LS as orthogonal projection

Renato Assunção
ESRI and DCC/UFMG

A prediction problem and its dataset

We want to predict apartment prices and collect a dataset

- Prices: a column-vector $\boldsymbol{Y}$ of dimension 1500. ○
- 30 features of the 1500 houses (a matrix of dimension 1500 × 30) ○
- We want to learn how to predict (new) apartment prices using a mathematical formula based on the 30 features of each apartment.

$$\boldsymbol{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_{1499} \\ y_{1500} \end{pmatrix} \qquad \begin{pmatrix} \text{area}_1 & \text{age}_1 & \text{rooms}_1 & \cdots & \text{gym}_1 \\ \text{area}_2 & \text{age}_2 & \text{rooms}_2 & \cdots & \text{gym}_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \text{area}_{1499} & \text{age}_{1499} & \text{rooms}_{1499} & \cdots & \text{gym}_{1499} \\ \text{area}_{1500} & \text{age}_{1500} & \text{rooms}_{1500} & \cdots & \text{gym}_{1500} \end{pmatrix}$$

Linear model implies a weighted sum

The linear model assumes that

$$Y = \underbrace{\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_{30}}_{\mathbb{E}(Y|\boldsymbol{X}=\boldsymbol{x})} + \underbrace{\epsilon}_{\text{random noise}}$$

This implies that the random price Y will be predicted by a weighted sum of the 30 features.

The random noise represents the effect of all other potential factors that may influence the price.

How do we use the data to learn (or estimate) these coefficients?

What we want

AIM: to learn a simple mathematical formula (a linear combination of the 30 features) that predicts well the prices we already know in the dataset, with the hope that the same formula will also predict well in future apartment cases.

So, we look for a linear combination such that:

$$
\begin{aligned}
y_1 &\approx \beta_0 + \beta_1 \, \text{área}_1 + \beta_2 \, \text{idade}_1 + \ldots + \beta_{30} \, \text{salão}_1 \\
y_2 &\approx \beta_0 + \beta_1 \, \text{área}_2 + \beta_2 \, \text{idade}_2 + \ldots + \beta_{30} \, \text{salão}_2 \\
&\vdots \\
y_{1500} &\approx \beta_0 + \beta_1 \, \text{área}_{1500} + \beta_2 \, \text{idade}_{1500} + \ldots + \beta_{30} \, \text{salão}_{1500}
\end{aligned}
$$

Equivalent matrix representation

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_{1499} \\ y_{1500} \end{pmatrix} \approx \beta_0 \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \\ 1 \end{pmatrix} + \beta_1 \begin{pmatrix} \text{área}_1 \\ \text{área}_2 \\ \vdots \\ \text{área}_{1499} \\ \text{área}_{1500} \end{pmatrix} + \beta_2 \begin{pmatrix} \text{idade}_1 \\ \text{idade}_2 \\ \vdots \\ \text{idade}_{1499} \\ \text{idade}_{1500} \end{pmatrix} + \ldots + \beta_{30} \begin{pmatrix} \text{salão}_1 \\ \text{salão}_2 \\ \vdots \\ \text{salão}_{1499} \\ \text{salão}_{1500} \end{pmatrix}$$

Each of the column vectors above is in $\mathbb{R}^{1500}$.

We want the best linear combination of the $30+1$ feature column vectors (that is, the best $\beta_0, \beta_1, \ldots, \beta_{30}$) that is the closest possible to the column vector $\boldsymbol{Y}$.

Another equivalent matrix representation

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_{1499} \\ y_{1500} \end{pmatrix} \approx \beta_0 \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \\ 1 \end{pmatrix} + \beta_1 \begin{pmatrix} \text{área}_1 \\ \text{área}_2 \\ \vdots \\ \text{área}_{1499} \\ \text{área}_{1500} \end{pmatrix} + \beta_2 \begin{pmatrix} \text{idade}_1 \\ \text{idade}_2 \\ \vdots \\ \text{idade}_{1499} \\ \text{idade}_{1500} \end{pmatrix} + \ldots + \beta_{30} \begin{pmatrix} \text{salão}_1 \\ \text{salão}_2 \\ \vdots \\ \text{salão}_{1499} \\ \text{salão}_{1500} \end{pmatrix}$$

is equivalent to

$$\underbrace{\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ \vdots \\ y_{1499} \\ y_{1500} \end{pmatrix}}_{\boldsymbol{Y},\ 1500\times 1} \approx \underbrace{\begin{pmatrix} 1 & \text{renda}_1 & \text{area}_1 & \cdots & \text{salao}_1 \\ 1 & \text{renda}_2 & \text{area}_2 & \cdots & \text{salao}_2 \\ \vdots & \vdots & \vdots & \vdots & \\ \vdots & \vdots & \vdots & \vdots & \\ 1 & \text{renda}_{1499} & \text{area}_{1499} & \cdots & \text{salao}_{1499} \\ 1 & \text{renda}_{1500} & \text{area}_{1500} & \cdots & \text{salao}_{1500} \end{pmatrix}}_{\boldsymbol{X},\ 1500\times 31} \underbrace{\begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{30} \end{pmatrix}}_{\boldsymbol{\beta},\ 31\times 1}$$

This looks similar to a linear system: $\boldsymbol{Y} \approx \boldsymbol{X}\,\boldsymbol{\beta}$

A solution

What if we write as a linear system, substituting the $\approx$ by an $=$ sign?

$$\underset{1500\times 1}{\boldsymbol{Y}} = \underbrace{\underset{1500\times 31}{\boldsymbol{X}} \; \underset{31\times 1}{\boldsymbol{\beta}}}_{1500\times 1}$$

This is an overdetermined system: it has more equations (more rows or apartments) than variables (features, the 31 columns of $\boldsymbol{X}$). Hence, it has no exact solution.

A solution

What if we write as a linear system, substituting the $\approx$ by an $=$ sign?

$$\underset{1500\times 1}{\boldsymbol{Y}} = \underbrace{\underset{1500\times 31}{\boldsymbol{X}} \; \underset{31\times 1}{\boldsymbol{\beta}}}_{1500\times 1}$$

This is an overdetermined system: it has more equations (more rows or apartments) than variables (features, the 31 columns of $\boldsymbol{X}$). Hence, it has no exact solution.

This leads us to look for the closest "solution": the vector $\boldsymbol{\beta}$ that minimizes the distance between $\boldsymbol{Y}$ and $\boldsymbol{X}\ \boldsymbol{\beta}$.

That is, we want

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \underbrace{||\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}||}_{||\boldsymbol{r}(\boldsymbol{\beta})||}$$

The least squares solution

As

$$||\boldsymbol{r}(\beta)|| = \sqrt{||\boldsymbol{r}(\beta)||^2} = \sqrt{||\boldsymbol{Y} - \boldsymbol{X}\beta||^2}$$

and the length of a vector is a non-negative number, we have

$$\hat{\beta} = \arg\min_{\boldsymbol{\beta}} \sqrt{||\boldsymbol{Y} - \boldsymbol{X}\beta||^2} = \arg\min_{\boldsymbol{\beta}} ||\boldsymbol{Y} - \boldsymbol{X}\beta||^2$$

The minimum value is different in each case, but the minimizing argument $\hat{\beta}$ is the same.

