Least squares

Example 43. Not all linear systems have solutions.

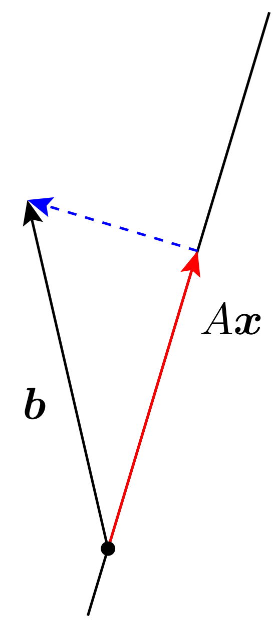
In fact, for many applications, data needs to be fitted and there is no hope for a perfect match.

For instance, $A\boldsymbol{x} = \boldsymbol{b}$ with

$$\begin{bmatrix} 1 & 2 \\ 3 & 1 \\ 0 & 5 \end{bmatrix} \boldsymbol{x} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

has no solution:

- $\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$ is not in $\operatorname{col}(A)$ since $\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \cdot \begin{bmatrix} -3 \\ 1 \\ 1 \end{bmatrix} \neq 0$ (see previous example).
- Instead of giving up, we want the $\boldsymbol{x}$ which makes $A\boldsymbol{x}$ and $\boldsymbol{b}$ as close as possible.
- Such $\boldsymbol{x}$ is characterized by the error $A\boldsymbol{x} - \boldsymbol{b}$ being **orthogonal** to $\operatorname{col}(A)$ (i.e. all possible $A\boldsymbol{x}$).



Definition 44. $\hat{\boldsymbol{x}}$ is a **least squares solution** of the system $A\boldsymbol{x} = \boldsymbol{b}$ if $\hat{\boldsymbol{x}}$ is such that $A\hat{\boldsymbol{x}} - \boldsymbol{b}$ is as small as possible (i.e. minimal norm).

- If $A\boldsymbol{x} = \boldsymbol{b}$ is consistent, then $\hat{\boldsymbol{x}}$ is just an ordinary solution. (in that case, $A\hat{\boldsymbol{x}} - \boldsymbol{b} = \boldsymbol{0}$)
- Interesting case: $A\boldsymbol{x} = \boldsymbol{b}$ is inconsistent. (in particular, if the system is overdetermined)

The normal equations

The following result provides a straightforward recipe (thanks to the FTLA) to find least squares solutions for all systems $A\boldsymbol{x} = \boldsymbol{b}$.

Theorem 45. $\hat{\boldsymbol{x}}$ is a least squares solution of $A\boldsymbol{x} = \boldsymbol{b}$
$\iff A^T A\hat{\boldsymbol{x}} = A^T\boldsymbol{b}$ (the **normal equations**)

Proof.

$\hat{\boldsymbol{x}}$ is a least squares solution of $A\boldsymbol{x} = \boldsymbol{b}$
$\iff A\hat{\boldsymbol{x}} - \boldsymbol{b}$ is as small as possible
$\iff A\hat{\boldsymbol{x}} - \boldsymbol{b}$ is orthogonal to $\operatorname{col}(A)$
$\overset{\text{FTLA}}{\iff} A\hat{\boldsymbol{x}} - \boldsymbol{b}$ is in $\operatorname{null}(A^T)$
$\iff A^T(A\hat{\boldsymbol{x}} - \boldsymbol{b}) = \boldsymbol{0}$
$\iff A^T A\hat{\boldsymbol{x}} = A^T\boldsymbol{b}$ □

Armin Straub
straub@southalabama.edu

Example 46. Find the least squares solution to $A\boldsymbol{x}=\boldsymbol{b}$, where

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

Solution. First, $A^TA=\begin{bmatrix} 1 & -1 & 0\\ 1 & 1 & 0 \end{bmatrix}\begin{bmatrix} 1 & 1\\ -1 & 1\\ 0 & 0 \end{bmatrix}=\begin{bmatrix} 2 & 0\\ 0 & 2 \end{bmatrix}$ and $A^T\boldsymbol{b}=\begin{bmatrix} 1 & -1 & 0\\ 1 & 1 & 0 \end{bmatrix}\begin{bmatrix} 2\\ 1\\ 1 \end{bmatrix}=\begin{bmatrix} 1\\ 3 \end{bmatrix}$.

