Lecture 33: Least Squares

Say you're in charge of redeveloping a public park, and you want to run a straight footpath as close as possible to 4 giant oaks.

That is, we want to find $\beta_0$ & $\beta_1$ coming "as close as possible" to

\[
\begin{align*}
\beta_0 + \beta_1 & \cdot (-3) = 2 \\
\beta_0 + \beta_1 & \cdot (-2) = 0 \\
\beta_0 + \beta_1 & \cdot (1) = -2 \\
\beta_0 + \beta_1 & \cdot (4) = 1
\end{align*}
\]

Or, in matrix form

\[
\begin{pmatrix}
1 & -3 \\
1 & -2 \\
1 & 1 \\
1 & 4
\end{pmatrix}
\begin{pmatrix}
\beta_0 \\
\beta_1
\end{pmatrix}
=
\begin{pmatrix}
2 \\
0 \\
-2 \\
1
\end{pmatrix}
\]

Of course, it would be more appropriate to take perpendicular distances, but for our first example let's not overcomplicate this.

As you can check, $\vec{y}$ does not belong to $\text{Col}(\vec{x})$, and so this system can't be solved. Instead, we aim to minimize the sum of squares of the vertical errors $E_i$. 
That is, we'd like to choose $\hat{\beta}$ to make
$$
\sqrt{\sum_{i=1}^{5} (y_i - \hat{y}_i)^2} = \|X\hat{\beta} - \hat{y}\| = \text{dist}(X\hat{\beta}, \hat{y})
$$
as small as possible. In fact, the vector in $\text{Col}(X)$ that does this is exactly
$$
\hat{y} = \text{Proj}_{\text{Col}(X)}(\hat{y})
$$

$$
(\text{valid if } \hat{w}_0 \perp \hat{w}_1)
$$

$$
= \frac{\hat{w}_0 \cdot \hat{y}}{\hat{w}_0 \cdot \hat{w}_0} \hat{w}_0 + \frac{\hat{w}_1 \cdot \hat{y}}{\hat{w}_1 \cdot \hat{w}_1} \hat{w}_1
$$

$$
= \frac{1}{4} \left( \begin{array}{c} \frac{39}{3} \\ \frac{31}{4} \end{array} \right) + \frac{-4}{30} \left( \begin{array}{c} -2 \\ \frac{1}{4} \end{array} \right) = \frac{1}{60} \left( \begin{array}{c} 39 \\ 31/4 \end{array} \right).
$$

In other words, the best-fit line will pass through $(-3, \frac{29}{40}), (-2, \frac{21}{60}), (1, \frac{3}{60}), (4, -\frac{9}{60})$. To find its equation, we need to solve
$$
X\hat{\beta} = \hat{y}
$$
for $\hat{\beta}$ — i.e., find $\beta_0$ and $\beta_1$ such that $\beta_0 \hat{w}_0 + \beta_1 \hat{w}_1 = \hat{y}$.

But looking at the calculations, we see that we've already done this, and $\beta_0 = \frac{1}{4}$ and $\beta_1 = -\frac{2}{15}$! So we use the line

$$
\hat{y} = \frac{1}{4} - \frac{2}{15} x.
$$

Now, how do things change if the trees are at $(0,2), (1,0), (2,-2),$ and $(3,1)$ instead? Then $\hat{y}$ is the same, but our matrix $X$ becomes
\[
X = \begin{pmatrix}
1 & 0 \\
1 & 1 \\
1 & 2 \\
1 & 3 \\
\end{pmatrix},
\]

and \( \vec{w}_0, \vec{w}_1 \) are no longer orthogonal. Perhaps a QR-decomposition could come in handy?

**Step 1:** Run the Gram-Schmidt algorithm (on columns of \( X \))

\[
\vec{v}_0 = \vec{w}_0 = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}
\]

\[
\vec{v}_1 = \vec{w}_1 - \frac{\vec{w}_1 \cdot \vec{v}_0}{\vec{v}_0 \cdot \vec{v}_0} \vec{v}_0 = \begin{pmatrix} 0 \\ 1/2 \\ 1/2 \\ 1/2 \end{pmatrix} - \begin{pmatrix} 1/4 \\ 1/4 \\ 1/4 \\ 1/4 \end{pmatrix} = \begin{pmatrix} -3/2 \\ -1/2 \\ -1/2 \\ -1/2 \end{pmatrix}
\]

\[
\Rightarrow \vec{v}_0 = \begin{pmatrix} 1/2 \\ 1/2 \\ 1/2 \\ 1/2 \end{pmatrix}, \quad \vec{v}_1 = \begin{pmatrix} -3/2 \\ -1/2 \\ -1/2 \\ -1/2 \end{pmatrix} \quad \text{and} \quad Q = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\
\frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} 
\end{pmatrix}
\]

As always with Gram-Schmidt, \( \text{Col}(Q) = \text{Col}(X) \).

**Step 2:** Find the decomposition \( X = QR \)

Since \( Q \) has orthonormal columns,

\[
Q^T Q = I_2.
\]

From \( QR = X \) we get

\[
R = Q^T QR = Q^T X
\]
\[ R = \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 3 \end{pmatrix} = \begin{pmatrix} 2 & 3 \\ 0 & \sqrt{5} \end{pmatrix} \]

Step 3: Use this to solve the least-squares problem.

Columns of \( Q \) provide an orthonormal basis of \( \text{Col}(X) \).