This justifies to avoid minimizing the vector length $||\boldsymbol{r}(\beta)||$ and rather minimizing its square length $||\boldsymbol{r}(\beta)||^2$, dropping the square-root function from consideration.

The least squares solution

As

$$||\boldsymbol{r}(\beta)|| = \sqrt{||\boldsymbol{r}(\beta)||^2} = \sqrt{||\boldsymbol{Y} - \boldsymbol{X}\beta||^2}$$

and the length of a vector is a non-negative number, we have

$$\hat{\beta} = \arg\min_{\boldsymbol{\beta}} \sqrt{||\boldsymbol{Y} - \boldsymbol{X}\beta||^2} = \arg\min_{\boldsymbol{\beta}} ||\boldsymbol{Y} - \boldsymbol{X}\beta||^2$$

The minimum value is different in each case, but the minimizing argument $\hat{\beta}$ is the same.

This justifies to avoid minimizing the vector length $||\boldsymbol{r}(\beta)||$ and rather minimizing its square length $||\boldsymbol{r}(\beta)||^2$, dropping the square-root function from consideration.

In ML, we use the Pythagoras theorem extensively, which works with the squares of the sides of a right-angle triangle.

Before the formal (rigorous) solution, a pit stop

A very simple mnemonic to remember the least squares solution is as follows: We want β such that $\boldsymbol{X}\, \beta \approx \boldsymbol{Y}$.

Before the formal (rigorous) solution, a pit stop

A very simple mnemonic to remember the least squares solution is as follows: We want $\boldsymbol{\beta}$ such that $\boldsymbol{X}\ \boldsymbol{\beta} \approx \boldsymbol{Y}$.

Multiply both sides of this approximation by the constant matrix $\boldsymbol{X}^\top$. Then:

$$\underbrace{\boldsymbol{X}^\top\ \boldsymbol{X}}_{31\times 31}\ \underbrace{\boldsymbol{\beta}}_{31\times 1} \approx \underbrace{\boldsymbol{X}^\top\ \boldsymbol{Y}}_{31\times 1}.$$

This is neither an overdetermined nor an underdetermined linear system, but a square system that has a unique and exact solution: multiply both sides by $(\boldsymbol{X}^\top\ \boldsymbol{X})^{-1}$ and then

$$\underbrace{(\boldsymbol{X}^\top\ \boldsymbol{X})^{-1}\boldsymbol{X}^\top\ \boldsymbol{X}}_{\boldsymbol{I}}\ \boldsymbol{\beta} = (\boldsymbol{X}^\top\ \boldsymbol{X})^{-1}\boldsymbol{X}^\top\ \boldsymbol{Y}.$$

or

$$\boldsymbol{\beta} = (\boldsymbol{X}^\top\ \boldsymbol{X})^{-1}\boldsymbol{X}^\top\ \boldsymbol{Y}.$$

This is the least squares solution, a simple matrix multiplication. We will see a deeper view of this LS solution next.

Another pit stop

We want to find $\beta_0, \ldots, \beta_p$ that minimizes

$$\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip})^2$$

The usual solution: Take the partial derivative with respect to each β_j and equate each of them to zero.

This leads to a linear system, called the normal equations, that we have just met in the previous slide:

$$\underbrace{\boldsymbol{X}^\top \boldsymbol{X}}_{31\times31} \underbrace{\boldsymbol{\beta}}_{31\times1} = \underbrace{\boldsymbol{X}^\top \boldsymbol{Y}}_{31\times1}.$$

Another pit stop

We want to find $\beta_0, \ldots, \beta_p$ that minimizes

$$\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip})^2$$

The usual solution: Take the partial derivative with respect to each β_j and equate each of them to zero.

This leads to a linear system, called the normal equations, that we have just met in the previous slide:

$$\underbrace{\boldsymbol{X}^\top \boldsymbol{X}}_{31\times 31} \underbrace{\boldsymbol{\beta}}_{31\times 1} = \underbrace{\boldsymbol{X}^\top \boldsymbol{Y}}_{31\times 1}.$$

If $(\boldsymbol{X}^\top \boldsymbol{X})$ is invertible, then the least-squares estimator is

$$\boldsymbol{\beta} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Y}.$$

General linear regression model

$$\boldsymbol{Y} \in \mathbb{R}^n, \qquad \boldsymbol{X} \in \mathbb{R}^{n \times (p+1)}, \qquad \boldsymbol{\beta} \in \mathbb{R}^{p+1}.$$

The first column of $\boldsymbol{X}$ is usually the vector of ones.

General linear regression model

$$\boldsymbol{Y} \in \mathbb{R}^n, \qquad \boldsymbol{X} \in \mathbb{R}^{n\times(p+1)}, \qquad \boldsymbol{\beta} \in \mathbb{R}^{p+1}.$$

The first column of $\boldsymbol{X}$ is usually the vector of ones.

Objective

Minimize

$$\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip})^2 = \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^\top (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})$$

A more theoretical viewpoint

Instead of seeing least squares only as numerical optimization, we can view it geometrically.

A more theoretical viewpoint

Instead of seeing least squares only as numerical optimization, we can view it geometrically.

- We need to work in vector spaces.
- We need a notion of distance.
- We need a notion of orthogonality.

A more theoretical viewpoint

Instead of seeing least squares only as numerical optimization, we can view it geometrically.

- We need to work in vector spaces.
- We need a notion of distance.
- We need a notion of orthogonality.

This perspective extends naturally to spaces of functions and motivates the phrase:

$$\text{least squares} = \text{orthogonal projection.}$$

Vector spaces and subspaces

Let $\mathcal{V}$ be any vector space.
A subset $\mathcal{W} \subseteq \mathcal{V}$ is a vector subspace if:

- $0 \in \mathcal{W}$;
- whenever $\boldsymbol{u}, \boldsymbol{v} \in \mathcal{W}$, then $u + v \in \mathcal{W}$;
- whenever $\boldsymbol{u} \in \mathcal{W}$ and $c \in \mathbb{R}$, then $c\boldsymbol{u} \in \mathcal{W}$.

Vector spaces and subspaces

Let $\mathcal{V}$ be any vector space.
A subset $\mathcal{W} \subseteq \mathcal{V}$ is a vector subspace if:

- $0 \in \mathcal{W}$;
- whenever $\boldsymbol{u}, \boldsymbol{v} \in \mathcal{W}$, then $u + v \in \mathcal{W}$;
- whenever $\boldsymbol{u} \in \mathcal{W}$ and $c \in \mathbb{R}$, then $c\boldsymbol{u} \in \mathcal{W}$.

Least-squares setting

The relevant subspace is the set of all linear combinations of the columns of $\boldsymbol{X}$.

