Statistics of least squares and Lasso estimators (part I)
Dual computational approaches in machine learning (part II)

Saverio Salzo

20 Gennaio 2023

School on “The mathematics of Machine Learning”
Scuola Normale Superiore, Pisa

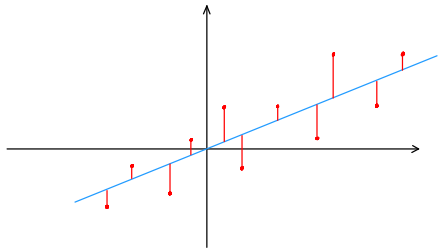
CLASSICAL LINEAR REGRESSION

Given n data points

$$(x_1, y_1), \ldots, (x_n, y_n),$$

such that

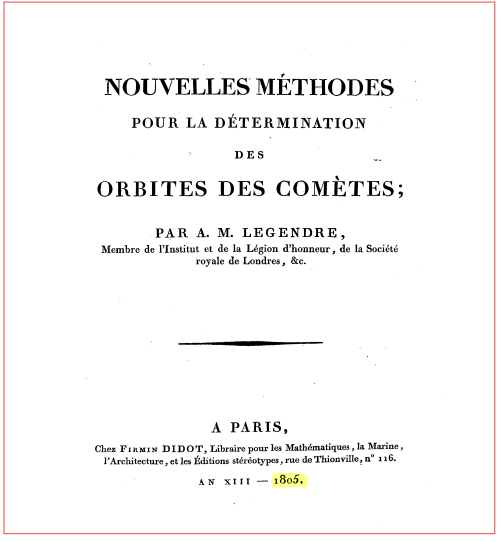
$$\langle \theta, x_i \rangle \approx y_i \quad (i = 1, \ldots, n),$$



the **method of least squares** (Legendre 1805, Gauss 1809-1795?) consists in determining the parameter θ by solving

$$\min_{\theta \in \mathbb{R}^d} \sum_{i=1}^{n} |\langle \theta, x_i \rangle - y_i|^2$$

CLASSICAL LINEAR REGRESSION



NOUVELLES MÉTHODES

POUR LA DÉTERMINATION

DES

ORBITES DES COMÈTES;

PAR A. M. LEGENDRE,

Membre de l'Institut et de la Légion d'honneur, de la Société royale de Londres, &c.

A PARIS,

Chez FIRMIN DIDOT, Libraire pour les Mathématiques, la Marine, l'Architecture, et les Éditions stéréotypes, rue de Thionville, n° 116.

AN XIII — 1805.

CLASSICAL LINEAR REGRESSION

APPENDICE.

Sur la Méthode des moindres quarrés.

DANS la plupart des questions où il s'agit de tirer des mesures données par l'observation, les résultats les plus exacts qu'elles peuvent offrir, on est presque toujours conduit à un système d'équations de la forme

$$E = a + bx + cy + fz + \&c.$$

dans lesquelles a, b, c, f, &c. sont des coëfficiens connus, qui varient d'une équation à l'autre, et x, y, z, &c. sont des inconnues qu'il faut déterminer par la condition que la valeur de E se réduise, pour chaque équation, à une quantité ou nulle ou très-petite.

Si l'on a autant d'équations que d'inconnues x, y, z, &c., il n'y a aucune difficulté pour la détermination de ces inconnues, et on peut rendre les erreurs E absolument nulles. Mais le plus souvent, le nombre des équations est supérieur à celui des inconnues, et il est impossible d'anéantir toutes les erreurs.

Dans cette circonstance, qui est celle de la plupart des problèmes physiques et astronomiques, où l'on cherche à déterminer quelques élémens importans, il entre nécessairement de l'arbitraire dans la distribution des erreurs, et on ne doit pas s'attendre que toutes les hypothèses conduiront exactement aux mêmes résultats; mais il faut sur-tout faire en sorte que les erreurs extrêmes, sans avoir égard à leurs signes, soient renfermées dans les limites les plus étroites qu'il est possible.

