Least squares method

Used to approximate solutions of overdetermined systems of equations, i.e., systems where the number of equations is bigger than the number of unknowns:

$$Ax = b, \qquad A \in \mathbb{R}^{m\times n},\ b \in \mathbb{R}^m,\ m > n$$

In general, it is impossible to find an exact solution of the linear system above (i.e., an x such that $b - Ax = 0$), but we can look for an approximate solution x such that $b - Ax \approx 0$. The name "least squares" means that the solution minimises the sum of the squares of the errors made in every single equation.

In data fitting, the best fit in the least square sense minimises the sum of the squares of the *residuals*, each residual being the difference between the observed value and the value provided by the model used.

most popular example: linear regression

We have a set of data $(x_1, y_1), (x_2, y_2), \cdots, (x_m, y_m)$, where x_i are the independent variables, (all distinct), and y_i are the observations; m is big.

most popular example: linear regression

We have a set of data $(x_1, y_1), (x_2, y_2), \cdots, (x_m, y_m)$, where x_i are the independent variables, (all distinct), and y_i are the observations; m is big.

We want to fit the data with a straight line $p(x) = a_1 + a_2 x$; finding a straight line such that $p(x_i) = y_i,\ i = 1, \cdots, m$ is impossible, unless all the y_i are aligned. Then we look for the straight line that minimises

$$S = \sum_{i=1}^{m} (y_i - p(x_i))^2.$$

most popular example: linear regression

We have a set of data $(x_1, y_1), (x_2, y_2), \cdots, (x_m, y_m)$, where x_i are the independent variables, (all distinct), and y_i are the observations; m is big.

We want to fit the data with a straight line $p(x) = a_1 + a_2 x$; finding a straight line such that $p(x_i) = y_i, \ i = 1, \cdots, m$ is impossible, unless all the y_i are aligned. Then we look for the straight line that minimises

$$S = \sum_{i=1}^{m} (y_i - p(x_i))^2.$$

Each quantity $y_i - p(x_i)$ is a *residual*, that is, the difference between the observed value y_i and the value $p(x_i)$ predicted by our model (a straight line in this case).

most popular example: linear regression

We have a set of data $(x_1, y_1), (x_2, y_2), \cdots, (x_m, y_m)$, where x_i are the independent variables, (all distinct), and y_i are the observations; m is big.

We want to fit the data with a straight line $p(x) = a_1 + a_2 x$; finding a straight line such that $p(x_i) = y_i,\ i = 1, \cdots, m$ is impossible, unless all the y_i are aligned. Then we look for the straight line that minimises

$$S = \sum_{i=1}^{m} (y_i - p(x_i))^2.$$

Each quantity $y_i - p(x_i)$ is a *residual*, that is, the difference between the observed value y_i and the value $p(x_i)$ predicted by our model (a straight line in this case). By using the expression of $p(x)$ in S, we see that we have to minimise a function of two unknowns, a_1, and a_2:

$$F(a_1, a_2) = \sum_{i=1}^{m} (y_i - (a_1 + a_2 x_i))^2.$$

$$\boxed{F(a_1, a_2) = \sum_{i=1}^{m}(y_i - (a_1 + a_2 x_i))^2}$$

$$\boxed{F(a_1, a_2) = \sum_{i=1}^{m} (y_i - (a_1 + a_2 x_i))^2}$$

F, as a function of a_1 and a_2, is a polynomial of degree 2:

$$F(a_1, a_2) = \sum_{i=1}^{m} (y_i^2 + a_1^2 + x_i^2 a_2^2 + 2x_i a_1 a_2 - 2y_i a_1 - 2y_i x_i a_2)$$

$$= m a_1^2 + \left(\sum_{i=1}^{m} x_i^2\right) a_2^2 + 2\left(\sum_{i=1}^{m} x_i\right) a_1 a_2 - 2\left(\sum_{i=1}^{m} y_i\right) a_1 - 2\left(\sum_{i=1}^{m} y_i x_i\right) a_2 + \sum_{i=1}^{m} y_i^2.$$