Vector subspace, in a nutshell

Informally: A vector subspace $\mathcal{W}$ of a vector space $\mathcal{V}$ is a subset of $\mathcal{V}$ such that:

- the sum of two vectors in $\mathcal{W}$ remains in $\mathcal{W}$
- multiplying a vector in $\mathcal{W}$ by a scalar remains in $\mathcal{W}$
- The vector 0 (zero) belongs to $\mathcal{W}$

Vector Subspace

That's all: $\mathcal{W}$ is a vector space and we don't leave it when manipulating its vectors with addition or scalar multiplication.

Column space of $\boldsymbol{X}$

Define

$$\mathcal{C}(\boldsymbol{X}) = \{\boldsymbol{v} \in \mathbb{R}^n : \boldsymbol{v} = \boldsymbol{X}\beta \text{ for some } \beta \in \mathbb{R}^{p+1}\}.$$

Column space of $\boldsymbol{X}$

Define

$$\mathcal{C}(\boldsymbol{X}) = \{\boldsymbol{v} \in \mathbb{R}^n : \boldsymbol{v} = \boldsymbol{X}\beta \text{ for some } \beta \in \mathbb{R}^{p+1}\}.$$

- $\mathcal{C}(\boldsymbol{X})$ is the column space of $\boldsymbol{X}$.
- It is a subspace of $\mathbb{R}^n$.
- The subspace $\mathcal{C}(\boldsymbol{X})$ has dimension $\leq p+1$ (the number of columns of $\boldsymbol{X}$)
- The approximation to $\boldsymbol{Y}$ is a vector $\boldsymbol{X}\beta$ that lies in $\mathcal{C}(\boldsymbol{X})$.

Column space of $\boldsymbol{X}$

Define

$$\mathcal{C}(\boldsymbol{X}) = \{\boldsymbol{v} \in \mathbb{R}^n : \boldsymbol{v} = \boldsymbol{X}\beta \text{ for some } \beta \in \mathbb{R}^{p+1}\}.$$

- $\mathcal{C}(\boldsymbol{X})$ is the column space of $\boldsymbol{X}$.
- It is a subspace of $\mathbb{R}^n$.
- The subspace $\mathcal{C}(\boldsymbol{X})$ has dimension $\leq p+1$ (the number of columns of $\boldsymbol{X}$)
- The approximation to $\boldsymbol{Y}$ is a vector $\boldsymbol{X}\beta$ that lies in $\mathcal{C}(\boldsymbol{X})$.

Geometric problem

Find the point (vector) $\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\beta} \in \mathcal{C}(\boldsymbol{X})$ that is as close as possible to $\boldsymbol{Y}$.

What is the column space of $\boldsymbol{X}$ in a specific case?

Consider the price apartment prediction example: $\boldsymbol{X}$ has 31 columns, each column being a vector in $\mathbb{R}^{1500}$.

$$\mathcal{C}(\boldsymbol{X}) = \left\{ \beta_0 \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \\ 1 \end{pmatrix} + \beta_1 \begin{pmatrix} \text{área}_1 \\ \text{área}_2 \\ \vdots \\ \text{área}_{1499} \\ \text{área}_{1500} \end{pmatrix} + \beta_2 \begin{pmatrix} \text{idade}_1 \\ \text{idade}_2 \\ \vdots \\ \text{idade}_{1499} \\ \text{idade}_{1500} \end{pmatrix} + \ldots + \beta_{30} \begin{pmatrix} \text{salão}_1 \\ \text{salão}_2 \\ \vdots \\ \text{salão}_{1499} \\ \text{salão}_{1500} \end{pmatrix} \right\}$$

$\mathcal{C}(\boldsymbol{X})$ is a subset of $\mathbb{R}^{1500}$ vectors.

Vector $0 = (0, \ldots, 0)^\top \in \mathbb{R}^{1500}$ belongs to $\mathcal{C}(\boldsymbol{X})$

Adding two vectors of $\mathcal{C}(\boldsymbol{X})$ we still have a linear combination in $\mathcal{C}(\boldsymbol{X})$.

Multiplying a vector of $\mathcal{C}(\boldsymbol{X})$ by $c \in \mathbb{R}$ we still have a linear combination in $\mathcal{C}(\boldsymbol{X})$.

Geometry of least squares

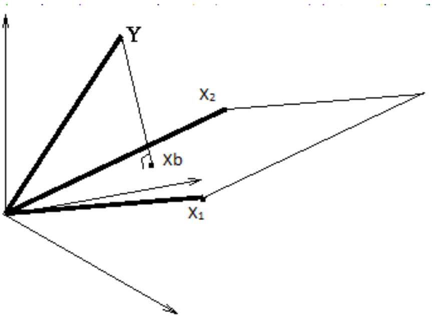


Figure: Representation of the vector $\boldsymbol{Y} \in \mathbb{R}^{1500}$. The inclined plane represents the vector subspace $\mathcal{C}(\boldsymbol{X})$ generated by a matrix $\boldsymbol{X}$, with only two columns, the vectors $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$, both in $\mathbb{R}^{1500}$. The vector subspace $\mathcal{C}(\boldsymbol{X})$ has dimension 2. Identify visually the point (vector) in $\mathcal{C}(\boldsymbol{X})$ that minimizes $\|\boldsymbol{Y} - \boldsymbol{X}\beta\|^2$.

Dimension of $\mathcal{C}(\boldsymbol{X})$

From linear algebra, if the columns (or vectors) of $\boldsymbol{X}$ are linearly independent, the dimension of $\mathcal{C}(\boldsymbol{X})$ is the number of columns.

When are they linearly dependent? When we can find a set of values for $\beta_0, \ldots, \beta_p$ such that they are not all zeros and

$$\beta_0 \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \\ 1 \end{pmatrix} + \beta_1 \begin{pmatrix} \text{área}_1 \\ \text{área}_2 \\ \vdots \\ \text{área}_{1499} \\ \text{área}_{1500} \end{pmatrix} + \beta_2 \begin{pmatrix} \text{idade}_1 \\ \text{idade}_2 \\ \vdots \\ \text{idade}_{1499} \\ \text{idade}_{1500} \end{pmatrix} + \ldots + \beta_{30} \begin{pmatrix} \text{salão}_1 \\ \text{salão}_2 \\ \vdots \\ \text{salão}_{1499} \\ \text{salão}_{1500} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix}$$

Dimension of $\mathcal{C}(\boldsymbol{X})$

Assume that $\beta_2 \neq 0$. Then, we can divide by β_2 and therefore rewrite this equation as

$$\begin{pmatrix} \text{idade}_1 \\ \text{idade}_2 \\ \vdots \\ \text{idade}_{1499} \\ \text{idade}_{1500} \end{pmatrix} = -\frac{\beta_0}{\beta_2} \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \\ 1 \end{pmatrix} - \frac{\beta_1}{\beta_2} \begin{pmatrix} \text{área}_1 \\ \text{área}_2 \\ \vdots \\ \text{área}_{1499} \\ \text{área}_{1500} \end{pmatrix} - \ldots - \frac{\beta_{30}}{\beta_2} \begin{pmatrix} \text{salão}_1 \\ \text{salão}_2 \\ \vdots \\ \text{salão}_{1499} \\ \text{salão}_{1500} \end{pmatrix}$$

It is practically impossible to write all 1500 ages as an *exact* linear combination of the remaining features (areas, number of rooms, etc.)