De tous les principes qu'on peut proposer pour cet objet, je pense qu'il n'en est pas de plus général, de plus exact, ni d'une application plus facile que celui dont nous avons fait usage dans les recherches précédentes, et qui consiste à rendre

CLASSICAL LINEAR REGRESSION

LEGENDRE

On Least Squares

(Translated from the French by Professor Henry A. Ruger and Professor Helen M. Walker, Teachers College, Columbia University, New York City.)

The great advances in mathematical astronomy made during the early year) of the nineteenth century were due in no small part to the development of the method of least squares. The same method is the foundation for the calculus of errors of observation now occupying a place of great importance in the scientific study of social, economic, biological, and psychological problems. Gauss says in his work on the *Theory of the Motions of the Heavenly Bodies* (1809) that he had made use of this principle since 1795 but that it was first published by Legendre. The first statement of the method appeared as an appendix entitled "Sur la Méthode des moindres quarrés" in Legendre's *Nouvelles méthodes pour la détermination des orbites des comètes*, Paris, 1805. The portion of the work translated here is found on pages 72–75.

Adrien-Marie Legendre (1752–1833) was for five years a professor of mathematics in the École Militaire at Paris, and his early studies on the paths of projectiles provided a background for later work on the paths of heavenly bodies. He wrote on astronomy, the theory of numbers, elliptic functions, the calculus, higher geometry, mechanics, and physics. His work on geometry, in which he rearranged the propositions of Euclid, is one of the most successful textbooks ever written.

On the Method of Least Squares

In the majority of investigations in which the problem is to get from measures given by observation the most exact results which they can furnish, there almost always arises a system of equations of the form

$$E = a + bx + cy + fz + \&c.$$

CLASSICAL LINEAR REGRESSION

On the Method of Least Squares

In the majority of investigations in which the problem is to get from measures given by observation the most exact results which they can furnish, there almost always arises a system of equations of the form

$$E = a + bx + cy + fz + \&c.$$

in which a, b, c, f, &c. are the known coefficients which vary from one equation to another, and x, y, z, &c. are the unknowns which must be determined in accordance with the condition that the value of E shall for each equation reduce to a quantity which is either zero or very small.

If there are the same number of equations as unknowns x, y, z, &c., there is no difficulty in determining the unknowns, and the error E can be made absolutely zero. But more often the number

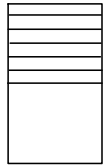
576

CLASSICAL LINEAR REGRESSION

of equations is greater than that of the unknowns, and it is impossible to do away with all the errors.

In a situation of this sort, which is the usual thing in physical and astronomical problems, where there is an attempt to determine certain important components, a degree of arbitrariness necessarily enters in the distribution of the errors, and it is not to be expected that all the hypotheses shall lead to exactly the same results; but it is particularly important to proceed in such a way that extreme errors, whether positive or negative, shall be confined within as narrow limits as possible.

Of all the principles which can be proposed for that purpose, I think there is none more general, more exact, and more easy of application, than that of which we have made use in the preceding researches, and which consists of rendering the sum of the squares of the errors a *minimum*. By this means there is established among the errors a sort of equilibrium which, preventing the extremes from exerting an undue influence, is very well fitted to reveal that state of the system which most nearly approaches the truth.



The design matrix at the Legendre's time.

CLASSICAL LINEAR REGRESSION

The sum of the squares of the errors $E^2 + E'^2 + E''^2 + \&c.$ being

$$\begin{aligned}&(a \ + bx \ + cy \ + fz \ + \&c.)^2\\ +\ &(a' \ + b'x \ + c'y \ + f'z \ + \&c.)^2\\ +\ &(a'' + b''x + c''y + f''z + \&c.)^2\\ +\ &\&c.,\end{aligned}$$

if its *minimum* is desired, when x alone varies, the resulting equation will be

$$o = \int ab + x\int b^2 + y\int bc + z\int bf + \&c.,$$

in which by $\int ab$ we understand the sum of similar products, i.e., $ab + a'b' + a''b'' + \&c.$; by $\int b^2$ the sum of the squares of the coefficients of x, namely $b^2 + b'^2 + b''^2 + \&c.$, and similarly for the other terms.

