orthogonalityAndOptimization

September 9, 2019

We saw that, given a model, the parameters that best explain the data can be found by minimizing the sum squares error **(SSE)**. SSE was written as a function of our parameters and minimized by taking the derivative with respect to each parameter.

Our end result was the following equation

$$\beta_{\text{optimal}} = (X'X)^{-1}X'y.$$

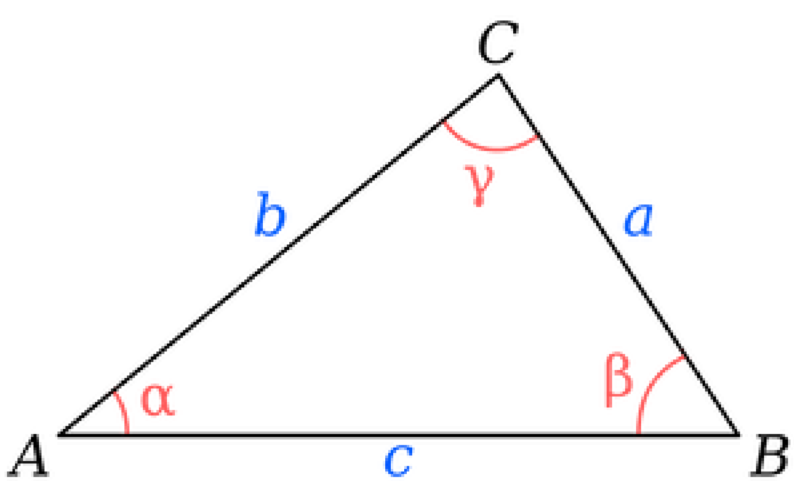
This was a calculus approach to finding optima. We can also understand optima, and arrive at this same equation, by using linear algebra.

0.1 orthogonality

Two vectors, x and y are perpendicular to one another, or **orthogonal**, if their inner product equals 0

$$x'y = 0$$

We can use the law of cosines to see why this is the case. The law of cosines says, give a triangle with edge-lengths a, b, and c, and the angle γ made by the vector CA and CB, the edge-lengths are related like



$$C^2 = A^2 + B^2 - 2AB\cos(\gamma)$$

where capital letters denote the **length** of the triangle's sides.

If we consider a and b vectors, then

$$c = a - b$$

and the length of c is

$$\begin{aligned}
c^2 = c'c = [a_1 - b_1 a_2 - b_2] \begin{bmatrix} a_1 - b_1 \\ a_2 - b_2 \end{bmatrix} \quad (1) \\
= (a_1 - b_1)^2 + (a_2 - b_2)^2 \quad (2) \\
= (a_1^2 + a_2^2) + (b_1^2 + b_2^2) - 2(a_1 b_1 + a_2 b_2) \quad (3)
\end{aligned}$$

The above can be rewritten as the inner product of three vectors

$$\begin{aligned}
c^2 = (a_1^2 + a_2^2) + (b_1^2 + b_2^2) - 2(a_1 b_1 + a_2 b_2) \quad (4) \\
= a'a + b'b - 2a'b \quad (5)
\end{aligned}$$

then we can relate this vector equation to our original cosine law.

$$a'a + b'b - 2a'b = A^2 + B^2 - 2AB\cos(\gamma)$$

We see that $a'a$ corresponds to the the length A squared and $b'b$ corresponds to the length B squared.

We define a vector's length

$$||v|| = (v'v)^{1/2},$$

and note that the length of a vector is always positive, and can only be zero if the vector has entries all zero.

The last term then relates the inner product between a and b to their lengths and the cosine of the angle they make

$$\begin{aligned}
-2a'b = -2AB\cos(\gamma) \quad (6) \\
a'b = AB\cos(\gamma) \quad (7) \\
a'b = ||a||||b||\cos(\gamma) \quad (8)
\end{aligned}$$