Hence, in most applied problems, the columns of $\boldsymbol{X}$ are linearly independent and the dimension of $\mathcal{C}(\boldsymbol{X})$ is the number of columns. We will assume this from now on.

Orthogonal projection theorem

Let $\mathcal{W} \subseteq \mathcal{V}$ be any vector subspace of the vector space $\mathcal{V} = \mathbb{R}^n$.

Theorem

For every $\boldsymbol{v} \in \mathbb{R}^n$, there exists a unique vector $\hat{\boldsymbol{v}} \in \mathcal{W}$ that minimizes

$$\|\boldsymbol{v} - \boldsymbol{w}\|, \qquad \boldsymbol{w} \in \mathcal{W}.$$

Moreover, the residual vector $\boldsymbol{v} - \hat{\boldsymbol{v}}$ is orthogonal to the subspace $\mathcal{W}$.

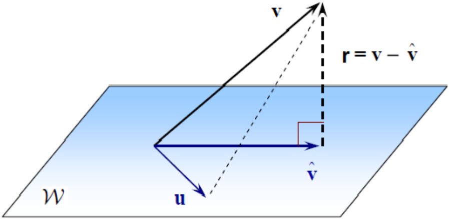


Figure: Representation of the vector $\boldsymbol{v} \in \mathbb{R}^n$, the subspace $\mathcal{W}$, the optimal $\hat{\boldsymbol{v}} \in \mathcal{W}$, and the residual vector $\boldsymbol{r} = \boldsymbol{v} - \hat{\boldsymbol{v}}$.

Notation

In linear regression, $\mathcal{W} = \mathcal{C}(\boldsymbol{X})$ and the orthogonal projection is $\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\beta}$. Remind the inner product with column vectors:

For column vectors $\boldsymbol{v}, \boldsymbol{w} \in \mathbb{R}^n$,

$$\langle \boldsymbol{v}, \boldsymbol{w} \rangle = \sum_{i=1}^{n} v_i w_i = \boldsymbol{v}^\top \cdot \boldsymbol{w} = \boxed{v_1 \quad v_2 \quad \cdots \quad v_n} \cdot \boxed{\begin{matrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{matrix}}$$

Orthogonality means

$$\boldsymbol{v} \perp \boldsymbol{w} \iff \langle \boldsymbol{v}, \boldsymbol{w} \rangle = \boxed{\quad \boldsymbol{v}^\top \quad} \cdot \boxed{\begin{matrix} \\ \\ \boldsymbol{w} \\ \\ \\ \end{matrix}} = 0$$

The theorem in the regression setting

Theorem

For every $\boldsymbol{Y} \in \mathbb{R}^n$, *there exists a unique vector* $\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} \in \mathcal{W} = \mathcal{C}(\boldsymbol{X})$ *that minimizes* $\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|$.

Moreover,

The vector $\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$ that minimizes the distance must satisfy

$$\hat{\boldsymbol{Y}} \perp (\boldsymbol{Y} - \hat{\boldsymbol{Y}}) \quad \text{or, equivalently,} \quad \boldsymbol{X}\hat{\boldsymbol{\beta}} \perp (\boldsymbol{Y} - \boldsymbol{X}\hat{\beta}).$$

To prove the theorem, we:

- show that, to have $\boldsymbol{X}\hat{\boldsymbol{\beta}} \perp (\boldsymbol{Y} - \boldsymbol{X}\hat{\beta})$, we must have $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{Y}$.
- Then, we show that, with this $\hat{\boldsymbol{\beta}}$, we minimize the distance $\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|^2$.

Proof: Orthogonality implies that $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Y}$.

Let $\boldsymbol{Y}^* = \boldsymbol{X}\beta$ for a non-null β.
Suppose that, for this $\boldsymbol{Y}^* = \boldsymbol{X}\beta$ we have

$$\boldsymbol{Y}^* \perp (\boldsymbol{Y} - \boldsymbol{Y}^*)$$

Therefore, their inner product is zero:

$$\begin{aligned} 0 &= (\boldsymbol{Y}^*)^\top \cdot (\boldsymbol{Y} - \boldsymbol{Y}^*) \\ &= (\boldsymbol{X}\beta)^\top \cdot (\boldsymbol{Y} - \boldsymbol{X}\beta) \\ &= \beta^\top \boldsymbol{X}^\top \cdot (\boldsymbol{Y} - \boldsymbol{X}\beta) \\ &= \beta^\top \left(\boldsymbol{X}^\top \boldsymbol{Y} - \boldsymbol{X}^\top \boldsymbol{X}\beta \right) \end{aligned}$$

For this expression to be zero for all possible $\boldsymbol{Y}$ and $\boldsymbol{X}$, we must have either $\beta = 0$ or the second expression in parenthesis equal to zero.

But we assume that β is a non-null vector. Therefore,

Proof: Orthogonality implies that $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{Y}$.

We have

$$0 = \boldsymbol{X}^\top \boldsymbol{Y} - \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta}$$

and therefore,

$$\boldsymbol{X}^\top \boldsymbol{Y} = \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta}$$

Multiplying both sides by $(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$, we have

$$(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{Y} = \boldsymbol{\beta}$$

Therefore, if $\boldsymbol{\beta}$ is a coefficient vector such that the orthogonality condition

$$\boldsymbol{X}\boldsymbol{\beta} \perp (\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})$$

is valid, then we must have only a single value for $\boldsymbol{\beta}$, and it is equal to

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{Y}$$

We will show now that, with this value for $\hat{\boldsymbol{\beta}}$, we minimize the distance between $\boldsymbol{Y}$ and $\boldsymbol{X}\boldsymbol{\beta}$.

