[MATH-22] The least squares method

Miguel-Angel Canela
Associate Professor, IESE Business School

Projection on a line

Orthogonal projection is a classic of optimization. In its simplest version, it can be stated, as an unrestricted optimization problem: given two n-vectors $\mathbf{x}$ and $\mathbf{y}$, search for the coefficient b such that $\|\mathbf{y} - b\mathbf{x}\|$ is minimum. Since the squared norm of a vector is the sum of the squares of its components, we call this a **least squares problem**. The vector $b\mathbf{x}$ is called the orthogonal projection of $\mathbf{y}$ on $\mathbf{x}$.

We immediately guess, after representing $\mathbf{x}$ and $\mathbf{y}$ as two arrows in a plane, that the minimum is attained when $\mathbf{x}$ and $\mathbf{y} - b\mathbf{x}$ are orthogonal. Then,

$$\mathbf{x} \cdot \big(\mathbf{y} - b\mathbf{x}\big) = \mathbf{x} \cdot \mathbf{y} - b\|\mathbf{x}\|^2 = 0 \quad \Longrightarrow \quad b = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\|^2}\,.$$

This can also be proved formally, using differential calculus. The objective function

$$f(b) = \big\|\mathbf{y} - b\mathbf{x}\big\|^2 = \|\mathbf{x}\|^2 b^2 - 2\big(\mathbf{x} \cdot \mathbf{y}\big)b + \|\mathbf{y}\|^2$$

is quadratic (a parabola), with a positive quadratic term. Therefore, there is a unique minimum, which is the solution of the equation

$$f'(b) = 2\big(\|\mathbf{x}\|^2 b - \mathbf{x} \cdot \mathbf{y}\big) = 0.$$

This leads to the formula given above for b. If we replace $\mathbf{x}$ by a collinear vector $\mathbf{z} = \alpha\mathbf{x}$, the coefficient for $\mathbf{z}$ is

$$\frac{\mathbf{z} \cdot \mathbf{y}}{\|\mathbf{z}\|^2} = \frac{b}{\alpha}\,.$$

So, the projection of $\mathbf{z}$ is $(b/\alpha)\alpha\mathbf{x} = b\,\mathbf{x}$. This means that the projection is the same for all vectors collinear with $\mathbf{x}$, so it should be properly called the projection of $\mathbf{y}$ *on the line* defined by $\mathbf{x}$.

Example 1. For

$$\mathbf{x} = \begin{bmatrix} 2 \\ 1 \\ 2 \end{bmatrix}, \qquad \mathbf{y} = \begin{bmatrix} 1 \\ 2 \\ -1 \end{bmatrix},$$

the orthogonal projection of $\mathbf{y}$ on $\mathbf{x}$ is $b\mathbf{x}$, with

$$b = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\|^2} = \frac{2}{9}\,.$$

Projection on a multidimensional subspace

The definition of the orthogonal projection of a vector $\mathbf{y}$ on a vector $\mathbf{x}$ (that is, on the line generated by $\mathbf{x}$) is easily extended to the projection on a linear subspace. If $\mathbf{x}_1$, …, $\mathbf{x}_k$, $\mathbf{y}$ are n-vectors, the orthogonal projection of $\mathbf{y}$ on $\mathbf{x}_1$, …, $\mathbf{x}_k$ (on the subspace generated by these vectors) is the linear combination $b_1\mathbf{x}_1 + \cdots + b_k\mathbf{x}_k$ for which the norm of the vector $\mathbf{e} = \mathbf{y} - b_1\mathbf{x}_1 - \cdots - b_k\mathbf{x}_k$ is minimum. I call $\mathbf{e}$ as the **residual vector**. Since

$$\|\mathbf{e}\|^2 = \sum \big(y_i - b_1\mathbf{x}_{1i} - \cdots - b_k\mathbf{x}_{ki}\big)^2,$$

this is a least squares problem. Extending the argument used for the case $k = 1$, it can be easily seen that the projection is such that $\mathbf{e}$ is orthogonal to all the $\mathbf{x}$'s. This can be translated into a system of equations,

$$\mathbf{x}_1 \cdot \big(\mathbf{y} - b_1\mathbf{x}_1 - \cdots - b_k\mathbf{x}_k\big) = \cdots = \mathbf{x}_k \cdot \big(\mathbf{y} - b_1\mathbf{x}_1 - \cdots - b_k\mathbf{x}_k\big) = 0.$$

If the $\mathbf{x}$'s are linearly independent, there is only one way of writing the optimal linear combination, and its coefficients are determined.

The projection is given, in matrix notation, by a simple formula. We pack the $\mathbf{x}$'s as the columns of an (n, k)-matrix $\mathbf{X}$. So, the transpose $\mathbf{X}^\mathsf{T}$ is the same pack, but with the $\mathbf{x}$'s as rows. Reminder: the product of two vectors can be written as $\mathbf{x} \cdot \mathbf{y} = \mathbf{x}^\mathsf{T}\mathbf{y}$ (if $\mathbf{x}$ is a column vector, $\mathbf{x}^\mathsf{T}$ will be a row vector). Now, writing $\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$, the orthogonality condition gives

$$\mathbf{X}^\mathsf{T}\big(\mathbf{y} - \mathbf{X}\mathbf{b}\big) = \mathbf{0} \implies \mathbf{X}^\mathsf{T}\mathbf{y} = \mathbf{X}^\mathsf{T}\mathbf{X}\mathbf{b} \implies \mathbf{b} = \big(\mathbf{X}^\mathsf{T}\mathbf{X}\big)^{-1}\mathbf{X}^\mathsf{T}\mathbf{y}.$$