Hence, the normal equations $A^TA\hat{\boldsymbol{x}}=A^T\boldsymbol{b}$ take the form $\begin{bmatrix} 2 & 0\\ 0 & 2 \end{bmatrix}\hat{\boldsymbol{x}}=\begin{bmatrix} 1\\ 3 \end{bmatrix}$.
Solving, we immediately find $\hat{\boldsymbol{x}}=\begin{bmatrix} 1/2\\ 3/2 \end{bmatrix}$.

Check. Since $A\hat{\boldsymbol{x}}=\begin{bmatrix} 2\\ 1\\ 0 \end{bmatrix}$, the error is $A\hat{\boldsymbol{x}}-\boldsymbol{b}=\begin{bmatrix} 0\\ 0\\ -1 \end{bmatrix}$. Recall that the error must be orthogonal to $\operatorname{col}(A)$!

This error is indeed orthogonal to $\operatorname{col}(A)$ because $\begin{bmatrix} 0\\ 0\\ -1 \end{bmatrix}\cdot\begin{bmatrix} 1\\ -1\\ 0 \end{bmatrix}=0$ and $\begin{bmatrix} 0\\ 0\\ -1 \end{bmatrix}\cdot\begin{bmatrix} 1\\ 1\\ 0 \end{bmatrix}=0$.

Comment. Why are the normal equations so particularly simple (compare with example below for the typical case) here? Note how each entry of the product A^TA is computed as the dot product of two columns of A (matrix products of a row of A^T times a column of A). That A^TA is a diagonal matrix reflects the fact that the two columns of A are orthogonal to each other.

Example 47. Find the least squares solution to $A\boldsymbol{x}=\boldsymbol{b}$, where

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

Solution. First, $A^TA=\begin{bmatrix} 1 & 3 & 0\\ 2 & 1 & 5 \end{bmatrix}\begin{bmatrix} 1 & 2\\ 3 & 1\\ 0 & 5 \end{bmatrix}=\begin{bmatrix} 10 & 5\\ 5 & 30 \end{bmatrix}$ and $A^T\boldsymbol{b}=\begin{bmatrix} 1 & 3 & 0\\ 2 & 1 & 5 \end{bmatrix}\begin{bmatrix} 1\\ 1\\ 1 \end{bmatrix}=\begin{bmatrix} 4\\ 8 \end{bmatrix}$.

Hence, the normal equations $A^TA\hat{\boldsymbol{x}}=A^T\boldsymbol{b}$ take the form $\begin{bmatrix} 10 & 5\\ 5 & 30 \end{bmatrix}\hat{\boldsymbol{x}}=\begin{bmatrix} 4\\ 8 \end{bmatrix}$.

Since $\begin{bmatrix} 10 & 5\\ 5 & 30 \end{bmatrix}^{-1}=\frac{1}{275}\begin{bmatrix} 30 & -5\\ -5 & 10 \end{bmatrix}=\frac{1}{55}\begin{bmatrix} 6 & -1\\ -1 & 2 \end{bmatrix}$, we find $\hat{\boldsymbol{x}}=\frac{1}{55}\begin{bmatrix} 6 & -1\\ -1 & 2 \end{bmatrix}\begin{bmatrix} 4\\ 8 \end{bmatrix}=\frac{1}{55}\begin{bmatrix} 16\\ 12 \end{bmatrix}$.

Check. Since $A\hat{\boldsymbol{x}}=\frac{1}{55}\begin{bmatrix} 40\\ 60\\ 60 \end{bmatrix}$, the error $A\hat{\boldsymbol{x}}-\boldsymbol{b}=\frac{1}{55}\begin{bmatrix} -15\\ 5\\ 5 \end{bmatrix}=\frac{1}{11}\begin{bmatrix} -3\\ 1\\ 1 \end{bmatrix}$ must be orthogonal to $\operatorname{col}(A)$.

The error is indeed orthogonal to $\operatorname{col}(A)$ because $\begin{bmatrix} 1\\ 3\\ 0 \end{bmatrix}\cdot\frac{1}{11}\begin{bmatrix} -3\\ 1\\ 1 \end{bmatrix}=0$ and $\begin{bmatrix} 2\\ 1\\ 5 \end{bmatrix}\cdot\frac{1}{11}\begin{bmatrix} -3\\ 1\\ 1 \end{bmatrix}=0$.

Any serious linear algebra problems are done by a machine. Let us see how to use the open-source computer algebra system **Sage** to do basic computations for us.

Sage is freely available at sagemath.org. Instead of installing it locally (it's huge!) we can conveniently use it in the cloud at cocalc.com from any browser. For short computations, like the one below, you can also just use the input field on our course website.

Sage is built as a **Python** library, so any Python code is valid. Here, we will just use it as a fancy calculator.