Proof: $\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\beta}$ minimizes the distance to $\boldsymbol{Y}$

- Using the value for $\hat{\beta}$ that we just found, let $\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\beta} = \boldsymbol{X}\,(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{Y}\hat{\beta}$
- Consider $\|\boldsymbol{Y} - \boldsymbol{X}\beta\|^2$ for an arbitrary β.
- Add and subtract $\boldsymbol{X}\hat{\beta}$

$$\boldsymbol{Y} - \boldsymbol{X}\beta = \boldsymbol{Y} - \boldsymbol{X}\hat{\beta} + \boldsymbol{X}\hat{\beta} - \boldsymbol{X}\beta$$

$$\begin{aligned}
\|\boldsymbol{Y} - \boldsymbol{X}\beta\|^2 &= (\boldsymbol{Y} - \boldsymbol{X}\beta)^\top(\boldsymbol{Y} - \boldsymbol{X}\beta) = \\
&= \left((\boldsymbol{Y} - \boldsymbol{X}\hat{\beta}) + (\boldsymbol{X}\hat{\beta} - \boldsymbol{X}\beta)\right)^\top\left((\boldsymbol{Y} - \boldsymbol{X}\hat{\beta}) + (\boldsymbol{X}\hat{\beta} - \boldsymbol{X}\beta)\right) \\
&= (\boldsymbol{Y} - \boldsymbol{X}\hat{\beta})^\top(\boldsymbol{Y} - \boldsymbol{X}\hat{\beta}) + (\boldsymbol{Y} - \boldsymbol{X}\hat{\beta})^\top(\boldsymbol{X}\hat{\beta} - \boldsymbol{X}\beta) + \\
&\quad + (\boldsymbol{X}\hat{\beta} - \boldsymbol{X}\beta)^\top(\boldsymbol{Y} - \boldsymbol{X}\hat{\beta}) + (\boldsymbol{X}\hat{\beta} - \boldsymbol{X}\beta)^\top(\boldsymbol{X}\hat{\beta} - \boldsymbol{X}\beta) \\
&= \|\boldsymbol{Y} - \boldsymbol{X}\hat{\beta}\|^2 + \underbrace{2\,(\boldsymbol{Y} - \boldsymbol{X}\hat{\beta})^\top(\boldsymbol{X}\hat{\beta} - \boldsymbol{X}\beta)}_{A} + \|\boldsymbol{X}\hat{\beta} - \boldsymbol{X}\beta\|^2
\end{aligned}$$

Proof (cont)

We show now that $A = 0$:

$$\begin{aligned}
A &= (\boldsymbol{Y} - \boldsymbol{X}\hat{\beta})^\top (\boldsymbol{X}\hat{\beta} - \boldsymbol{X}\beta) \\
&= \left(\boldsymbol{Y} - \boldsymbol{X}\,(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{Y}\right)^\top (\boldsymbol{X}\hat{\beta} - \boldsymbol{X}\beta) \\
&= \left(((\boldsymbol{I} - \boldsymbol{X}\,(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top)\,\boldsymbol{Y})^\top\right)(\boldsymbol{X}\hat{\beta} - \boldsymbol{X}\beta) \\
&= \left(\boldsymbol{Y}^\top(\boldsymbol{I} - \boldsymbol{X}\,(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top)^\top\right)\left(\boldsymbol{X}(\hat{\beta} - \beta)\right) \\
&= (*)
\end{aligned}$$

We have

$$\begin{aligned}
\left(\boldsymbol{I} - \boldsymbol{X}\,(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\right)^\top &= \boldsymbol{I}^\top - (\boldsymbol{X}^\top)^\top\left((\boldsymbol{X}^\top\boldsymbol{X})^{-1}\right)^\top\boldsymbol{X}^\top \\
&= \boldsymbol{I} - \boldsymbol{X}\left((\boldsymbol{X}^\top\boldsymbol{X})^\top\right)^{-1}\boldsymbol{X}^\top \\
&= \boldsymbol{I} - \boldsymbol{X}\,(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top
\end{aligned}$$

Proof (cont)

Hence,

$$
\begin{aligned}
(*) &= \boldsymbol{Y}^\top \left(\boldsymbol{I} - \boldsymbol{X}\left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top\right)\boldsymbol{X}\,(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \\
&= \boldsymbol{Y}^\top \left[(I - \boldsymbol{X}\,(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top)\,X\right](\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \\
&= \boldsymbol{Y}^\top \left[\boldsymbol{X} - \boldsymbol{X}\underbrace{(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{X}}_{\boldsymbol{I}}\right](\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \\
&= \boldsymbol{Y}^\top\,[\boldsymbol{X} - \boldsymbol{X}]\,(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \\
&= \boldsymbol{Y}^\top\,[0]\,(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \\
&= 0
\end{aligned}
$$

Proof (cont)

Therefore,

$$\|\boldsymbol{Y} - \boldsymbol{X}\beta\|^2 = \|\boldsymbol{Y} - \boldsymbol{X}\hat{\beta}\|^2 + 0 + \underbrace{\|\boldsymbol{X}\hat{\beta} - \boldsymbol{X}\beta\|^2}_{\geq 0} \geq \|\boldsymbol{Y} - \boldsymbol{X}\hat{\beta}\|^2$$

As $\|\boldsymbol{X}\hat{\beta} - \boldsymbol{X}\beta\|^2$ is a distance between vectors, it must be ≥ 0.

In conclusion,

$$\hat{\beta} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Y}$$

is the vector of coefficients β that minimizes

$$\|\boldsymbol{Y} - \boldsymbol{X}\beta\|^2 .$$

Summary: Regression is orthogonal projection

- Our problem is to find $\hat{\beta}$ such that $\boldsymbol{X}\hat{\beta}$ is closest to $\boldsymbol{Y}$.
- The orthogonal projection theorem guarantees existence and uniqueness of such $\hat{\beta}$.
- The solution is the unique $\hat{\beta}$ such that $\boldsymbol{X}\hat{\beta}$ is the orthogonal projection of $\boldsymbol{Y}$ into $\mathcal{C}(\boldsymbol{X})$
- But how to find this unique $\hat{\beta}$ that provides the orthogonal projection?
- The orthogonal projection theorem also helps by saying that it must be the unique $\hat{\beta}$ such that $\boldsymbol{X}\hat{\beta}$ is orthogonal to $\boldsymbol{Y} - \boldsymbol{X}\hat{\beta}$.
- This implies that

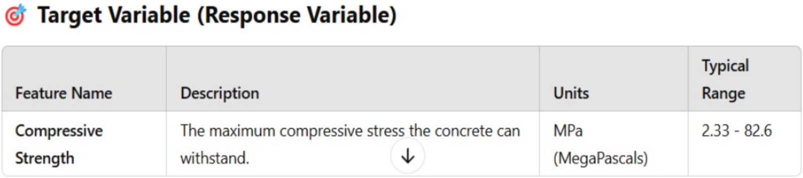
$$\hat{\beta} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Y}.$$

Concrete Compressive Strength dataset

- Dataset from Kaggle
- Concrete compressive strength is a measure of concrete's ability to withstand loads that tend to compress or reduce its volume.
- Watch some videos from YouTube
- It is one of the most critical properties of concrete, indicating its ability to withstand structural loads without breaking.
- It is defined as the maximum compressive stress that concrete can withstand before failure.
- It is determined by applying a compressive load to a concrete sample (usually a cube or cylinder) under controlled conditions until it breaks.

Kaggle dataset

- Aim: to predict compressive strength of concrete from material composition.
- Number of samples: 1030.
- Number of predictors: 8.

