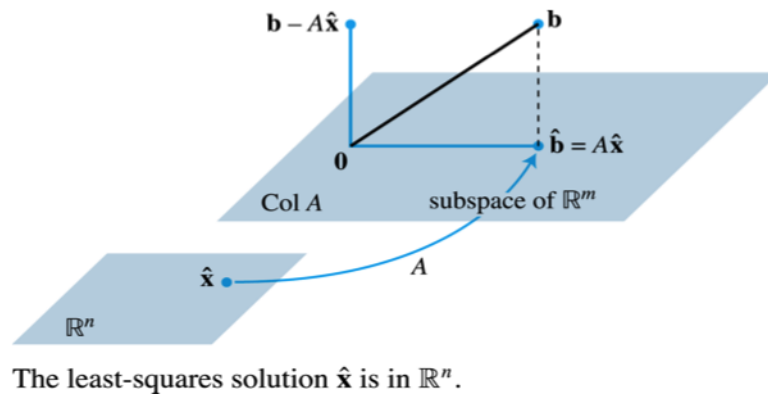
The key is that  $\hat{\mathbf{b}}$  is in the column space of  $A$ . So this equation is consistent, and we can solve it:

$$A\hat{\mathbf{x}} = \hat{\mathbf{b}}.$$

Since  $\hat{\mathbf{b}}$  is the closest point in  $\text{Col } A$  to  $\mathbf{b}$ , a vector  $\hat{\mathbf{x}}$  is a least-squares solution of  $A\mathbf{x} = \mathbf{b}$  if and only if  $\hat{\mathbf{x}}$  satisfies  $A\hat{\mathbf{x}} = \hat{\mathbf{b}}$ .

(Note: we know that  $A\hat{\mathbf{x}} = \hat{\mathbf{b}}$  is consistent (by definition), so there exists at least one solution. However note that if  $A$  has free variables – the columns of  $A$  are not independent – then there would be many solutions of  $A\hat{\mathbf{x}} = \hat{\mathbf{b}}$ .)

Here is another picture of what we are doing:



Now let's look at a specific case.

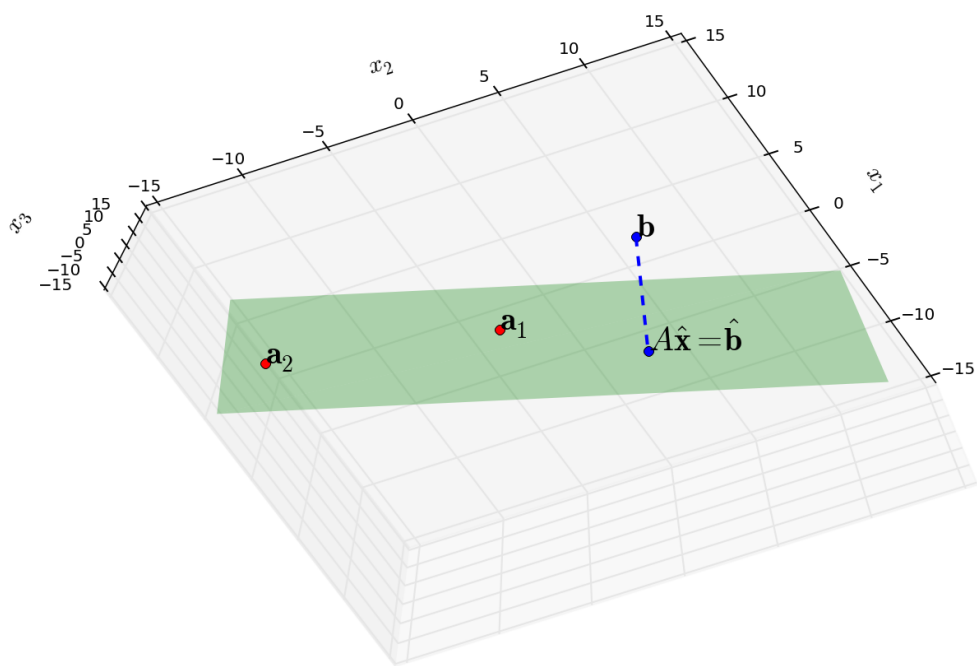
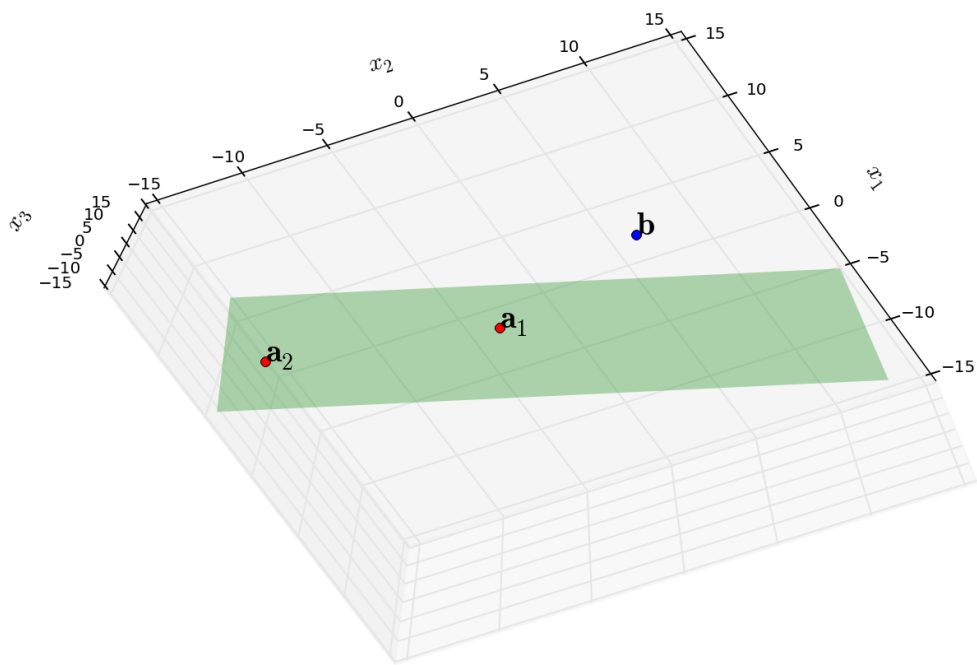
Say that