$$\boxed{F(a_1, a_2) = \sum_{i=1}^{m}(y_i - (a_1 + a_2 x_i))^2}$$

F, as a function of a_1 and a_2, is a polynomial of degree 2:

$$F(a_1, a_2) = \sum_{i=1}^{m}(y_i^2 + a_1^2 + x_i^2 a_2^2 + 2x_i a_1 a_2 - 2y_i a_1 - 2y_i x_i a_2)$$

$$= ma_1^2 + \left(\sum_{i=1}^{m} x_i^2\right) a_2^2 + 2\left(\sum_{i=1}^{m} x_i\right) a_1 a_2 - 2\left(\sum_{i=1}^{m} y_i\right) a_1 - 2\left(\sum_{i=1}^{m} y_i x_i\right) a_2 + \sum_{i=1}^{m} y_i^2.$$

What kind of surface is $z = F(a_1, a_2)$, in the space of variables a_1, a_2 and z? We have to look at the determinant of the second-order part of F:

$$det \begin{pmatrix} m & \sum_{i=1}^{m} x_i \\ \sum_{i=1}^{m} x_i & \sum_{i=1}^{m} (x_i)^2 \end{pmatrix} = m \sum_{i=1}^{m} (x_i)^2 - \left(\sum_{i=1}^{m} x_i\right)^2$$

$$\boxed{F(a_1, a_2) = \sum_{i=1}^{m}(y_i - (a_1 + a_2 x_i))^2}$$

F, as a function of a_1 and a_2, is a polynomial of degree 2:

$$F(a_1, a_2) = \sum_{i=1}^{m}(y_i^2 + a_1^2 + x_i^2 a_2^2 + 2x_i a_1 a_2 - 2y_i a_1 - 2y_i x_i a_2)$$

$$= ma_1^2 + \Big(\sum_{i=1}^{m} x_i^2\Big)a_2^2 + 2\Big(\sum_{i=1}^{m} x_i\Big)a_1 a_2 - 2\Big(\sum_{i=1}^{m} y_i\Big)a_1 - 2\Big(\sum_{i=1}^{m} y_i x_i\Big)a_2 + \sum_{i=1}^{m} y_i^2.$$

We use the classical inequality

$$\sum_{i=1}^{m} x_i \equiv \sum_{i=1}^{m} x_i \cdot 1 \leq \Big(\sum_{i=1}^{m}(x_i)^2\Big)^{1/2}\Big(\sum_{i=1}^{m} 1^2\Big)^{1/2},$$

$$\boxed{F(a_1, a_2) = \sum_{i=1}^m (y_i - (a_1 + a_2 x_i))^2}$$

F, as a function of a_1 and a_2, is a polynomial of degree 2:

$$F(a_1, a_2) = \sum_{i=1}^{m} (y_i^2 + a_1^2 + x_i^2 a_2^2 + 2x_i a_1 a_2 - 2y_i a_1 - 2y_i x_i a_2)$$

$$= m a_1^2 + \Big(\sum_{i=1}^{m} x_i^2\Big) a_2^2 + 2\Big(\sum_{i=1}^{m} x_i\Big) a_1 a_2 - 2\Big(\sum_{i=1}^{m} y_i\Big) a_1 - 2\Big(\sum_{i=1}^{m} y_i x_i\Big) a_2 + \sum_{i=1}^{m} y_i^2 .$$

We use the classical inequality

$$\sum_{i=1}^{m} x_i \equiv \sum_{i=1}^{m} x_i \cdot 1 \leq \Big(\sum_{i=1}^{m} (x_i)^2\Big)^{1/2} \Big(\sum_{i=1}^{m} 1^2\Big)^{1/2},$$

