We left off talking about least–squares, specifically the solution

$$A^T A\vec{x} = A^T\vec{b}$$

but what if $A^T A^{-1}$ is not invertible?

$$A = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix} \qquad \vec{b} = \begin{bmatrix} -3 \\ -1 \\ 0 \\ 2 \\ 5 \\ 1 \end{bmatrix}$$

$$A^T A = \begin{bmatrix} 6 & 2 & 2 & 2 \\ 2 & 2 & 0 & 0 \\ 2 & 0 & 2 & 0 \\ 2 & 0 & 0 & 2 \end{bmatrix} \qquad A^T\vec{b} = \begin{bmatrix} 4 \\ -4 \\ 2 \\ 6 \end{bmatrix}$$

$A^T A$ is not invertible here, so we'll do Gaussian elimination.

$$\begin{bmatrix} 6 & 2 & 2 & 2 & 4 \\ 2 & 2 & 0 & 0 & -4 \\ 2 & 0 & 2 & 0 & 2 \\ 2 & 0 & 0 & 2 & 6 \end{bmatrix} \sim \begin{bmatrix} 1 & 0 & 0 & 1 & 3 \\ 0 & 1 & 0 & -2 & -5 \\ 0 & 0 & 1 & -1 & -2 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$\begin{aligned} x_1 &= 3 - x_4 \\ x_2 &= -5 + x_4 \\ x_3 &= -2 + x_4 \\ x_4 &\text{ is free} \end{aligned}$$

Therefore,

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

Theorem 1 *Let $A \in \mathbb{R}^{m\times n}$. The following statements are logically equivalent*

1. *$A\vec{x} = \vec{b}$ is a unique least–squares solution for all $\vec{b} \in \mathbb{R}^m$*
2. *The columns of A are linearly independent*
3. *The matrix $A^T A$ is invertible*

(In the example above, because $A^T A$ is not invertible (3), we don't get a unique solution (1). Instead, there are many solutions.)

How do we measure how "good" a least–squares solution is? The least–squares error

$$||\vec{b} - A\vec{x}||$$

Does orthogonality change how we compute a least–squares solution?

Let $A = \begin{bmatrix} 1 & -6 \\ 1 & -2 \\ 1 & 1 \\ 1 & 7 \end{bmatrix}$, $\vec{b} = \begin{bmatrix} -1 \\ 2 \\ 1 \\ 6 \end{bmatrix}$

$$\hat{b} = \frac{8}{4}\vec{a_1} + \frac{45}{90}\vec{a_2}$$

$$= \begin{bmatrix} 2 \\ 2 \\ 2 \\ 2 \end{bmatrix} + \begin{bmatrix} -3 \\ -1 \\ 1/2 \\ 7/2 \end{bmatrix} = \begin{bmatrix} -1 \\ 1 \\ 5/2 \\ 11/2 \end{bmatrix}$$

Because the basis vectors of A are orthogonal, we already had our solution when we calculated the projection.

$$\hat{x} = \begin{bmatrix} 8/4 \\ 45/90 \end{bmatrix} = \begin{bmatrix} 2 \\ 1/2 \end{bmatrix}$$

Some systems are called *ill–conditioned*, which means that a very small change in input can lead to a very large change in output. This can be problematic when using least–squares. The normal equations, depending on the input, can be ill–conditioned.

Theorem 2 *Given $A \in \mathbb{R}^{m\times n}$ with linearly independent columns, let $A = QR$ be a QR factorization of A. Then, for each $\vec{b} \in \mathbb{R}^m$, $A\vec{x} = \vec{b}$ has a unique least–squares solution*

$$\hat{x} = R^{-1}Q^T\vec{b}$$

Example

Let $A = \begin{bmatrix} 1 & 3 & 5 \\ 1 & 1 & 0 \\ 1 & 1 & 2 \\ 1 & 3 & 5 \end{bmatrix}$, $\vec{b} = \begin{bmatrix} 3 \\ 5 \\ 7 \\ -3 \end{bmatrix}$

1. Compute the QR factorization using Gram–Schmidt

$$A = \begin{bmatrix} 1/2 & 1/2 & 1/2 \\ 1/2 & -1/2 & -1/2 \\ 1/2 & -1/2 & 1/2 \\ 1/2 & 1/2 & -1/2 \end{bmatrix} \begin{bmatrix} 2 & 4 & 5 \\ 0 & 2 & 3 \\ 0 & 0 & 2 \end{bmatrix}$$

$$Q^T\vec{b} = \begin{bmatrix} 6 \\ -6 \\ 4 \end{bmatrix}$$

Now we can solve

$$R\hat{x} = Q^T\vec{b}$$

$$\begin{bmatrix} 2 & 4 & 5 \\ 0 & 2 & 3 \\ 0 & 0 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 6 \\ -6 \\ 4 \end{bmatrix}$$

$$\hat{x} = \begin{bmatrix} 10 \\ -6 \\ 2 \end{bmatrix}$$

1 Linear Models

The Simplest Model

Take $A\vec{x} = \vec{b}$. In older textbooks, we would see $x\vec{\beta} = \vec{y}$