To project \( \hat{y} \) to this, we have the formula from Lec. 32:

\[ \hat{y} = \text{Proj}_{\text{Col}(X)} \hat{y} = Q Q^T \hat{y}. \]

So we must solve

\[ X \bar{\beta} = \hat{y}, \]

\[ QR \bar{\beta} = QQ^T \hat{y}, \]

\[ Q^T QR \bar{\beta} = Q^T QQ^T \hat{y}. \]

(\#) \[ R \bar{\beta} = Q^T \hat{y}. \]

Now

\[ Q^T \hat{y} = \begin{pmatrix} 1 & 2 & 3 \\ -2 & -2 & -2 \end{pmatrix} \begin{pmatrix} 2 \\ 0 \\ -2 \end{pmatrix} = \begin{pmatrix} \frac{1}{2} \\ -\frac{\sqrt{5}}{2} \end{pmatrix}. \]

So (\#) reads

\[ \begin{pmatrix} 2 & 3 \\ 0 & \sqrt{5} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} = \begin{pmatrix} \frac{1}{2} \\ -\frac{\sqrt{5}}{2} \end{pmatrix} \implies \beta_1 = -\frac{1}{2}, \quad \beta_0 = 1. \]

and the least-squares line is \( y = 1 - \frac{1}{2} x \).
Summary

We found

1. How to get QR factorization of $A$

   **Step 1:** Perform full Gram-Schmidt on columns of $A$ — i.e. normalize the vectors at the end!

   **Step 2:** Compute $R = Q^T A$ (should be upper-triangular)

2. A least-squares solution of an equation $A\hat{x} = \hat{b}$

   is $\hat{x} \in \mathbb{R}^n$ such that $\|\hat{b} - A\hat{x}\| \leq \|\hat{b} - A\hat{x}\|$ for all $\hat{x} \in \mathbb{R}^n$.

   It can be obtained by solving $A\hat{x} = \hat{b}$, where $\hat{b} = \text{Proj}_\text{col}(A)(\hat{b})$, since

   - $\text{Col}(A)$ is the set of vectors that can be written $A\hat{x}$
   - $\text{Proj}_\text{col}(A)(\hat{b})$ is the closest vector to $\hat{b}$ in $\text{Col}(A)$

   (in particular, $A\hat{x} = \hat{b}$ is consistent).

3. If the columns of $A$ are independent and you know a QR decomposition of $A$, then a least-squares solution may be found by

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

   Since then

   $$QR\hat{x} = QQ^T \hat{b}.$$
Q1: What can you say about the least-squares solution of $A\hat{x} = \hat{b}$ when $\hat{b}$ is $\perp$ to $A$’s columns?

Ans: It’s $\hat{0}$! (More precisely, $\hat{0}$ is a least-squares solution.)

Q2: When is “least-squares solution” unique, i.e. the least-squares solution?

Ans: When columns of $A$ are independent, so that $\text{Null}(A) = \{0\}$.

Finding $\hat{x}$ w/o Gram-Schmidt or QR

Set $\hat{b} := \text{Proj}_{\text{Col}(A)} \hat{b}$. Then

\[
\hat{x} \text{ is a least-squares solution to } A\hat{x} = \hat{b} \quad \iff \quad A\hat{x} = \hat{b} \\
\iff \quad A\hat{x} - \hat{b} \in (\text{Col}(A)^{\perp}) \\
\iff \quad 0 = c_j \cdot (A\hat{x} - \hat{b}) = c_j^{T}(A\hat{x} - \hat{b}) \\
\quad \text{for each } j \\
\iff \quad A^{T}(A\hat{x} - \hat{b}) = 0 \quad \iff \quad A^{T}A\hat{x} = A^{T}\hat{b} \\
\quad \text{“Normal equations”}
\]

Claim: $A^{T}A$ is invertible $\iff$ columns of $A$ are independent.

In this case, $\hat{x} = (A^{T}A)^{-1}A^{T}\hat{b}$ is the unique least-squares solution.
Ex/ Find a (or the!) least-squares solution to
\[
\begin{pmatrix}
1 & 0 \\
1 & 2 \\
1 & 3
\end{pmatrix}
\begin{pmatrix}
\hat{x}_1 \\
\hat{x}_2
\end{pmatrix}
= 
\begin{pmatrix}
2 \\
0 \\
-2
\end{pmatrix}
\quad \left[ A\hat{x} = b \right]
\]
in this way.

\[
A^T A = 
\begin{pmatrix}
1 & 0 & 1 \\
0 & 1 & 2 \\
1 & 2 & 3
\end{pmatrix}
\begin{pmatrix}
1 & 1 & 1 \\
1 & 2 & 2 \\
1 & 3 & 3
\end{pmatrix} = 
\begin{pmatrix}
4 & 6 \\
6 & 14
\end{pmatrix}
\]

\[
A^T b = 
\begin{pmatrix}
1 & 0 & 1 \\
0 & 1 & 2 \\
1 & 2 & 3
\end{pmatrix}
\begin{pmatrix}
2 \\
0 \\
-2
\end{pmatrix} = 
\begin{pmatrix}
1 \\
-1
\end{pmatrix}
\]

Normal equations: 
\[
\begin{pmatrix}
4 & 6 \\
6 & 14
\end{pmatrix}
\begin{pmatrix}
\hat{x}_1 \\
\hat{x}_2
\end{pmatrix} = 
\begin{pmatrix}
1 \\
-1
\end{pmatrix}
\]

\[
\begin{pmatrix}
\hat{x}_1 \\
\hat{x}_2
\end{pmatrix} = \frac{1}{20} \begin{pmatrix}
14 & -6 \\
-6 & 4
\end{pmatrix} \begin{pmatrix}
1 \\
-1
\end{pmatrix} = 
\begin{pmatrix}
1 \\
-1/2
\end{pmatrix}
\]

Which is the same solution as before, but from a very
different-looking 2 x 2 system!

While the absence of square-roots in the
normal equation makes it appear better suited for
computers, the opposite is true: taking the inverse
of A^T A turns out to be (numerically) a bit of
a disaster for large matrices.