$((\underline{x}, \underline{1}) \leq \|\underline{x}\| \|\underline{1}\|)$. In our case $(\underline{x}, \underline{1}) < \|\underline{x}\| \|\underline{1}\|$ and we square it to obtain

$$(\sum_{i=1}^{m} x_i)^2 < m \Big(\sum_{i=1}^{m} (x_i)^2\Big)$$

$$\boxed{F(a_1, a_2) = \sum_{i=1}^{m} (y_i - (a_1 + a_2 x_i))^2}$$

F, as a function of a_1 and a_2, is a polynomial of degree 2:

$$F(a_1, a_2) = \sum_{i=1}^{m} (y_i^2 + a_1^2 + x_i^2 a_2^2 + 2x_i a_1 a_2 - 2y_i a_1 - 2y_i x_i a_2)$$

$$= m a_1^2 + \Big(\sum_{i=1}^{m} x_i^2\Big) a_2^2 + 2\Big(\sum_{i=1}^{m} x_i\Big) a_1 a_2 - 2\Big(\sum_{i=1}^{m} y_i\Big) a_1 - 2\Big(\sum_{i=1}^{m} y_i x_i\Big) a_2 + \sum_{i=1}^{m} y_i^2.$$

We use the classical inequality

$$\sum_{i=1}^{m} x_i \equiv \sum_{i=1}^{m} x_i \cdot 1 \leq \Big(\sum_{i=1}^{m} (x_i)^2\Big)^{1/2} \Big(\sum_{i=1}^{m} 1^2\Big)^{1/2},$$

$((\underline{x}, \underline{1}) \leq \|\underline{x}\| \|\underline{1}\|)$. In our case $(\underline{x}, \underline{1}) < \|\underline{x}\| \|\underline{1}\|$ and we square it to obtain

$$(\sum_{i=1}^{m} x_i)^2 < m \Big(\sum_{i=1}^{m} (x_i)^2\Big) \Rightarrow m \Big(\sum_{i=1}^{m} (x_i)^2\Big) - (\sum_{i=1}^{m} x_i)^2 > 0$$

This ensures that

$$det \begin{pmatrix} m & \sum_{i=1}^{m} x_i \\ \sum_{i=1}^{m} x_i & \sum_{i=1}^{m} (x_i)^2 \end{pmatrix} = m \sum_{i=1}^{m} (x_i)^2 - \Big(\sum_{i=1}^{m} x_i \Big)^2$$

is positive, and since the coefficients of a_1^2 and a_2^2 are positive, $z = F(a_1, a_2)$ is a paraboloid going to $+\infty$ when $a_1^2 + a_2^2$ goes to infinity (roughly speaking, it is "a cup"). Hence F has a unique minimum, and the point of minimum is the point where the gradient of F vanishes.

∇F is given by

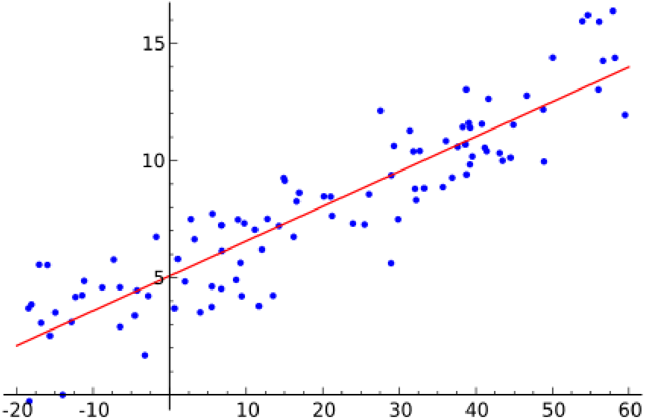
$$\nabla F(a_1, a_2) = \begin{bmatrix} \dfrac{\partial F}{\partial a_1} \\ \dfrac{\partial F}{\partial a_2} \end{bmatrix} = \begin{bmatrix} 2\displaystyle\sum_{i=1}^{m}(y_i - (a_1 + a_2 x_i))(-1) \\ 2\displaystyle\sum_{i=1}^{m}(y_i - (a_1 + a_2 x_i))(-x_i) \end{bmatrix}$$

