**The Least Squares Method of Linear Regression**

A common problem that arises in the analysis of statistical data is to find the best line to fit a bunch of data points. That is, given a bunch of ordered pairs \((x_1, y_1), \ldots, (x_n, y_n)\), we want to find the line \(L\), given by an equation \(y = cx + d\), that will come closest to fitting the data.

We are hoping that \(y\) really does depend on \(x\) in an essentially linear fashion \(y = cx + d\), although we don’t (yet) know what those numbers \(c\) and \(d\) should be. But in practice, factors like experimental error mean that our observed data points \((x_1, y_1), \ldots, (x_n, y_n)\) probably don’t quite fit on a line. So we won’t be able to find a line that fits the data perfectly; instead, we want the line that fits it best.

What exactly does “best” mean here? Well, it depends what you want to do, but one of the standard meanings is that we want to find numbers \(c\) and \(d\) so that the function \(f(x) = cx + d\) **minimizes the total squared error**. To explain that phrase:

- The **error** at the \(i\)-th data point \((x_i, y_i)\) is \(|f(x_i) - y_i|\), i.e., \(|cx_i + d - y_i|\)
- The **squared error** at the \(i\)-th data point \((x_i, y_i)\) is \(|f(x_i) - y_i|^2\), i.e., \((cx_i + d - y_i)^2\)
- The **total squared error** is the sum of all the squared errors, i.e.

\[
E = \sum_{i=1}^{n} (cx_i + d - y_i)^2. \tag{1}
\]

Remember, in equation (1), we have been given all the numbers \(x_1, \ldots, x_n\) and \(y_1, \ldots, y_n\), and we are trying to find numbers \(c\) and \(d\) to minimize \(E\). In theory, we could do this using multivariable calculus — think of equation (1) as \(E(c, d)\), set the partial derivatives \(E_c\) and \(E_d\) to 0 and try to solve — but linear algebra provides a better way, via projections. Here goes:

Note that equation (1) can be rewritten in vector and inner product notation as

\[
E = \| (cx_1 + d - y_1, cx_2 + d - y_2, \ldots, cx_n + d - y_n) \|^2 = \| A\vec{v} - \vec{y} \|^2,
\]

where

\[
A = \begin{bmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{bmatrix}, \quad \vec{v} = \begin{bmatrix} c \\ d \end{bmatrix}, \quad \text{and} \quad \vec{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}. \tag{2}
\]

Thus, we want to find the vector \(\vec{v} \in \mathbb{R}^2\) that minimizes the distance (in \(\mathbb{R}^n\)) from \(A\vec{v}\) to \(\vec{y}\).

**Key Observation:** The vector \(A\vec{v} \in \mathbb{R}^n\) is in \(W = \text{Im}(A)\), the span of the (two) columns of \(A\). So to get the best \(\vec{v}\), make sure that \(A\vec{v}\) is the closest point in \(W\) to \(\vec{y} \in \mathbb{R}^n\). But think about it: [that closest point is precisely \(P_W(\vec{y})\), the orthogonal projection of \(\vec{y}\) onto \(W\)!!!]

Rather than try to compute a matrix for \(P_W\), here’s a direct way to find the specific vector \(P_W(\vec{y})\), which is all we need. Recall (Proposition 4.4.3.d) that there are unique vectors \(\vec{w} \in W\) and \(\vec{u} \in W^\perp\) such that \(\vec{y} = \vec{w} + \vec{u}\), and that this \(\vec{w}\) is precisely \(P_W(\vec{y})\). Rearranging this statement, we are saying that \(\vec{w} = P_W(\vec{y})\) is the unique element of \(W = \text{Im}(A)\) such that \(\vec{w} - \vec{y}\) belongs to \(W^\perp\). Since \(\vec{w} \in W\) can be written as \(\vec{w} = A\vec{v}\), we are saying that \(A\vec{v} - \vec{y} \in W^\perp\), or in other words, that \(A\vec{v} - \vec{y}\) is orthogonal to both columns of \(A\). (Remember that \(W = \text{Im}(A)\) is the span of the columns of \(A\).)

But that is the same as saying that \(A^t(A\vec{v} - \vec{y}) = \vec{0}\). In other words, \(\vec{v}\) is the unique element of \(\mathbb{R}^2\) such that \(A^tA\vec{v} = A^t\vec{y}\). To summarize, then, we have the following Theorem:
Theorem: The choice $\vec{v} = \begin{bmatrix} c \\ d \end{bmatrix} \in \mathbb{R}^2$ minimizing the total squared error $E$ of equation (1) is the unique solution $\vec{v}$ to the $2 \times 2$ system

$$A^t A \vec{v} = A^t \vec{y},$$

where $A$ and $\vec{y}$ are as in equation (2).

Example. Find the least squares line to fit the following data:

<table>
<thead>
<tr>
<th>$x$</th>
<th>4.1</th>
<th>8.2</th>
<th>3.6</th>
<th>5.0</th>
<th>5.9</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y$</td>
<td>6.3</td>
<td>13.0</td>
<td>6.1</td>
<td>7.2</td>
<td>8.1</td>
</tr>
</tbody>
</table>

Answer. Define $A = \begin{bmatrix} 4.1 & 1 \\ 8.2 & 1 \\ 3.6 & 1 \\ 5.0 & 1 \\ 5.9 & 1 \end{bmatrix}$ and $\vec{y} = \begin{bmatrix} 6.3 \\ 13.0 \\ 6.1 \\ 7.2 \\ 8.1 \end{bmatrix}$. Then

$$A^t A = \begin{bmatrix} 4.1 & 8.2 & 3.6 & 5.0 & 5.9 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 4.1 & 1 \\ 8.2 & 1 \\ 3.6 & 1 \\ 5.0 & 1 \\ 5.9 & 1 \end{bmatrix} = \begin{bmatrix} 156.82 & 26.8 \\ 26.8 & 5 \end{bmatrix},$$

and

$$A^t \vec{y} = \begin{bmatrix} 4.1 & 8.2 & 3.6 & 5.0 & 5.9 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 6.3 \\ 13.0 \\ 6.1 \\ 7.2 \\ 8.1 \end{bmatrix} = \begin{bmatrix} 238.18 \\ 40.7 \end{bmatrix}.$$ 

Solving $A^t A \vec{v} = A^t \vec{y}$, i.e., $\begin{bmatrix} 156.82 & 26.8 \\ 26.8 & 5 \end{bmatrix} \vec{v} = \begin{bmatrix} 238.18 \\ 40.7 \end{bmatrix}$ gives (approximately) $\vec{v} = \begin{bmatrix} 1.52 \\ -0.01 \end{bmatrix}$.

So the line best approximating the data is $y = 1.52x - 0.01$.

FYI: $y = 1.52x - 0.01$ for the five $x$ values gives:

<table>
<thead>
<tr>
<th>$x$</th>
<th>4.1</th>
<th>8.2</th>
<th>3.6</th>
<th>5.0</th>
<th>5.9</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y$</td>
<td>6.22</td>
<td>12.45</td>
<td>5.46</td>
<td>7.59</td>
<td>8.96</td>
</tr>
</tbody>
</table>

And here’s how that line looks relative to the original five points: