Least Squares

In terms of linear algebra we want to solve an overdetermined system of equations (i.e sets of linear system of equations in which there are more equations than unknowns)

i.e $\underset{\sim}{A}\underset{\sim}{x} = \underset{\sim}{b}$ where $\underset{\sim}{A}$ having more rows than columns.

The idea of least squares solution is to find $\underset{\sim}{x}$ that minimizes 2-norm of the residual $\underset{\sim}{r} = \underset{\sim}{b} - \underset{\sim}{A}\underset{\sim}{x}$

<u>Example</u>:- Let us consider the data

x log(GDP)	y % urbanization
x_1	y_1
x_2	y_2
$\vdots$	
x_m	y_m

Say we want to fit a model

$$y = c_1 + c_2 x + c_3 x^2 + \cdots c_n x^n$$

$n < m$

It would be very nice to have

$$\left.\begin{aligned} c_1 + c_2 x_1 + c_3 x_1^2 + \cdots c_n x_1^n &= y_1 \\ c_1 + c_2 x_2 + c_3 x_2^2 + \cdots c_n x_2^n &= y_2 \\ &\vdots \\ c_1 + c_2 x_m + c_3 x_m^2 + \cdots c_n x_m^n &= y_m \end{aligned}\right\}$$

$$\underbrace{\begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^n \\ 1 & x_2 & x_2^2 & \cdots & x_2^n \\ 1 & x_3 & x_3^2 & \cdots & x_3^n \\ \vdots & & & & \\ 1 & x_m & x_m^2 & \cdots & x_m^n \end{bmatrix}}_{\underset{\sim}{A} \quad m \times n} \underbrace{\begin{bmatrix} c_1 \\ c_2 \\ c_3 \\ \vdots \\ c_n \end{bmatrix}}_{\underset{\sim}{x} \; n \times 1} = \underbrace{\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_m \end{bmatrix}}_{\underset{\sim}{b} \; m \times 1}$$

i.e we want to solve

$$\underset{\sim}{A}\,\underset{\sim}{x} = \underset{\sim}{b}$$

$$\underset{\sim}{A} \in \mathbb{R}^{m \times n}; \; \underset{\sim}{x} \in \mathbb{R}^{n \times 1}; \; \underset{\sim}{b} \in \mathbb{R}^{m \times 1}$$

$m > n$

$\underset{\sim}{A}$ is a full rank matrix

In general there is no solution to this problem unless $\underset{\sim}{b} \in \text{range}(\underset{\sim}{A})$

and this will be true for special choices of $\underset{\sim}{b}$

$$\underset{\sim}{A}\,\underset{\sim}{x} \approx \underset{\sim}{b}$$

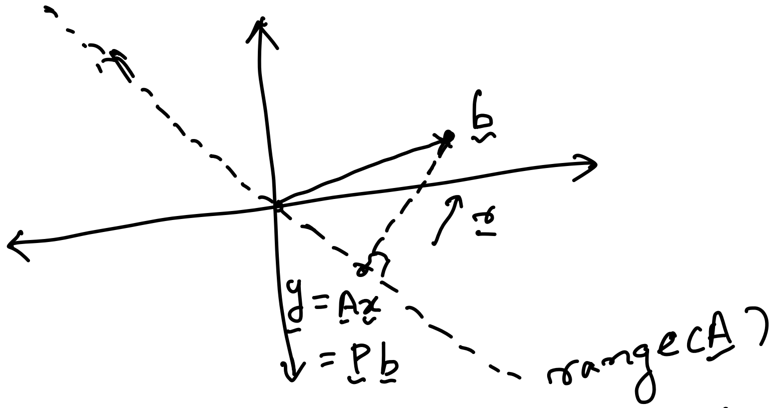
but can $\underset{\sim}{r} = \underset{\sim}{b} - \underset{\sim}{A}\,\underset{\sim}{x}$ be made smaller? Smallness of $\underset{\sim}{r}$ hints us to use a norm, and if we choose 2-norm, the problem becomes

$$\boxed{\begin{array}{l}\text{Given } \underset{\sim}{A} \in \mathbb{R}^{m\times n};\ m \geq n,\ b \in \mathbb{R}^{m} \\ \underset{\sim}{A} \text{ is a full rank matrix, then find} \\ \underset{\sim}{x} \in \mathbb{R}^{n} \text{ such that} \\ \|\underset{\sim}{b} - \underset{\sim}{A}\,\underset{\sim}{x}\|_2^2 \text{ is minimized}\end{array}}$$

$$\boxed{\min_{c_j} \sum_i \left(p(x_i) - y_i\right)^2}$$

The 2-norm corresponds to Euclidean distance and the geometric interpretation is that we want to find vector $\underset{\sim}{x}$ such the vector $\underset{\sim}{A}\,\underset{\sim}{x} \in \mathbb{R}^{m}$ in range $(\underset{\sim}{A})$ is closest to $\underset{\sim}{b}$

Orthogonal projection and normal equations!