Let's revisit Example 38 and let Sage do the work for us:

```
>>> A = matrix([[1,2,1,4],[2,4,0,2],[3,6,0,3]])
```

```
>>> A.rref()
```

$$\begin{pmatrix} 1 & 2 & 0 & 1 \\ 0 & 0 & 1 & 3 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

Similarly, if we wanted to compute a basis for $\operatorname{null}(A^T)$, we can simply do:

```
>>> A.transpose().rref()
```

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & \frac{3}{2} \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

Here are some other standard things we might be interested in (compare with Example 17):

```
>>> A = matrix([[4,0,2],[2,2,2],[1,0,3]])
```

```
>>> A.eigenvalues()
```

$$[5, 2, 2]$$

```
>>> A.eigenvectors_right()
```

$$\left[\left(5, \left[\left(1,\ 1,\ \frac{1}{2}\right)\right], 1\right), (2, [(1,\ 0,\ -1), (0,\ 1,\ 0)], 2)\right]$$

```
>>> A.eigenmatrix_right()
```

$$\left(\begin{bmatrix} 5 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{bmatrix}, \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ \frac{1}{2} & -1 & 0 \end{bmatrix}\right)$$

```
>>> A.rank()
```

3

```
>>> A.determinant()
```

$$20$$

```
>>> A.inverse()
```

$$\begin{pmatrix} \frac{3}{10} & 0 & -\frac{1}{5} \\ -\frac{1}{5} & \frac{1}{2} & -\frac{1}{5} \\ -\frac{1}{10} & 0 & \frac{2}{5} \end{pmatrix}$$

Armin Straub
straub@southalabama.edu

Application: least squares lines

Given data points (x_i, y_i), we wish to find optimal parameters a, b such that $y_i \approx a + bx_i$ for all i.

Example 48. Determine the line that "best fits" the data points $(2,1), (5,2), (7,3), (8,3)$.

Comment. Can you see that there is no line fitting the data perfectly? (Check out the last two points!)

Solution. We need to determine the values a, b for the best-fitting line $y = a + bx$.
If there was a line that fit the data perfectly, then:

$$\begin{aligned} a + 2b &= 1 \qquad (2,1)\\ a + 5b &= 2 \qquad (5,2)\\ a + 7b &= 3 \qquad (7,3)\\ a + 8b &= 3 \qquad (8,3) \end{aligned}$$

In matrix form, this is: $\underbrace{\begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ 1 & x_4 \end{bmatrix}}_{\text{design matrix } X} \begin{bmatrix} a \\ b \end{bmatrix} = \underbrace{\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix}}_{\text{observation vector } \boldsymbol{y}}$ (writing the points as (x_i, y_i))

Using our points, these equations become $\begin{bmatrix} 1 & 2 \\ 1 & 5 \\ 1 & 7 \\ 1 & 8 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 3 \end{bmatrix}$. [This system is inconsistent (as expected).]

We compute a least squares solution.

$$X^T X = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 2 & 5 & 7 & 8 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 1 & 5 \\ 1 & 7 \\ 1 & 8 \end{bmatrix} = \begin{bmatrix} 4 & 22 \\ 22 & 142 \end{bmatrix}, \qquad X^T \boldsymbol{y} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 2 & 5 & 7 & 8 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \\ 3 \\ 3 \end{bmatrix} = \begin{bmatrix} 9 \\ 57 \end{bmatrix}.$$

Solving the normal equations $\begin{bmatrix} 4 & 22 \\ 22 & 142 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} 9 \\ 57 \end{bmatrix}$, we find $\begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} 2/7 \\ 5/14 \end{bmatrix}$.

Hence, the least squares line is $y = \frac{2}{7} + \frac{5}{14}x$.

The plot above shows our points together with this line. It does look like a very good fit!

Important comment. In what sense is this the line of "best fit"? By computing a least squares solution the way we do, we are minimizing the error $\boldsymbol{y} - X\begin{bmatrix} a \\ b \end{bmatrix}$. The components of that error are $y_i - (a + bx_i)$.

Hence, we see that we are minimizing the **residual sum of squares** $\mathrm{SS}_{\mathrm{res}} = \sum_i [y_i - (a + bx_i)]^2$.

Also see the discussion after the next example (where we swap the role of x and y) as well as the example at the beginning of next class (where we discuss making predictions and why minimizing $\mathrm{SS}_{\mathrm{res}}$ corresponds to minimizing the error of those predictions).

Armin Straub
straub@southalabama.edu