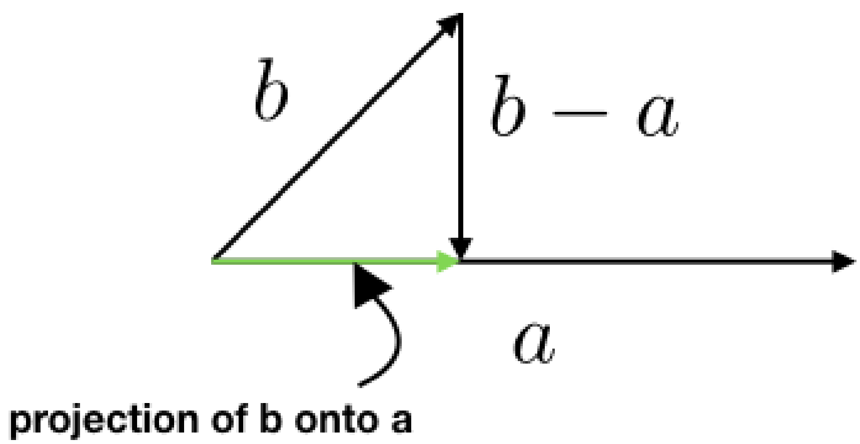
if the inner product $a'b$ is zero

$$0 = ||a||||b|| \cos(\gamma) \tag{9}$$

and assuming a and b are not zero vectors, it must be the case that $\cos(\gamma) = 0$ and this happens when $\gamma = \frac{\pi}{2}$, a perpendicular (orthogonal) angle.

0.2 projection

A vector b is a **orthogonal** projection onto a if the inner product between $b - a$ and a is 0.



We can derive a formula for this "green" vector. The goal is to find the number ω, in the same direction as a, so that $b - a$ and a are orthogonal.

$$(b - \omega a)'(\omega a) = \omega b'a - \omega^2 a'a = 0 \tag{10}$$
$$b'a - \omega a'a = 0 \tag{11}$$
$$\omega = \frac{b'a}{a'a} \tag{12}$$

This value $\omega = b'a \Big/ a'a$ is the distance along a we need to travel until a and $b - a$ are orthogonal to one another.

0.3 orthogonal projection as minimizer

What does orthogonality and minima have to do with each other?

Suppose we want to find the vector $p \in S$ such that p is closer to $y \in B$ than any other vector $v \in S$, and we assume $S \subset B$, that y and any vector v cannot lie in th same space.

The distance between y and any vector v is

$$||y - v|| = [(y - v)'(y - v)]^{1/2}, \tag{13}$$

and if a vector p is closest in distance $||.||$ than it will be closes in squared distance too $||.||^2$.

So we're searching for a vector p so that

$$||y - v||^2 = (y - v)'(y - v) \tag{14}$$

is as small as possible.

First we introduce this smallest vector p without changing the above equation

$$\begin{aligned} ||y - p + p - v||^2 &= \{[(y - p) + (p - v)]'[(y - p) + (p - v)]\} && (15) \\ &= (y - p)'(y - p) + (p - v)'(p - v) + 2(p - v)'(y - p) && (16) \\ &= ||y - p||^2 + ||p - v||^2 + 2(p - v)'(y - p) && (17) \end{aligned}$$

the first two terms here cannot be changed much, but lets look at the third term. p and v are both vectors in S and so their subtraction is a vector in S. y is in B and p is in S. if we suppose p is the vector such that the difference $y - p$ is orthogonal to **every** possible vector in S then the third term would equal 0.

The vector p is smallest if and only if the difference between y and p is orthogonal to every vector in S.

0.3.1 (aside) span

We can represent any vector in a space S through a basis. A basis is a set of independent vectors such that every vector in S is the weighted sum of basis vectors.

Suppose a is in some space V. Then a basis is a set of vectors $v_1, v_2, \cdots, v_n$ such that

$$a = \sum_{i=1}^{N} \alpha_i v_i$$

for every vector $a \in V$.

Returning back to our vector p, this vector is the one so that $y - p$ is orthogonal to every vector in S, or

$$(y - p)' \left(\sum_{i=1}^{N} \alpha_i v_i \right) = \sum_{i=1}^{N} \alpha_i (y - p)' v_i = 0$$

of $y - p$ must be orthogonal to every basis vector.

0.4 reframing our problem in linear algebra

We can use material on orthogonal projections to help us understand the optimal β.