→ Orthogonal projection will minimize norm $\underset{\sim}{r} = \underset{\sim}{b} - \underset{\sim}{A}\underset{\sim}{x}$ in the 2-norm.

→ That magical $\underset{\sim}{x}$ that minimizes $\|r\|_2$ satisfies $\underset{\sim}{A}\underset{\sim}{x} = \underset{\sim}{P}\underset{\sim}{b}$

where $\underset{\sim}{P} \in \mathbb{R}^{m \times m}$ is an orthogonal projector onto range(A)

i.e residual $\underset{\sim}{r}$ must be orthogonal to range(A))

Thm 1:- Let $\underset{\sim}{A} \in \mathbb{R}^{m \times n}$ $(m \geq n)$ and full-rank and $\underset{\sim}{b} \in \mathbb{R}^m$ be given. Then a vector

$\underset{\sim}{x} \in \mathbb{R}^n$ that minimizes $\|\underset{\sim}{r}\|_2 = \|\underset{\sim}{b} - \underset{\sim}{A}\underset{\sim}{x}\|$

(i.e $\underset{\sim}{x}$ is a least squares solution) if and only if

$\underset{\sim}{r} \perp \text{range}(\underset{\sim}{A})$

Remarks:- Let y be any vector in the range$(\underset{\sim}{A})$, then there exists a $\underset{\sim}{d} \in \mathbb{R}^n$ such that $\underset{\sim}{y} = \underset{\sim}{A}\underset{\sim}{d}$

Since $\underset{\sim}{r} \perp \text{range}(\underset{\sim}{A})$

$$\underset{\sim}{y}^T \underset{\sim}{r} = 0$$

$$\Rightarrow \underset{\sim}{d}^T \underset{\sim}{A}^T \underset{\sim}{r} = 0 \quad \forall\, \underset{\sim}{d} \in \mathbb{R}^n$$

$$\Rightarrow \underset{\sim}{A}^T \underset{\sim}{r} = 0$$

$$\Rightarrow \underset{\sim}{A}^T(\underset{\sim}{b} - \underset{\sim}{A}\underset{\sim}{x}) = 0$$

$$\Rightarrow \boxed{\underset{\sim}{A}^T \underset{\sim}{A} \underset{\sim}{x} = \underset{\sim}{A}^T \underset{\sim}{b}} \quad (*)$$

$$\underset{\sim}{x} = (\underset{\sim}{A}^T\underset{\sim}{A})^{-1}\underset{\sim}{A}^T\underset{\sim}{b}$$

$$\underset{\sim}{A}\underset{\sim}{x} = \underbrace{\underset{\sim}{A}(\underset{\sim}{A}^T\underset{\sim}{A})^{-1}\underset{\sim}{A}^T}\underset{\sim}{b}$$

$$= \underset{\sim}{P}\underset{\sim}{b}$$

where $\underset{\sim}{P} \in \mathbb{R}^{m\times m}$

orthogonal projector onto range $(\underset{\sim}{A})$

$$\Rightarrow \boxed{\underset{\sim}{A}^T \underset{\sim}{A} \underset{\sim}{x} = \underset{\sim}{A}^T b} \longrightarrow \text{Normal Equations}$$

$$(A^T A) \underset{\sim}{x} = \underset{\sim}{A}^T \underset{\sim}{b}$$

is $n \times n$ system of equations that has a unique solution if and only if A has full rank!

(*) When $\underset{\sim}{A}$ has a ~~f~~ full rank, the solution $\underset{\sim}{x}$ to the least squares problem is unique and formally can be written as

$$\underset{\sim}{x} = (\underset{\sim}{A}^T \underset{\sim}{A})^{-1} \underset{\sim}{A}^T \underset{\sim}{b}$$

This allows us to define pseudo-inverse of $\underset{\sim}{A}$ denote by $\underset{\sim}{A}^+ = (\underset{\sim}{A}^T \underset{\sim}{A})^{-1} \underset{\sim}{A}^T \in \mathbb{R}^{n \times m}$

$$\boxed{\underset{\sim}{A}^+ \underset{\sim}{A} = \underset{\sim}{I}} \qquad \boxed{\underset{\sim}{x} = \underset{\sim}{A}^+ \underset{\sim}{b}}$$