By imposing $\nabla F = 0$ we obtain a system of two equations in the two unknowns a_1, a_2:

$$(LS1) \qquad \begin{pmatrix} m & \displaystyle\sum_{i=1}^{m} x_i \\ \displaystyle\sum_{i=1}^{m} x_i & \displaystyle\sum_{i=1}^{m} (x_i)^2 \end{pmatrix} \begin{pmatrix} a_1 \\ \\ a_2 \end{pmatrix} = \begin{pmatrix} \displaystyle\sum_{i=1}^{m} y_i \\ \displaystyle\sum_{i=1}^{m} x_i y_i \end{pmatrix}$$

Let $\overline{a}_1, \overline{a}_2$ be the solution, then

$$p(x) = \overline{a}_1 + \overline{a}_2 x \quad \text{is the linear regression line}$$


15
10
5
-20
-10
10
20
30
40
50
60

Other models

Least square method is widely applied in many fields (economics, statistics, stock-market and the like) to predict the behaviour of a phenomenon for which the values (x_i, y_i), for $i = 1, ..., m$) are samples (or experimental data).

In different cases, our "guess" could be different from the linear case discussed so far. Actually different models are used, in different circumstances, in order to have a better fitting of the data.

For example, if the data show a quadratic distribution we might use a parabola $p(x) = a_1 + a_2 x + a_3 x^2$. In this case S would become

$$F(a_1, a_2, a_3) = \sum_{i=1}^{m} (y_i - (a_1 + a_2 x_i + a_3 x_i^2))^2.$$

Other models

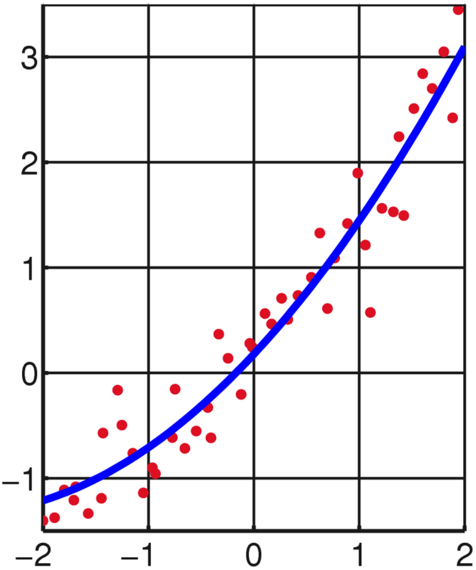
$$\boxed{F(a_1, a_2, a_3) = \sum_{i=1}^m (y_i - (a_1 + a_2 x_i + a_3 x_i^2))^2.}$$

Proceeding as before, computing ∇F and imposing $\nabla F = 0$ to find the point of minimum we obtain a 3×3 system in the 3 unknowns a_1, a_2, a_3:

$$(LS2) \qquad \begin{pmatrix} m & \sum_{i=1}^m x_i & \sum_{i=1}^m (x_i)^2 \\ \sum_{i=1}^m x_i & \sum_{i=1}^m (x_i)^2 & \sum_{i=1}^m (x_i)^3 \\ \sum_{i=1}^m (x_i)^2 & \sum_{i=1}^m (x_i)^3 & \sum_{i=1}^m (x_i)^4 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^m y_i \\ \sum_{i=1}^m x_i y_i \\ \sum_{i=1}^m (x_i)^2 y_i \end{pmatrix}$$

The solution of the system is the parabola that gives the best fit of the given data.

Example of a least square parabola



More general models

More generally, least square method can be used to fit a set of data with a linear combination of functions (not necessarily monomials like 1 and x) chosen to best fit the distribution of a given cloud of data.

More general models

More generally, least square method can be used to fit a set of data with a linear combination of functions (not necessarily monomials like 1 and x) chosen to best fit the distribution of a given cloud of data.