$$A = [\mathbf{a}_1 \quad \mathbf{a}_2] = \begin{bmatrix} 1 & 5 \\ -2 & -13 \\ 3 & -3 \end{bmatrix}$$

and

$$\mathbf{b} = \begin{bmatrix} 6 \\ 8 \\ -5 \end{bmatrix}.$$

We have two columns  $\mathbf{a}_1$  and  $\mathbf{a}_2$  so they do not span  $\mathbb{R}^3$ . So  $\mathbf{b}$  may not lie in  $\text{Col } A$ , and indeed it does not:



OK, how are we going to find this projection  $\hat{\mathbf{b}}$ ?

Here is the key idea: we know that the projection  $\hat{\mathbf{b}}$  has the property that  $\hat{\mathbf{b}} - \mathbf{b}$  is orthogonal to  $\text{Col } A$ . Suppose  $\hat{\mathbf{b}}$  is  $\text{proj}_{\text{Col } A} \mathbf{b}$ , and that  $\hat{\mathbf{x}}$  satisfies  $A\hat{\mathbf{x}} = \hat{\mathbf{b}}$ .

So  $A\hat{\mathbf{x}} - \mathbf{b}$  is orthogonal to each column of  $A$ .

If  $\mathbf{a}_j$  is any column of  $A$ , then

$$\mathbf{a}_j^T (A\hat{\mathbf{x}} - \mathbf{b}) = 0.$$

Now, each  $\mathbf{a}_j^T$  is a row of  $A^T$ . We can collect all of the equations for all the  $\mathbf{a}_j$  as:

$$A^T (A\hat{\mathbf{x}} - \mathbf{b}) = \mathbf{0}.$$

So

$$A^T A\hat{\mathbf{x}} - A^T \mathbf{b} = \mathbf{0}$$

So

$$A^T A\hat{\mathbf{x}} = A^T \mathbf{b}$$

Looking at this, we see that  $A^T \mathbf{b}$  is a vector, and  $A^T A$  is a matrix, so this is a standard linear system.

This linear system is called the **normal equations** for  $A\mathbf{x} = \mathbf{b}$ .

It's solution is usually denoted  $\hat{\mathbf{x}}$ .

**Theorem** The set of least-squares solutions of  $A\mathbf{x} = \mathbf{b}$  is equal to the (nonempty) set of solutions of the normal equations  $A^T A\mathbf{x} = A^T \mathbf{b}$ .

**Proof.**

- (1) The set of solutions is nonempty. The matrix on the left has the same column space as  $A^T$  and the vector on the right is a vector in  $A^T$ .

And, by the arguments above, any least-squares solution of  $A\mathbf{x} = \mathbf{b}$  must satisfy the normal equations  $A^T A\mathbf{x} = A^T \mathbf{b}$ .

- (2) Now let's show that any solution of  $A^T A\mathbf{x} = A^T \mathbf{b}$  is a least squares solution of  $A\mathbf{x} = \mathbf{b}$ .