Example 2. Let

$$\mathbf{x}_1 = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}, \qquad \mathbf{x}_2 = \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix}, \qquad \mathbf{y} = \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix}.$$

The orthogonal projection is easy to find here. Writing $\mathbf{y} = b_1\mathbf{x}_1 + b_2\mathbf{x}_2 + \mathbf{e}$, the third component of $\mathbf{e}$ must be 1, because of the zeros in the last row of $\mathbf{x}_1$ and $\mathbf{x}_2$. Then, the first two coordinates of $\mathbf{y}$ are a linear combination of those of $\mathbf{x}_1$ and $\mathbf{x}_2$, that is

$$b_1 = 3/2, \qquad b_2 = -1/2, \qquad \mathbf{e} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}.$$

The same can be obtained with the matrix formula. Here,

$$\mathbf{X} = \begin{bmatrix} 1 & 1 \\ 1 & -1 \\ 0 & 0 \end{bmatrix},$$

hence

$$\mathbf{b} = \left(\begin{bmatrix} 1 & 1 & 0 \\ 1 & -1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & -1 \\ 0 & 0 \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 & 1 & 0 \\ 1 & -1 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix} = \begin{bmatrix} 3/2 \\ -1/2 \end{bmatrix}. \square$$

Example 3. One could believe that the coefficient b_j is "associated" to the pair $\mathbf{y}$, $\mathbf{x}_j$. This is wrong, since b_j is also dependent on the other $\mathbf{x}$ vectors. This fact may come out unexpectedly in Statistics, but linear algebra makes it plainly evident. Take

$$\mathbf{x}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \qquad \mathbf{x}_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \qquad \mathbf{y} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}.$$

The orthogonal projection of $\mathbf{y}$ on the subspace generated by $\mathbf{x}_1$ and $\mathbf{x}_2$ is equal to $\mathbf{y}$, since $\mathbf{y} = \mathbf{x}_1 + \mathbf{x}_2$. Nevertheless, projecting $\mathbf{y}$ on $\mathbf{x}_1$ gives $2\mathbf{x}_1$. Thus, the coefficient of $\mathbf{x}_1$ is different in the two projections. Nevertheless, in Example 2, the coefficient of $\mathbf{x}_1$ did not change when I removed $\mathbf{x}_2$. Why is this so? Because the two $\mathbf{x}$'s vectors were orthogonal there.

Linear regression

Let me suppose, now, a numeric data set whose columns are the vectors $\mathbf{y}$, $\mathbf{x}_1$, ..., $\mathbf{x}_k$. We want to approximate $\mathbf{y}$ by a linear expression $b_0\mathbf{1} + b_1\mathbf{x}_1 + \cdots + b_k\mathbf{x}_k$, in which $\mathbf{1}$ denotes a vector of 1's, in the least squares sense. We write this as a **linear regression equation**

$$y = b_0 + b_1x_1 + \cdots + b_kx_k + e,$$

in which e is the residual term and the b_i's are called **regression coefficients**.

The regression coefficients are calculated with the projection formula $\mathbf{b} = \left(\mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1}\mathbf{X}^\mathsf{T}\mathbf{y}$, but here the coefficient vector includes the constant term b_0 and the matrix X contains a column of 1's. In this context,the projection formula is called the **ordinary least squares** (OLS) formula. We use the term *ordinary* here to distinguish this formula from a refinement called *generalized* least squares (GLS) that will find in the Econometrics course.

If $k = 1$, we call this **simple regression** and, if $k > 1$, **multiple regression**. In the case $k = 0$, the equation only contains a constant term. The special case $k = 0$ is called **regression without regressors**. Which is the value of the constant term in that equation? It must be the coefficient of the vector $\mathbf{1}$ in the orthogonal projection of $\mathbf{y}$ on $\mathbf{1}$. So,

$$b_0 = \frac{\mathbf{y}\cdot\mathbf{1}}{\|\mathbf{1}\|^2} = \frac{\sum y_i}{n}$$

is the mean of the vector $\mathbf{y}$.

Dropping terms

Since in Statistics the analyst typically performs different analyses, dropping some of the columns, an interesting question is: what is the effect of dropping one term of a regression equation? Mathematically, the answer is easy:

- The coefficients of the other terms, in general, change (the change may or not be relevant for the analysis).
- If the term dropped is orthogonal to the other terms (in Statistics we will say uncorrelated), the coefficients of the other terms do not change. To show you why, let me consider a simple case, of a regression equation

  $$y = b_0 + b_1x_1 + b_2x_2 + e$$

  in which x_2 is orthogonal to x_1. Putting $u = b_2x_2 + e$, we get an equation

  $$y = b_0 + b_1x_1 + u$$

  in which u is orthogonal to the other two term, so it is a linear regression equation. u is now the residual term, and the coefficient b_1 is the same as before.
- Dropping one term is equivalent to setting the corresponding coefficient to zero. The other coefficients change, to get the optimal solution with this constraint, which will be suboptimal with respect to the complete equation. So, the residual sum of squares increases. What matters, is Statistics, is how relevant is the loss. We will discuss this soon.