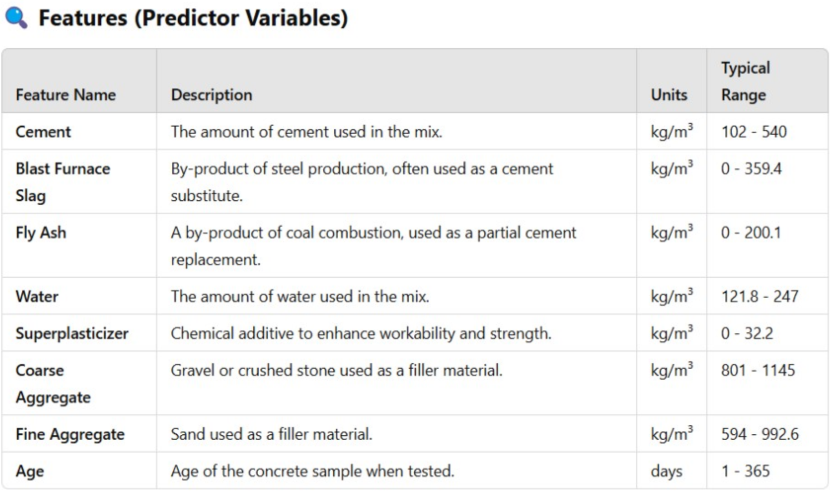


🎯 Target Variable (Response Variable)

Feature Name	Description	Units	Typical Range
Compressive Strength	The maximum compressive stress the concrete can withstand.	MPa (MegaPascals)	2.33 - 82.6

Figure: Response variable

Available features

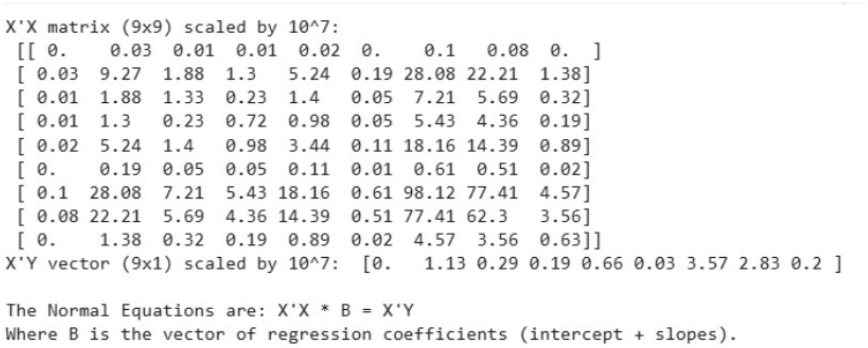


🔍 Features (Predictor Variables)

Feature Name	Description	Units	Typical Range
Cement	The amount of cement used in the mix.	kg/m³	102 - 540
Blast Furnace Slag	By-product of steel production, often used as a cement substitute.	kg/m³	0 - 359.4
Fly Ash	A by-product of coal combustion, used as a partial cement replacement.	kg/m³	0 - 200.1
Water	The amount of water used in the mix.	kg/m³	121.8 - 247
Superplasticizer	Chemical additive to enhance workability and strength.	kg/m³	0 - 32.2
Coarse Aggregate	Gravel or crushed stone used as a filler material.	kg/m³	801 - 1145
Fine Aggregate	Sand used as a filler material.	kg/m³	594 - 992.6
Age	Age of the concrete sample when tested.	days	1 - 365

Figure: Nine Features

Matrix ($\boldsymbol{X}^\top \boldsymbol{X}$)

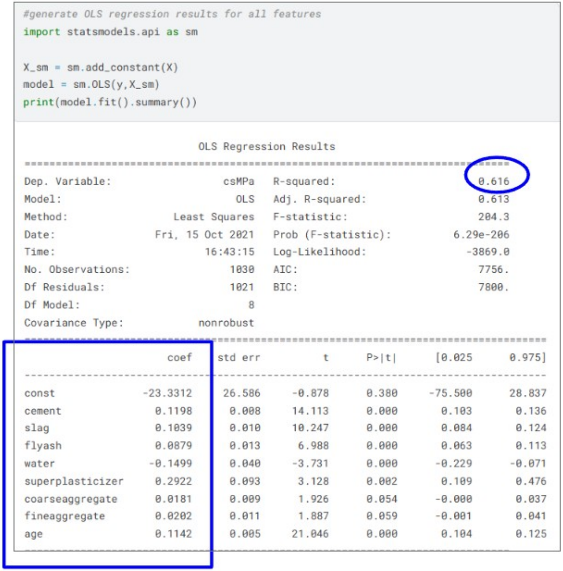


```
X'X matrix (9x9) scaled by 10^7:
 [[ 0.    0.03  0.01  0.01  0.02  0.    0.1   0.08  0.  ]
 [ 0.03  9.27  1.88  1.3   5.24  0.19 28.08 22.21  1.38]
 [ 0.01  1.88  1.33  0.23  1.4   0.05  7.21  5.69  0.32]
 [ 0.01  1.3   0.23  0.72  0.98  0.05  5.43  4.36  0.19]
 [ 0.02  5.24  1.4   0.98  3.44  0.11 18.16 14.39  0.89]
 [ 0.    0.19  0.05  0.05  0.11  0.01  0.61  0.51  0.02]
 [ 0.1  28.08  7.21  5.43 18.16  0.61 98.12 77.41  4.57]
 [ 0.08 22.21  5.69  4.36 14.39  0.51 77.41 62.3   3.56]
 [ 0.    1.38  0.32  0.19  0.89  0.02  4.57  3.56  0.63]]
X'Y vector (9x1) scaled by 10^7:  [0.   1.13 0.29 0.19 0.66 0.03 3.57 2.83 0.2 ]

The Normal Equations are: X'X * B = X'Y
Where B is the vector of regression coefficients (intercept + slopes).
```

Figure: Matrix ($\boldsymbol{X}^\top \boldsymbol{X}$)

Output with stasmodels Python library



```python
#generate OLS regression results for all features
import statsmodels.api as sm

X_sm = sm.add_constant(X)
model = sm.OLS(y,X_sm)
print(model.fit().summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  csMPa   R-squared:                       0.616
Model:                            OLS   Adj. R-squared:                  0.613
Method:                 Least Squares   F-statistic:                     204.3
Date:                Fri, 15 Oct 2021   Prob (F-statistic):          6.29e-206
Time:                        16:43:15   Log-Likelihood:                -3869.0
No. Observations:                1030   AIC:                             7756.
Df Residuals:                    1021   BIC:                             7800.
Df Model:                           8
Covariance Type:            nonrobust
====================================================================================
                       coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------
const              -23.3312     26.586     -0.878      0.380     -75.500      28.837
cement               0.1198      0.008     14.113      0.000       0.103       0.136
slag                 0.1039      0.010     10.247      0.000       0.084       0.124
flyash               0.0879      0.013      6.988      0.000       0.063       0.113
water               -0.1499      0.040     -3.731      0.000      -0.229      -0.071
superplasticizer     0.2922      0.093      3.128      0.002       0.109       0.476
coarseaggregate      0.0181      0.009      1.926      0.054      -0.000       0.037
fineaggregate        0.0202      0.011      1.887      0.059      -0.001       0.041
age                  0.1142      0.005     21.046      0.000       0.104       0.125
```