If  $\hat{\mathbf{x}}$  satisfies  $A^T A\hat{\mathbf{x}} = A^T \mathbf{b}$ , then  $A^T (A\hat{\mathbf{x}} - \mathbf{b}) = \mathbf{0}$ ,

which shows that  $A\hat{\mathbf{x}} - \mathbf{b}$  is orthogonal to the rows of  $A^T$ , and so is orthogonal to the columns of  $A$ .

So the vector  $A\hat{\mathbf{x}} - \mathbf{b}$  is orthogonal to  $\text{Col } A$ .

So the equation

$$\mathbf{b} = A\hat{\mathbf{x}} - (\mathbf{b} - A\hat{\mathbf{x}})$$

is a decomposition of  $\mathbf{b}$  into the sum of a vector in  $\text{Col } A$  and a vector orthogonal to  $\text{Col } A$ .

Since the orthogonal decomposition is unique,  $A\hat{\mathbf{x}}$  must be the orthogonal projection of  $\mathbf{b}$  onto the column space of  $A$ .

So  $A\hat{\mathbf{x}} = \hat{\mathbf{b}}$  and  $\hat{\mathbf{x}}$  is a least-squares solution.

**Example.** Find the least squares solution of the inconsistent system  $A\mathbf{x} = \mathbf{b}$  for

$$A = \begin{bmatrix} 4 & 0 \\ 0 & 2 \\ 1 & 1 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 2 \\ 0 \\ 11 \end{bmatrix}.$$

**Solution.**

We will use the normal equations  $A^T A\mathbf{x} = A^T \mathbf{b}$ .

$$A^T A = \begin{bmatrix} 4 & 0 & 1 \\ 0 & 2 & 1 \end{bmatrix} \begin{bmatrix} 4 & 0 \\ 0 & 2 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 17 & 1 \\ 1 & 5 \end{bmatrix}$$

$$A^T \mathbf{b} = \begin{bmatrix} 4 & 0 & 1 \\ 0 & 2 & 1 \end{bmatrix} \begin{bmatrix} 2 \\ 0 \\ 11 \end{bmatrix} = \begin{bmatrix} 19 \\ 11 \end{bmatrix}$$

So the normal equations are:

$$\begin{bmatrix} 17 & 1 \\ 1 & 5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 19 \\ 11 \end{bmatrix}$$

We can solve this using row operations, or by inverting  $A^T A$ .

$$(A^T A)^{-1} = \frac{1}{84} \begin{bmatrix} 5 & -1 \\ -1 & 17 \end{bmatrix}$$

And we can then solve  $A^T A \mathbf{x} = A^T \mathbf{b}$  as

$$\begin{aligned} \hat{\mathbf{x}} &= (A^T A)^{-1} A^T \mathbf{b} \\ &= \frac{1}{84} \begin{bmatrix} 5 & -1 \\ -1 & 17 \end{bmatrix} \begin{bmatrix} 19 \\ 11 \end{bmatrix} = \frac{1}{84} \begin{bmatrix} 84 \\ 168 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}. \end{aligned}$$

So we conclude that  $\hat{\mathbf{x}} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$  is the vector that minimizes  $\|A\mathbf{x} - \mathbf{b}\|$ .

**When there are multiple solutions.**

We have seen that the normal equations always have a solution. Is there always a unique solution?

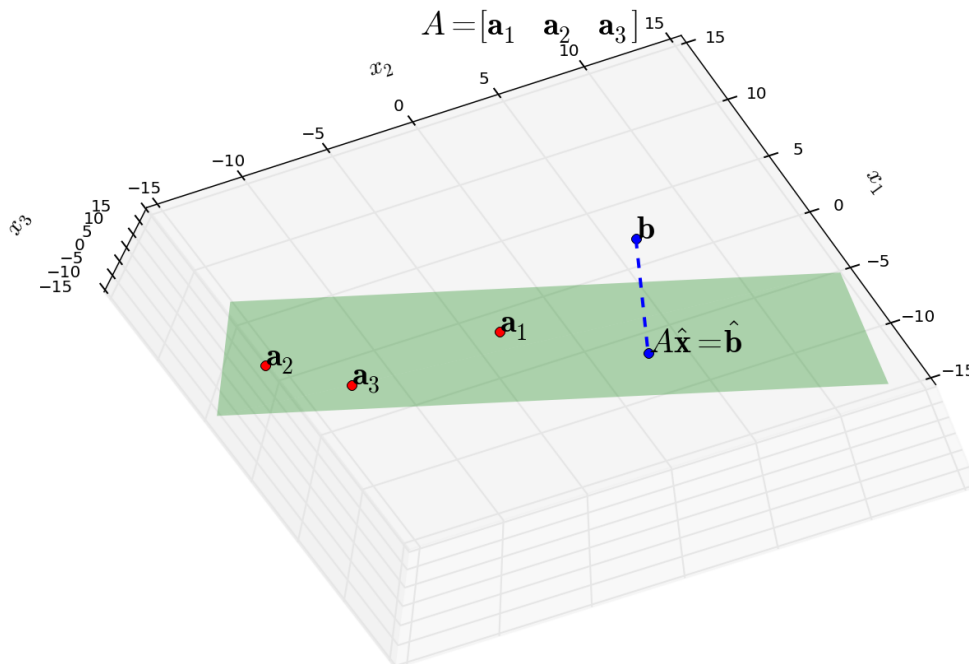
No, there can be multiple solutions that **all** minimize  $\|A\mathbf{x} - \mathbf{b}\|$ .