Algorithms to solve least squares:-

$$\underset{\sim}{A}^T \underset{\sim}{A} \underset{\sim}{x} = \underset{\sim}{A}^T \underset{\sim}{b} \longrightarrow \text{(Normal equations)}$$

$$\underset{\sim}{A} \underset{\sim}{x} = \underset{\sim}{P} \underset{\sim}{b} \longrightarrow \text{(least squares)}$$

(i) Cholesky Factorization:- $\underset{\sim}{A} \in \mathbb{R}^{m \times n}$ $m \geq n$

If $\underset{\sim}{A}$ has full rank, then $\underset{\sim}{A}^T \underset{\sim}{A}$ is square, symmetric and positive definite

Use cholesky factorization, which factors a symmetric positive definite matrix into the form $\underset{\sim}{R}^T \underset{\sim}{R}$ where $\underset{\sim}{R}$ is upper triangular

$$\underset{\sim}{A}^T \underset{\sim}{A} = \underset{\sim}{R}^T \underset{\sim}{R}$$

$$\underset{\sim}{A}^T \underset{\sim}{A} \underset{\sim}{x} = \underset{\sim}{A}^T \underset{\sim}{b}$$

$$\underset{\sim}{R}^T \underset{\sim}{R} \underset{\sim}{x} = \underset{\sim}{A}^T \underset{\sim}{b}$$

Algo:- (1) Form $\underset{\sim}{A}^T \underset{\sim}{A}$ and $\underset{\sim}{A}^T \underset{\sim}{b}$

(2) cholesky factorization of $\underset{\sim}{A}^T \underset{\sim}{A}$

$$\underset{\sim}{A}^T \underset{\sim}{A} = \underset{\sim}{R}^T \underset{\sim}{R}$$

to obtain $\underset{\sim}{R}^T \underset{\sim}{R} \underset{\sim}{x} = \underset{\sim}{A}^T \underset{\sim}{b}$

(3) Solve Lower triangular system

$$\underset{\sim}{R}^T \underset{\sim}{w} = \underset{\sim}{A}^T \underset{\sim}{b} \text{ for } \underset{\sim}{w}$$

$$(\underset{\sim}{w} = R\underset{\sim}{x})$$

(4) Solve upper triangular system

$$\underset{\sim}{R}\underset{\sim}{x} = \underset{\sim}{w} \text{ for } \underset{\sim}{x}$$

Work $\sim mn^2 + \frac{1}{3}n^3$ flops

(ii) -via- $\underline{QR}$ factorization

$\underset{\sim}{A} = \hat{Q}\hat{R}$ obtain Householder triangularization

$$\underset{\sim}{A}\underset{\sim}{x} = \underline{P}\underset{\sim}{b}$$

$\underset{\sim}{P} = \hat{Q}\hat{Q}^T$ is orthogonal projector onto range(A)

$$\underset{\sim}{A}\underset{\sim}{x} = \underline{P}\underset{\sim}{b}$$

$$\Rightarrow \hat{\underset{\sim}{Q}}\hat{\underline{R}}\underset{\sim}{x} = \hat{Q}\hat{Q}^T\underset{\sim}{b}$$

$$\Rightarrow \hat{\underset{\sim}{R}}\underset{\sim}{x} = \hat{\underline{Q}}^T\underline{b} \quad - (\#)$$

Algo :- (1) Compute reduced QR factorization

$$\underset{\sim}{A} = \hat{Q}\hat{R}$$

(2) Compute vector $\hat{\underline{Q}}^T\underset{\sim}{b}$

(3) Solve upper triangular system
$$\hat{R}x = \hat{Q}^T b \text{ for } x$$

Work $\sim 2mn^2 - \frac{2}{3}n^3$ flops

Least squares using SVD:-

We use reduced SVD
$$\underset{m\times n}{A} = \underset{m\times n}{\hat{U}}\hat{\Sigma}V^T$$ to least squares problem

$$Ax = Pb \quad \text{projector } P = \hat{U}\hat{U}^T$$

$$Ax = \hat{U}\hat{U}^T b$$

$$\Rightarrow \hat{U}\hat{\Sigma}V^T x = \hat{U}\hat{U}^T b$$

$$\Rightarrow \underbrace{\hat{\Sigma}V^T x = \hat{U}^T b}$$

Algo:- (1) Compute reduced SVD $A = \hat{U}\hat{\Sigma}V^T$

(2) Compute $\hat{U}^T b$

(3) Solve diagonal system
$$\hat{\Sigma} w = \hat{U}^T b$$
for $w = V^T x$

(4) Solve $\underset{\sim}{V}^T \underset{\sim}{x} = w$ for $\underset{\sim}{x}$

Work $\sim 2mn^2 + 11n^3$

Cholesky $\sim mn^2 + \frac{1}{3}n^2$ flops

QR factorization $\sim \underline{2mn^2 - \frac{2}{3}n^3}$ flops

SVD $\sim \underline{2mn^2 + 11n^3}$

Comparison of Algos:-

(1) Solving least squares by cholesky is cheapest roughly by a factor of 2. However algorithm may not be stable in presence of rounding errors