R^2 and the prediction quality

Projection and prediction quality

- If the columns of $\boldsymbol{X}$ are useful predictors...
- ...and the linear model is a reasonable approximation for $\mathbb{E}(Y \mid \boldsymbol{X} = \boldsymbol{x})$,
- then ...

$$Y \approx \hat{Y} = \boldsymbol{X}\hat{\beta}.$$

Projection and prediction quality

- If the columns of $\boldsymbol{X}$ are useful predictors...
- ...and the linear model is a reasonable approximation for $\mathbb{E}(Y \mid \boldsymbol{X} = \boldsymbol{x})$,
- then ...

$$Y \approx \hat{Y} = \boldsymbol{X}\hat{\beta}.$$

- A natural question is: *How good is this approximation? How well can we predict Y with this linear regression model?*
- How to get a measurement to work well in different problems, having widely different scales and variation ranges?
- Y can be a proportion such as unemployment rate ($Y \in (0, 1)$); in others, we talk about healthy "bad" cholesterol measurements ($<$ 130 mg/dL); some deal with house prices in millions of dollars.

Projection and prediction quality

- If the columns of $\boldsymbol{X}$ are useful predictors...
- ...and the linear model is a reasonable approximation for $\mathbb{E}(Y \mid \boldsymbol{X} = \boldsymbol{x})$,
- then ...
$$Y \approx \hat{Y} = \boldsymbol{X}\hat{\beta}.$$
- A natural question is: *How good is this approximation? How well can we predict Y with this linear regression model?*
- How to get a measurement to work well in different problems, having widely different scales and variation ranges?
- Y can be a proportion such as unemployment rate ($Y \in (0,1)$); in others, we talk about healthy "bad" cholesterol measurements ($<$ 130 mg/dL); some deal with house prices in millions of dollars.

Idea for a universal measurement

Get the total variability of Y and verify what percentage can be explained by the concomitant variation of $\boldsymbol{X}$.

Decomposing the variance of Y

- The response variable Y (apartment prices) varies widely, some apartments being extremely expensive while others have low prices. Why this happens?
- We hope that the features $\boldsymbol{X}$ (area, number of rooms, etc.) will be able to "explain" most of this variation in Y.
- We hope $\boldsymbol{X}$ contains the main factors causing the change in Y.
- An expensive apartment can have its high price "explained" by its large area, great number of rooms, etc.
- The low price of another is explained by its small area, only one room, etc.
- In a good linear regression model, the variation of prices is mostly "explained" by the variation of the features.

Decomposing the variance of Y

- We will be able to, in **any** regression problem, find a way to decompose the total variation of the response variable Y into two components:
  1. The first one captures the part of this total variation of Y that is "explained" by the variation of the features in $\boldsymbol{X}$.
  2. The second component contains the residual variation, what is not associated with the variation of $\boldsymbol{X}$.
  3. This second component is "explained" by other factors different from those in $\boldsymbol{X}$.
  4. These missing factors could be variables that are: impossible to measure, too expensive to collect, with so small an impact that it is not worthwhile to collect, or those that we are not even aware of their influence.
- A good linear regression model, that predicts well Y based on $\boldsymbol{X}$, should have the first component much larger than the second one.

Decomposition around the mean

- Let $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$, the average value of the response variable (a scalar number).
- Define the $n \times 1$ column-vector given by $\bar{\boldsymbol{Y}} = (\bar{Y}, \bar{Y}, \ldots, \bar{Y})^\top = \bar{Y}\,(1, 1, \ldots, 1)^\top = \bar{Y}\,\mathbf{1}^\top$
- By adding and subtracting the same vector, we can write $\boldsymbol{Y}$ as

$$\boldsymbol{Y} - \bar{\boldsymbol{Y}} = (\hat{\boldsymbol{Y}} - \bar{\boldsymbol{Y}}) + (\boldsymbol{Y} - \hat{\boldsymbol{Y}})$$

Decomposition around the mean

$\boldsymbol{Y} - \bar{\boldsymbol{Y}} = (\hat{\boldsymbol{Y}} - \bar{\boldsymbol{Y}}) + (\boldsymbol{Y} - \hat{\boldsymbol{Y}})$ means

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{1499} \\ y_{1500} \end{bmatrix} - \begin{bmatrix} \bar{y} \\ \bar{y} \\ \vdots \\ \bar{y} \\ \bar{y} \end{bmatrix} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_{1499} \\ \hat{y}_{1500} \end{bmatrix} - \begin{bmatrix} \bar{y} \\ \bar{y} \\ \vdots \\ \bar{y} \\ \bar{y} \end{bmatrix} + \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{1499} \\ y_{1500} \end{bmatrix} - \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_{1499} \\ \hat{y}_{1500} \end{bmatrix}$$

or, equivalently,

$$\underbrace{\begin{bmatrix} y_1 - \bar{y} \\ y_2 - \bar{y} \\ \vdots \\ y_{1499} - \bar{y} \\ y_{1500} - \bar{y} \end{bmatrix}}_{\boldsymbol{Y} - \bar{y}\mathbf{1}} = \underbrace{\begin{bmatrix} \hat{y}_1 - \bar{y} \\ \hat{y}_2 - \bar{y} \\ \vdots \\ \hat{y}_{1499} - \bar{y} \\ \hat{y}_{1500} - \bar{y} \end{bmatrix}}_{\hat{\boldsymbol{Y}} - \bar{y}\mathbf{1}} + \underbrace{\begin{bmatrix} y_1 - \hat{y}_1 \\ y_2 - \hat{y}_2 \\ \vdots \\ y_{1499} - \hat{y}_{1499} \\ y_{1500} - \hat{y}_{1500} \end{bmatrix}}_{\boldsymbol{Y} - \hat{\boldsymbol{Y}}}$$

We decompose the vector $\boldsymbol{Y}$ (centered on its mean) into two pieces. Let us look at what these components are.

Decomposition around the mean

- What do we gain by writing $\boldsymbol{Y} - \bar{\boldsymbol{Y}} = (\hat{\boldsymbol{Y}} - \bar{\boldsymbol{Y}}) + (\boldsymbol{Y} - \hat{\boldsymbol{Y}})$?
- The right-hand side vectors are orthogonal to each other (left as an exercise). Therefore,

$$\underbrace{||\boldsymbol{Y} - \bar{\boldsymbol{Y}}||^2}_{\text{SSTO}} = \underbrace{||\hat{\boldsymbol{Y}} - \bar{\boldsymbol{Y}}||^2}_{\text{SSReg}} + \underbrace{||\boldsymbol{Y} - \hat{\boldsymbol{Y}}||^2}_{\text{SSE}}$$

- The left-hand side does not depend on the linear model. It is the variation of Y around its global (marginal mean).