Let's remind ourselves of what is going on when a linear system has multiple solutions.

We know that a linear system has multiple solutions when there are columns that are not pivot columns.

Equivalently, when there are multiple solutions the columns of  $A$  are linearly dependent.

Here is a picture of what is going on:





**Example.**

Find a least-squares solution for  $A\mathbf{x} = \mathbf{b}$  for

$$A = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} -3 \\ -1 \\ 0 \\ 2 \\ 5 \\ 1 \end{bmatrix}.$$

**Solution.** Compute

$$A^T A = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 6 & 2 & 2 & 2 \\ 2 & 2 & 0 & 0 \\ 2 & 0 & 2 & 0 \\ 2 & 0 & 0 & 2 \end{bmatrix}$$

$$A^T \mathbf{b} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} -3 \\ -1 \\ 0 \\ 2 \\ 5 \\ 1 \end{bmatrix} = \begin{bmatrix} 4 \\ -4 \\ 2 \\ 6 \end{bmatrix}$$

To solve  $A^T A \mathbf{x} = A^T \mathbf{b}$ , we'll use row reduction. The augmented matrix  $[A^T A \ A^T \mathbf{b}]$  is:

$$\left[ \begin{array}{cccc|c} 6 & 2 & 2 & 2 & 4 \\ 2 & 2 & 0 & 0 & -4 \\ 2 & 0 & 2 & 0 & 2 \\ 2 & 0 & 0 & 2 & 6 \end{array} \right] \sim \left[ \begin{array}{cccc|c} 1 & 0 & 0 & 1 & 3 \\ 0 & 1 & 0 & -1 & -5 \\ 0 & 0 & 1 & -1 & -2 \\ 0 & 0 & 0 & 0 & 0 \end{array} \right]$$

Since there is a row of zeros, we know the columns of  $A^T A$  are linearly dependent. This occurs because the columns of  $A$  are linearly dependent.

The general solution is  $x_1 = 3 - x_4$ ,  $x_2 = -5 + x_4$ ,  $x_3 = -2 + x_4$ , and  $x_4$  is free.

So the general least-squares solution of  $A\mathbf{x} = \mathbf{b}$  has the form

$$\hat{\mathbf{x}} = \begin{bmatrix} 3 \\ -5 \\ -2 \\ 0 \end{bmatrix} + x_4 \begin{bmatrix} -1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

Keep in mind that the orthogonal projection  $\hat{\mathbf{b}}$  is always unique.

The reason that there are multiple solutions to this least squares problem is that there are **multiple ways** to construct  $\hat{\mathbf{b}}$ .

The reason that there are multiple ways to construct  $\hat{\mathbf{b}}$  is that the columns of  $A$  are linearly dependent, so **any** vector in the column space of  $A$  can be constructed in multiple ways.

Here is a theorem that allows use to identify when there are multiple least-squares solutions.

**Theorem.** Let  $A$  be an  $m \times n$  matrix. The following statements are equivalent:

1. The equation  $A\mathbf{x} = \mathbf{b}$  has a unique least-squares solution for each  $\mathbf{b}$  in  $\mathbb{R}^m$ .
2. The columns of  $A$  are linearly independent.
3. The matrix  $A^T A$  is invertible.

When these statements are true, the least-squares solution  $\hat{\mathbf{x}}$  is given by:

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

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

**Finding  $\hat{\mathbf{b}}$  directly.**

When  $A^T A$  is invertible, and  $\hat{\mathbf{b}}$  is unique, we can put together the two equations

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

and

$$A\hat{\mathbf{x}} = \hat{\mathbf{b}}$$

to get:

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

Let's stop and look at this another way. Up until now we have seen how to project a point onto a line, or on to a subspace with an orthogonal basis.

Now here is a formula for projection onto a subspace space given an **arbitrary** basis. This is a general formula that can be very useful.