Similarly the *minimum* with respect to y will be

$$o = \int ac + x\int bc + y\int c^2 + z\int fc + \&c.,$$

and the *minimum* with respect to z,

$$o = \int af + x\int bf + y\int cf + z\int f^2 + \&c.,$$

in which it is apparent that the same coefficients $\int bc$, $\int bf$, &c. are common to two equations, a fact which facilitates the calculation.

LINEAR REGRESSION

(Statistical assumptions)

FIXED DESIGN: $\left.\begin{matrix} x_1 \in \mathbb{R}^d \\ \vdots \\ x_n \in \mathbb{R}^d \end{matrix}\right\} \rightarrow X = \begin{bmatrix} x_1^\top \\ \vdots \\ x_n^\top \end{bmatrix} \in \mathbb{R}^{n \times d}$ no errors in the x_i's

design points ↑ — design matrix ↑

$\left.\begin{matrix} y_1 \in \mathbb{R} \\ \vdots \\ y_n \in \mathbb{R} \end{matrix}\right\}$ realizations of random variables $\begin{matrix} Y_1 \\ \vdots \\ Y_n \end{matrix}$

$Y_i = \langle x_i, \theta^* \rangle + \varepsilon_i$

ε_i are i.i.d. random variables

$\varepsilon_i \sim N(0, \sigma^2)$

LINEAR REGRESSION

(Statistical assumptions)

FIXED DESIGN: $\left.\begin{matrix} x_1 \in \mathbb{R}^d \\ \vdots \\ x_n \in \mathbb{R}^d \end{matrix}\right\} \rightarrow X = \begin{bmatrix} x_1^\top \\ \vdots \\ x_n^\top \end{bmatrix} \in \mathbb{R}^{n \times d}$ no errors in the x_i's

design points ↑ design matrix ↑

$\left.\begin{matrix} y_1 \in \mathbb{R} \\ \vdots \\ y_n \in \mathbb{R} \end{matrix}\right\}$ realizations of random variables $\begin{matrix} Y_1 \\ \vdots \\ Y_n \end{matrix}$

$$Y = X\theta^* + \varepsilon$$
$$\varepsilon \sim N(0, \sigma^2 I_n)$$

LINEAR REGRESSION
(Least square estimator)

The least squares estimator $\hat{\theta}_{LS}$ is computed by solving

$$\min_{\theta \in \mathbb{R}^d} \sum_{i=1}^{n} (\langle \theta, x_i \rangle - y_i)^2 \iff \min_{\theta \in \mathbb{R}^d} \|X\theta - Y\|^2.$$

The (nonempty) solution set of this problem is obtained by the normal equations

$$X^\top (X\theta - Y) = 0 \iff X^\top X\theta = X^\top Y \quad .$$

LINEAR REGRESSION

(Least square estimator)

The least squares estimator $\hat{\theta}_{LS}$ is computed by solving

$$\min_{\theta \in \mathbb{R}^d} \sum_{i=1}^{n} (\langle \theta, x_i \rangle - y_i)^2 \iff \min_{\theta \in \mathbb{R}^d} \|X\theta - Y\|^2.$$

The (nonempty) solution set of this problem is obtained by the normal equations

$$X^\top (X\theta - Y) = 0 \iff X^\top X\theta = X^\top Y \iff X\theta = P_{R(X)} Y.$$

LINEAR REGRESSION
(Least square estimator)

The least squares estimator $\hat{\theta}_{LS}$ is computed by solving

$$\min_{\theta\in\mathbb{R}^d}\sum_{i=1}^{n}(\langle\theta,x_i\rangle-y_i)^2 \iff \min_{\theta\in\mathbb{R}^d}\|X\theta-Y\|^2.$$