Decomposition around the mean

- What do we gain by writing $\boldsymbol{Y} - \bar{\boldsymbol{Y}} = (\hat{\boldsymbol{Y}} - \bar{\boldsymbol{Y}}) + (\boldsymbol{Y} - \hat{\boldsymbol{Y}})$?
- The right-hand side vectors are orthogonal to each other (left as an exercise). Therefore,

$$\underbrace{||\boldsymbol{Y} - \bar{\boldsymbol{Y}}||^2}_{\text{SSTO}} = \underbrace{||\hat{\boldsymbol{Y}} - \bar{\boldsymbol{Y}}||^2}_{\text{SSReg}} + \underbrace{||\boldsymbol{Y} - \hat{\boldsymbol{Y}}||^2}_{\text{SSE}}$$

- The left-hand side does not depend on the linear model. It is the variation of Y around its global (marginal mean).
- The last term represents what remains of variation after we use $\boldsymbol{X}$ to predict in Y. If this term is zero, we have a perfect predicition.

Decomposition around the mean

- What do we gain by writing $\boldsymbol{Y} - \bar{\boldsymbol{Y}} = (\hat{\boldsymbol{Y}} - \bar{\boldsymbol{Y}}) + (\boldsymbol{Y} - \hat{\boldsymbol{Y}})$?
- The right-hand side vectors are orthogonal to each other (left as an exercise). Therefore,

$$\underbrace{||\boldsymbol{Y} - \bar{\boldsymbol{Y}}||^2}_{\text{SSTO}} = \underbrace{||\hat{\boldsymbol{Y}} - \bar{\boldsymbol{Y}}||^2}_{\text{SSReg}} + \underbrace{||\boldsymbol{Y} - \hat{\boldsymbol{Y}}||^2}_{\text{SSE}}$$

- The left-hand side does not depend on the linear model. It is the variation of Y around its global (marginal mean).
- The last term represents what remains of variation after we use $\boldsymbol{X}$ to predict in Y. If this term is zero, we have a perfect predicition.
- The first term of the right-hand side represents how much variation we have in $\hat{\boldsymbol{Y}}$ around its mean (as an exercise, also show that the mean of $\bar{\boldsymbol{Y}}$ is $\bar{y}$.
- It represents the amount of variation present in the predicted Y's.

Decomposition around the mean

- What do we gain by writing $\boldsymbol{Y} - \bar{\boldsymbol{Y}} = (\hat{\boldsymbol{Y}} - \bar{\boldsymbol{Y}}) + (\boldsymbol{Y} - \hat{\boldsymbol{Y}})$?
- The right-hand side vectors are orthogonal to each other (left as an exercise). Therefore,

$$\underbrace{||\boldsymbol{Y} - \bar{\boldsymbol{Y}}||^2}_{\text{SSTO}} = \underbrace{||\hat{\boldsymbol{Y}} - \bar{\boldsymbol{Y}}||^2}_{\text{SSReg}} + \underbrace{||\boldsymbol{Y} - \hat{\boldsymbol{Y}}||^2}_{\text{SSE}}$$

- The left-hand side does not depend on the linear model. It is the variation of Y around its global (marginal mean).
- The last term represents what remains of variation after we use $\boldsymbol{X}$ to predict in Y. If this term is zero, we have a perfect predicition.
- The first term of the right-hand side represents how much variation we have in $\hat{\boldsymbol{Y}}$ around its mean (as an exercise, also show that the mean of $\bar{\boldsymbol{Y}}$ is $\bar{y}$.
- It represents the amount of variation present in the predicted Y's.
- The sum of the parts on the right is constant and does not depend on the model. If the model is good, it should make the last part very small compared to the total on the left-hand side.

Coefficient of determination R^2

A residual vector with a small length indicates a good fit of the linear regression model.

The usual summary measure is

$$\begin{aligned} R^2 &= 1 - \frac{\|\boldsymbol{Y} - \hat{\boldsymbol{Y}}\|^2}{\|\boldsymbol{Y} - \bar{y}\,\mathbf{1}\,\|^2} \\ &= 1 - \frac{\text{SSE}}{\text{SSTo}} \\ &= \frac{\|\hat{\boldsymbol{Y}} - \bar{\boldsymbol{Y}}\|^2}{\|\boldsymbol{Y} - \bar{y}\,\mathbf{1}\,\|^2} \\ &= \frac{\text{SSReg}}{\text{SSTo}}. \end{aligned}$$

Coefficient of determination R^2

A residual vector with a small length indicates a good fit of the linear regression model.

The usual summary measure is

$$
\begin{aligned}
R^2 &= 1 - \frac{\| \boldsymbol{Y} - \hat{\boldsymbol{Y}} \|^2}{\| \boldsymbol{Y} - \bar{y}\,\mathbf{1} \|^2} \\
&= 1 - \frac{\text{SSE}}{\text{SSTo}} \\
&= \frac{\| \hat{\boldsymbol{Y}} - \bar{\boldsymbol{Y}} \|^2}{\| \boldsymbol{Y} - \bar{y}\,\mathbf{1} \|^2} \\
&= \frac{\text{SSReg}}{\text{SSTo}}.
\end{aligned}
$$

- R^2 close to 1 indicates a strong fit.
- R^2 compares the explained variability (the variation of $\hat{\boldsymbol{Y}}$) with the total variability of $\boldsymbol{Y}$.

Hat matrix

Recall:

$$\hat{\beta} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Y}.$$

Therefore,

$$\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\beta} = \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Y} = \boldsymbol{H}\boldsymbol{Y},$$

where

$$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top.$$

Hat matrix

Recall:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Y}.$$

Therefore,

$$\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Y} = \boldsymbol{H}\boldsymbol{Y},$$

where

$$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top.$$

$\boldsymbol{H}$ is called the *hat matrix* because it puts a hat on $\boldsymbol{Y}$.

Basic properties of the hat matrix

- $\boldsymbol{H}$ is an $n \times n$ matrix.
- $\boldsymbol{H}$ is the orthogonal projection matrix onto $\mathcal{C}(\boldsymbol{X})$.
- $\boldsymbol{H}^2 = \boldsymbol{H}$ (idempotent).
- $\boldsymbol{H}^\top = \boldsymbol{H}$ (symmetric).
- The proof of these results is left as exercise.

Basic properties of the hat matrix

- $\boldsymbol{H}$ is an $n \times n$ matrix.
- $\boldsymbol{H}$ is the orthogonal projection matrix onto $\mathcal{C}(\boldsymbol{X})$.
- $\boldsymbol{H}^2 = \boldsymbol{H}$ (idempotent).
- $\boldsymbol{H}^\top = \boldsymbol{H}$ (symmetric).
- The proof of these results is left as exercise.

Geometric meaning

Applying $\boldsymbol{H}$ to any vector $\boldsymbol{Y}$ returns its projection $\boldsymbol{HY}$ onto the column space of $\boldsymbol{X}$.