(2) QR is cheaper than SVD when $m \gg n$. $m \approx n$ QR and SVD has similar costs.

SVD is a method of choice when $\underset{\sim}{A}$ is close to rank deficiency and QR is a method of choice when $\underset{\sim}{A}$ is not too close to rank deficiency.

Rank Deficient matrices

$m \geq n$
rank$(A) < n$

Use reduced SVD

$$\boxed{\underset{\sim}{A} = \underset{\sim}{U}_1 \underset{\sim}{\Sigma}_1 \underset{\sim}{V}_1^T}$$

$\underset{\sim}{A}\underset{\sim}{x} \simeq \underset{\sim}{b}$

$\underset{\sim}{U}_1 \in \mathbb{R}^{m \times r}$

$\underset{\sim}{\Sigma}_1 \in \mathbb{R}^{r \times r}$

$\underset{\sim}{V}_1 \in \mathbb{R}^{n \times r}$

then

$$x_{LS} = \underset{\sim}{V}_1 \underset{\sim}{\Sigma}_1^{-1} \underset{\sim}{U}_1^T \underset{\sim}{b}$$

is the minimum ℓ^2 norm minimizer $\|\underset{\sim}{A}\underset{\sim}{x} - \underset{\sim}{b}\|_2$

$x_{LS} = \underset{\sim}{A}^+ \underset{\sim}{b}$

Note:- If $\underset{\sim}{x}$ minimizes $\|\underset{\sim}{A}\underset{\sim}{x} - \underset{\sim}{b}\|_2$

then all vectors $\underset{\sim}{A}\underset{\sim}{z} = 0$ will minimize $\|\underset{\sim}{A}\underset{\sim}{x} - \underset{\sim}{b}\|_2$

and hence we seek a solution $\underset{\sim}{x}$ which has a minimum norm $\|x\|_2$ and

is given as above!

Pseudo-inverse:-

$A \in \mathbb{R}^{m \times n}$ a pseudoinverse of A is defined as a matrix $A^{+} \in \mathbb{R}^{n \times m}$ satisfying the criteria :-

1: $\underbrace{A A^{+}}A = A$ In general AA^{+} need not be identity matrix but maps all columnvectors of A to themselves

2. $\underbrace{A^{+}}AA^{+} = A^{+}$

3. $(AA^{+})^{T} = AA^{+}$ i.e AA^{+} is symmetric

4. $(A^{+}A)^{T} = A^{+}A$ i.e $A^{+}A$ is symmetric

Remarks :-

A^{+} exists for any matrix but

A is full rank $m > n$

$$A^{+} = \underbrace{(A^{T}A)^{-1}A^{T}}_{\text{left inverse}} \text{ is called}$$

$A^{+}A = I$

$\underset{\sim}{A}$ is full rank $m < n$

$$A^{+} = \underbrace{A^{T}(AA^{T})^{-1}}_{\text{right inverse}} \qquad \underset{\sim}{A}\underset{\sim}{A}^{+} = \underset{=}{I}$$

Pseudoinverse exists for any matrix $\underset{\sim}{A}$ and is unique.

In the case of rank deficient matrices we cannot use the above algebraic forms as described above but SVD offers a way to construct A^{+}

$$\underset{\sim}{A} = \underset{\sim}{U_1}\underset{\sim}{\Sigma_1}\underset{\sim}{V_1}^{T} \qquad \underset{\sim}{U_1} \in \mathbb{R}^{m\times r} \qquad \underset{\sim}{\Sigma_1} \in \mathbb{R}^{r\times r}$$

$$\underset{\sim}{V_1} \in \mathbb{R}^{n\times r}$$

$$\text{rank}(\underset{\sim}{A}) = r$$

and $\boxed{\underset{\sim}{A}^{+} = \underset{\sim}{V_1}\underset{\sim}{\Sigma_1}^{-1}\underset{\sim}{U_1}^{T}}$