Let S_n (with $n << m$) be a finite dimensional space:

$S_n = span\{\varphi_1, \varphi_2, \cdots, \varphi_n\}$. We look for a function $p(x) = \sum_{j=1}^{n} a_j \varphi_j(x)$

such that

$$F(a_1, a_2, \cdots, a_n) := \sum_{i=1}^{m} (y_i - \sum_{j=1}^{n} a_j \varphi_j(x_i))^2 = \text{ minimum}$$

More general models

As before, computing ∇F and imposing $\nabla F = 0$, the point of minimum will be the solution of the $n \times n$ linear system in the n unknowns $a_1, a_2, \cdots, a_n$:

$$\begin{pmatrix} \sum_{i=1}^{m} (\varphi_1(x_i))^2 & \sum_{i=1}^{m} \varphi_1(x_i)\varphi_2(x_i) & \cdots & \sum_{i=1}^{m} \varphi_1(x_i)\varphi_n(x_i) \\ & \sum_{i=1}^{m} (\varphi_2(x_i))^2 & \cdots & \sum_{i=1}^{m} \varphi_2(x_i)\varphi_n(x_i) \\ symm & & \ddots & \vdots \\ & & & \sum_{i=1}^{m} (\varphi_n(x_i))^2 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{m} y_i \varphi_1(x_i) \\ \sum_{i=1}^{m} y_i \varphi_2(x_i) \\ \vdots \\ \sum_{i=1}^{m} y_i \varphi_n(x_i) \end{pmatrix}$$

A different approach

The least square systems $(LS1), (LS2)$, and the general one here above can be obtained with a different procedure. Let us see how, in the simplest case of the linear regression line.

We are looking for a line of equation $p(x) = a_1 + a_2 x$ such that $p(x_i) = y_i,\ \ i = 1, \cdots, m$: we have m equations in 2 unknowns which, in matrix form, is the overdetermined system

$$\underbrace{\begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_m \end{bmatrix}}_{A} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \underbrace{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}}_{b} \tag{1}$$

By left-multiplying the system by A^T we obtain $A^T(A\underline{a} - \underline{b}) = 0$:

$$\underbrace{\begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_m \end{bmatrix}}_{A^T} \underbrace{\begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_m \end{bmatrix}}_{A} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_m \end{bmatrix}}_{A^T} \underbrace{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}}_{\underline{b}}$$

$$\Downarrow$$

$$(LS1) \qquad \underbrace{\begin{pmatrix} m & \sum_{i=1}^{m} x_i \\ \sum_{i=1}^{m} x_i & \sum_{i=1}^{m} (x_i)^2 \end{pmatrix}}_{A^T A} \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = \underbrace{\begin{pmatrix} \sum_{i=1}^{m} y_i \\ \sum_{i=1}^{m} x_i y_i \end{pmatrix}}_{A^T \underline{b}}$$

This is the same 2×2 system obtained with the least square approach.

The solution of $(LS1)$ is called the least square solution of (1), and it exists provided that $A^T A$ is non-singular. This is true if the matrix A has full rank (rank 2 in this case). Indeed:

$A^T A$ is always symmetric and positive semidefinite, for every matrix $A \not\equiv 0$:

$$(A^T A\underline{z}, \underline{z}) = (A\underline{z}, A\underline{z}) = \|A\underline{z}\|^2 \geq 0$$

$A^T A$ is positive definite if A has full rank (that is: if $A\underline{z} = \underline{0}$ implies that $\underline{z} = \underline{0}$).
In fact, if $A\underline{z} = \underline{0} \rightarrow \underline{z} = \underline{0}$ then

$$(A^T A\underline{z}, \underline{z}) = \|A\underline{z}\|^2 = 0 \Longleftrightarrow \underline{z} = \underline{0}.$$

Note: if A has full rank, $(LS1)$ has always a solution, even if system (1) has no solutions