The (nonempty) solution set of this problem is obtained by the normal equations

$$X^\top(X\theta-Y)=0 \iff X^\top X\theta=X^\top Y \iff X\theta=P_{R(X)}Y.$$

More precisely we set

$$\hat{\theta}_{LS}:=\operatorname*{argmin}_{X^\top X\theta=X^\top Y}\|\theta\|^2=X^\dagger Y=(X^\top X)^\dagger X^\top Y.$$

random vector

LINEAR REGRESSION
(Least square estimator)

$$\hat{\theta}_{LS} := X^{\dagger} Y$$

The Moore-Penrose pseudo-inverse of X is defined as follows:

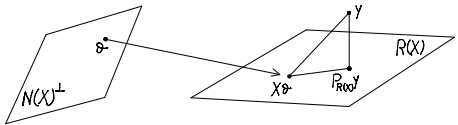
$X_{|N(X)^{\perp}} : N(X)^{\perp} \to R(X)$ is bijective

$P_{R(X)} : \mathbb{R}^n \to R(X)$ is the (orthogonal) projection onto the range of X

$$X^{\dagger} = (X_{|N(X)^{\perp}})^{-1} \circ P_{R(X)}.$$

The following properties hold:

- $X^{\dagger} = (X^{\top} X)^{\dagger} X^{\top}$
- $X X^{\dagger} = P_{R(X)}$



LINEAR REGRESSION
(Least square estimator)

$$\hat{\theta}_{LS} := X^{\dagger} Y$$

The statistical quantities of interest about a random vector $Z \in \mathbb{R}^n$.

$$\mathbb{E}[Z] = \begin{bmatrix} \mathbb{E}[Z_1] \\ \vdots \\ \mathbb{E}[Z_n] \end{bmatrix} \qquad \begin{aligned} \text{cov}[Z] &= \mathbb{E}[(Z - \mathbb{E}[Z])(Z - \mathbb{E}[Z])^{\top}] \\ (\text{cov}[Z])_{i,j} &= \mathbb{E}[(Z_i - \mathbb{E}[Z_i])(Z_j - \mathbb{E}[Z_j])] \end{aligned}$$

The following properties hold: If A is a matrix,

- $\mathbb{E}[AZ] = A\mathbb{E}[Z]$
- $\text{cov}[AZ] = A\text{cov}[Z]A^{\top}$.

LINEAR REGRESSION
(Statistical analysis of the least squares estimator)

$$\hat{\theta}_{LS} := X^{\dagger} Y$$

The quantities we really care of are

$$\hat{\mu}_{LS} := X\hat{\theta}_{LS} = \begin{bmatrix} \langle \hat{\theta}_{LS}, x_1 \rangle \\ \vdots \\ \langle \hat{\theta}_{LS}, x_n \rangle \end{bmatrix} \quad \text{and} \quad \mu^* := X\theta^* = \begin{bmatrix} \langle \theta^*, x_1 \rangle \\ \vdots \\ \langle \theta^*, x_n \rangle \end{bmatrix}$$

LINEAR REGRESSION
(Statistical analysis of the least squares estimator)

$$\hat{\theta}_{LS} := X^\dagger Y$$

The quantities we really care of are

$$\hat{\mu}_{LS} := X\hat{\theta}_{LS} = \begin{bmatrix} \langle \hat{\theta}_{LS}, x_1 \rangle \\ \vdots \\ \langle \hat{\theta}_{LS}, x_n \rangle \end{bmatrix} \quad \text{and} \quad \mu^* := X\theta^* = \begin{bmatrix} \langle \theta^*, x_1 \rangle \\ \vdots \\ \langle \theta^*, x_n \rangle \end{bmatrix}$$

Properties:

- $\mathbb{E}[\hat{\mu}_{LS}] = \mu^*$ (unbiased estimator)

$$\mathbb{E}[\hat{\mu}_{LS}] = \mathbb{E}[X\hat{\theta}_{LS}] = \mathbb{E}[XX^\dagger Y] = XX^\dagger \mathbb{E}[Y] = XX^\dagger X\theta^* = X\theta^*.$$