Our **design** matrix X times β can be thought of as a basis.

$$X\beta = \begin{bmatrix} x_{1,1}\beta_1 + x_{1,2}\beta_2 + \cdots + \beta_n x_{1,n} \\ x_{2,1}\beta_1 + x_{2,2}\beta_2 + \cdots + \beta_n x_{2,n} \\ x_{3,1}\beta_1 + x_{3,2}\beta_2 + \cdots + \beta_n x_{3,n} \\ \vdots \\ x_{m,1}\beta_1 + x_{m,2}\beta_2 + \cdots + \beta_n x_{m,n} \end{bmatrix} = \beta_1 \begin{bmatrix} x_{1,1} \\ x_{2,1} \\ \vdots \\ x_{m,1} \end{bmatrix} + \beta_2 \begin{bmatrix} x_{1,2} \\ x_{2,2} \\ \vdots \\ x_{m,2} \end{bmatrix} + \cdots + \beta_n \begin{bmatrix} x_{1,n} \\ x_{2,n} \\ \vdots \\ x_{m,n} \end{bmatrix} = \sum_{i=1}^{N} \beta_i x_{;,i}$$

The y observations can also be considered a single m-dimensional vector.

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}.$$

Instead of asking for the *beta* that minimizes the **SSE**, let's instead ask for the vector that is a member of the space spanned by the columns of X and closest to the vector y. This could be an alternative expression for "good fit to the data".

This best vector's (denoted b for best) difference from y must be orthogonal to all vectors in the space, or equivalently all vectors in the basis.

$$\begin{aligned}
(y-b)'x_{;1} &= 0 && (18)\\
(y-b)'x_{;2} &= 0 && (19)\\
(y-b)'x_{;3} &= 0 && (20)\\
\vdots &= 0 && (21)\\
(y-b)'x_{;1} &= 0 && (22)
\end{aligned}$$

or

$$\begin{aligned}
y'x_{;1} - b'x_{;1} &= 0 && (23)\\
y'x_{;2} - b'x_{;2} &= 0 && (24)\\
y'x_{;3} - b'x_{;3} &= 0 && (25)\\
\vdots &= 0 && (26)\\
y'x_{;n} - b'x_{;n} &= 0 && (27)\\
& && (28)
\end{aligned}$$

rearranging terms

$$
\begin{aligned}
y' x_{;1} &= b' x_{;1} && (29)\\
y' x_{;2} &= b' x_{;2} && (30)\\
y' x_{;3} &= b' x_{;3} && (31)\\
\vdots &= 0 && (32)\\
y' x_{;n} &= b' x_{;n} && (33)\\
& && (34)
\end{aligned}
$$

We can rewrite both sides of the above equation as a matrix times a vector.

$$
X'y = X'b \tag{35}
$$

We can take this equation further by remembering b must be a member of the space created by the columns of X. That is b is a weighted sum of the columns of X, for weights (let's say) β.

$$
\begin{aligned}
b &= \sum_{i=1}^{N} \beta_i x_{;i} && (36)\\
&= \beta_1 \begin{bmatrix} x_{1,1} \\ x_{2,1} \\ \vdots \\ x_{m,1} \end{bmatrix} + \beta_2 \begin{bmatrix} x_{1,2} \\ x_{2,2} \\ \vdots \\ x_{m,2} \end{bmatrix} + \cdots + \beta_n \begin{bmatrix} x_{1,n} \\ x_{2,n} \\ \vdots \\ x_{m,n} \end{bmatrix} = X\beta && (37)
\end{aligned}
$$

and the above equation now becomes

$$
\begin{aligned}
X'y &= X'b && (38)\\
X'y &= X'X\beta && (39)
\end{aligned}
$$

This is **exactly** the same equation as before. Minimizing the sum squares of error is the same as finding the vector b, constrained to be a weighted sum of the columns of X, closest to the vector y.

$$
\beta = (X'X)^{-1} X'y \tag{40}
$$

0.5 hat matrix

Now that we know how to compute optimal weights (β) for our vector b, we see the vector closest to y is

$$b = X\beta, \tag{41}$$

but this vector is just the functional form we specified for our model, minus the error. The vector b is used to make predictions about y given data X, so that

$$\hat{y} = Xb \tag{42}$$

$$\hat{y} = X\left[(X'X)^{-1}X'y\right] \tag{43}$$

$$\hat{y} = \left[X(X'X)^{-1}X'\right]y \tag{44}$$

$$\tag{45}$$

Considered a function, the matrix

$$H = \left[X(X'X)^{-1}X'\right]$$

is called the **hat matrix** because it transforms y into the vector $\hat{y}$, it places the "hat" on y.