x — the design matrix

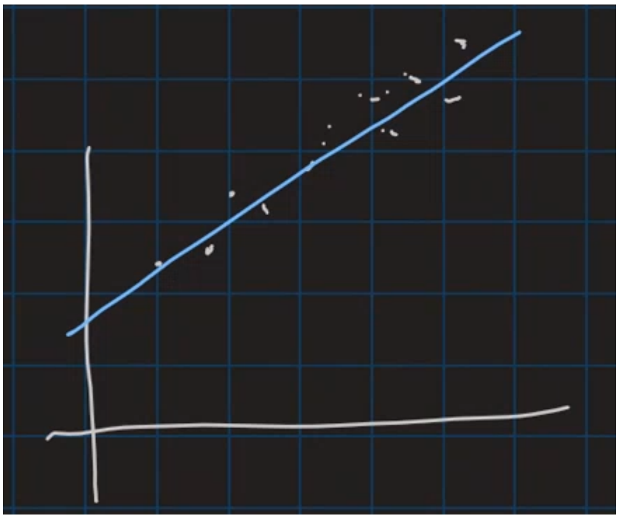
$\vec{\beta}$ — parameter vector

$\vec{y}$ — observation vector

The simplest version of this is

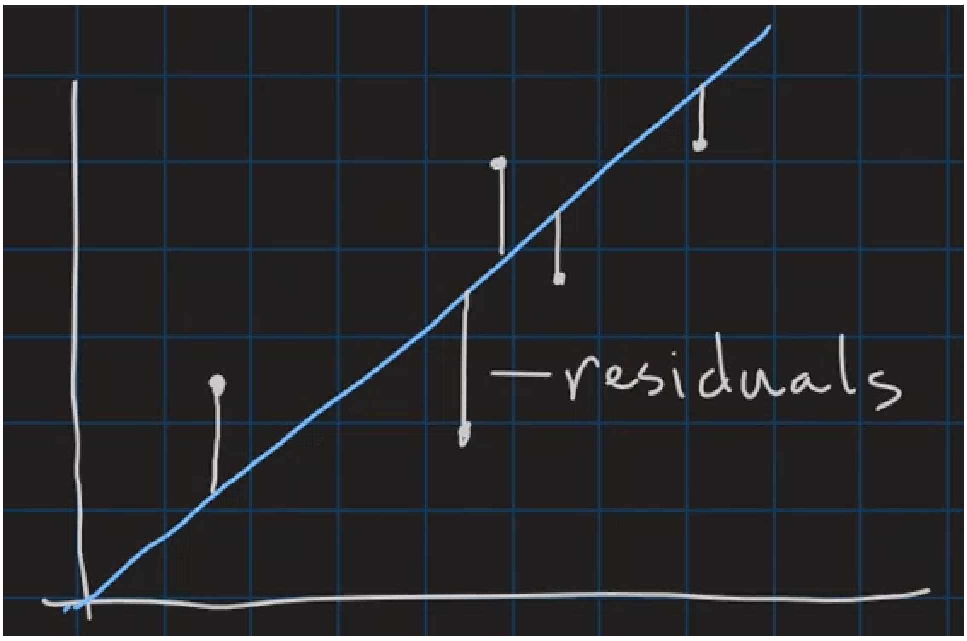
$$y = \beta_0 + \beta_1 x$$

Notice this is the equation for a line. The β_0 is the “bias” which is important in approximating data with a line. If that was not present, all approximations would have to pass through the origin. Here’s a visual example of a line approximating data:



An exact solution is not feasible. The best we can do is approximate. This method is called linear regression, because the true points are regressing towards the linear approximation.

What we can do is plot a line through our data and take the projection of the points into the line. The distance of the projection should be minimized. This will tie in with the least–squares solution. The distance between the point and the line is called a “residual”



We want to optimize β by computing the least–squares line. This is the line that minimizes the sum of squares of the residuals.

The parameters β_i are called regression coefficients.

Going back to the equation above...

$$x\vec{\beta} = \vec{y}$$

$$x = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \qquad \vec{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \qquad \vec{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$$

The square of the distance between $x\vec{\beta}$ and $\vec{y}$ is the sum of squares of the residuals.

Example

Find the equation $\vec{y} = \beta_0 + \beta_1 x$ of the least–squares line.

Data: (2, 1), (5, 2), (7, 3), (8, 3)

1. Build the design matrix

$$x = \begin{bmatrix} 1 & 2 \\ 1 & 5 \\ 1 & 7 \\ 1 & 8 \end{bmatrix} \qquad \vec{y} = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 3 \end{bmatrix}$$

2. Solve the normal equations

$$x^T x\vec{\beta} = x^T\vec{y}$$

$$x^T x = \begin{bmatrix} 4 & 22 \\ 22 & 142 \end{bmatrix} \qquad x^T\vec{y} = \begin{bmatrix} 9 \\ 57 \end{bmatrix}$$

$$\begin{bmatrix} 4 & 22 \\ 22 & 142 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} 9 \\ 57 \end{bmatrix}$$

$$\begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} 2/7 \\ 5/14 \end{bmatrix}$$

So the least–squares line is

$$y = \frac{2}{7} + \frac{5}{14}x$$