LINEAR REGRESSION
(Statistical analysis of the least squares estimator)

$$\hat{\theta}_{LS} := X^\dagger Y$$

The quantities we really care of are

$$\hat{\mu}_{LS} := X\hat{\theta}_{LS} = \begin{bmatrix} \langle \hat{\theta}_{LS}, x_1 \rangle \\ \vdots \\ \langle \hat{\theta}_{LS}, x_n \rangle \end{bmatrix} \quad \text{and} \quad \mu^* := X\theta^* = \begin{bmatrix} \langle \theta^*, x_1 \rangle \\ \vdots \\ \langle \theta^*, x_n \rangle \end{bmatrix}$$

Properties:

- $\mathbb{E}[\hat{\mu}_{LS}] = \mu^*$ (unbiased estimator)

$$\mathbb{E}[\hat{\mu}_{LS}] = \mathbb{E}[X\hat{\theta}_{LS}] = \mathbb{E}[XX^\dagger Y] = XX^\dagger \mathbb{E}[Y] = XX^\dagger X\theta^* = X\theta^*.$$

- $\text{cov}[\hat{\mu}_{LS}] = \sigma^2 XX^\dagger$

$$\text{cov}[\hat{\mu}_{LS}] = \text{cov}[XX^\dagger Y] = XX^\dagger \text{cov}[Y] XX^\dagger = XX^\dagger \sigma^2 I_n XX^\dagger = \sigma^2 XX^\dagger$$

LINEAR REGRESSION
(Statistical analysis of the least squares estimator)

The quality of the estimator $\hat{\mu}_{LS}$ si measured by the mean squared error

$$\mathrm{MSE}(\hat{\mu}_{LS}) = \frac{1}{n}\sum_{i=1}^{n}(\langle x_i, \hat{\theta}_{LS}\rangle - \langle x_i, \theta^*\rangle)^2 = \frac{1}{n}\|X\hat{\theta}_{LS} - X\theta^*\|^2 = \frac{1}{n}\|\hat{\mu}_{LS} - \mu^*\|^2$$

Theorem.
Let $r = \mathrm{rank}(X)$. Then

$$\mathbb{E}[\mathrm{MSE}(\hat{\mu}_{LS})] = \frac{1}{n}\mathbb{E}[\|\hat{\mu}_{LS} - \mu^*\|^2] = \frac{\sigma^2 r}{n}.$$

Remark.
If X is full rank, meaning $r = \min\{d, n\}$, we have

$$\mathbb{E}[\mathrm{MSE}(\hat{\mu}_{LS})] = \frac{\sigma^2 \min\{d, n\}}{n}$$

dimension of the space

number of measurements

Proof.
Recall that, in general, if $Z \in \mathbb{R}^n$ is a random vector, then

$$\begin{aligned} \mathrm{cov}[Z] &= \mathbb{E}[(Z - \mathbb{E}[Z])(Z - \mathbb{E}[Z])^\top] \\ (\mathrm{cov}[Z])_{i,j} &= \mathbb{E}[(Z_i - \mathbb{E}[Z_i])(Z_j - \mathbb{E}[Z_j])] \end{aligned}$$

Hence

$$\mathbb{E}[\|Z - \mathbb{E}[Z]\|^2] = \sum_{i=1}^{n} \mathbb{E}[(Z_i - \mathbb{E}[Z_i])^2] = \mathrm{Tr}(\mathrm{cov}[Z]).$$

Therefore, since $\hat{\mu}_{LS}$ is an unbiased estimator of μ^*, we have

$$\mathbb{E}[\mathrm{MSE}(\hat{\mu}_{LS})] = \frac{1}{n}\mathbb{E}[\|\hat{\mu}_{LS} - \mu^*\|^2] = \frac{1}{n}\mathrm{Tr}(\mathrm{cov}[\hat{\mu}_{LS}]) = \frac{\sigma^2}{n}\mathrm{Tr}(XX^\dagger) = \frac{\sigma^2 r}{n}$$

It continues on the blackboard →