Least Square Problem

Consider observation data $b_1, b_2, \ldots, b_m \in \mathbb{R}$ of some quantities, e.g., temperature, test scores, Dow Jones index, accidents, you name it. Meanwhile, for each observation b_i, we pair with some independent data $a_{i1}, a_{i2}, \ldots, a_{in}$, such as, humidity, age, unemployment rate, eyesight, etc.. An example: we observe that at 32 Fahrenheit (b_i, dependent variable), local humidity is 5%, local wind speed is 32 mph, and local pressure is 1 atm. In fact, we may draw up data in a table.

Temperature in F $\boldsymbol{b}$	Humidity $\boldsymbol{a}_1$	Wind Speed $\boldsymbol{a}_2$	Pressure $\boldsymbol{a}_3$	Air Quality Index$\boldsymbol{a}_4$
32	5	17	1	55
37	7	13	1.02	65
44	8	11	0.99	14
47	11	6	0.96	36

We seek the answer to the question: how does temperature depend on all these variables with the provided data? It is completely natural to posit a linear relationship between $\boldsymbol{b}$ and the $\boldsymbol{a}_i$'s. More precisely, for each b_i, we seek the coefficients $x_1, \ldots, x_n$ that

$$b_i = x_0 + x_1 a_{i1} + x_2 a_{i2} + \cdots + x_n a_n, \quad i = 1, 2, \ldots, m.$$

Is it utterly possible that we can find the exact $x_{i1}, \ldots, x_{in}$ that satisfy this relationship? We probably need n equations at least to determine these unknowns. So,

$$\begin{aligned} b_1 &= x_0 + x_1 a_{11} + \cdots + x_n a_{1n} = (1, a_{11}, \ldots, a_{1n}) \cdot (x_0, x_1, \ldots, x_n) \\ &.. \\ &.. \\ b_m &= x_0 + x_1 a_{m1} + \cdots + x_n a_{mn} = (1, a_{m1}, \ldots, a_{mn}) \cdot (x_0, x_1, \ldots, x_n) \end{aligned}$$

or more compactly,

$$\boldsymbol{b} = \begin{bmatrix} 1 & a_{11} & a_{12} & . & . & a_{1n} \\ 1 & a_{21} & . & . & . & . \\ 1 & . & . & . & . & . \\ 1 & . & . & . & . & . \\ 1 & . & . & . & . & . \\ 1 & a_{m1} & . & . & . & a_{mn} \end{bmatrix}_{m \times (n+1)} \begin{bmatrix} x_0 \\ x_1 \\ . \\ . \\ . \\ x_n \end{bmatrix}_{(n+1) \times 1} = \boldsymbol{A}\boldsymbol{x}.$$

So, we have $(n+1)$ unknowns but only m equations. This is only possibly exactly solvable when $m = n+1$ which means we have exactly the same number of observation data points as the number of independent variables. This is unrealistic. In practice, $m \gg n+1$, that is, we have massive amount of data, but only a handful of features we seek the extent of dependence on. Therefore, the system $\boldsymbol{Ax} = \boldsymbol{b}$ here is not solvable in general.

Then what? Game over? We go for the next best thing. Now, if $\boldsymbol{Ax} = \boldsymbol{b}$ is not solvable, we may find some $\boldsymbol{x}$ that achieves $\boldsymbol{Ax} \approx \boldsymbol{b}$, which is a reasonable request. In fact, we seek a solution that minimizes the square of the l^2-error

$$\phi(\boldsymbol{x}) = \|\boldsymbol{b} - \boldsymbol{Ax}\|_2^2$$

where $\phi : \mathbb{R}^{n+1} \to \mathbb{R}$. Recall that $\boldsymbol{b} - \boldsymbol{Ax}$ is known as the residual vector. All we are trying to do is to come up with a solution that minimizes the l^2-norm of this residual vector.

We expand the l^2-norm by definition,

$$\phi(\boldsymbol{x}) = \phi(x_0, x_1, \ldots, x_n) = \sum_{i=1}^{m} (b_i - x_0 - x_1 a_{i1} - x_2 a_{i2} - \cdots - x_n a_{in})^2.$$

How do we minimize a function of multiple variables? We compute its gradient and set it equal to $\boldsymbol{0}$ to find the critical points.

$$0 = \frac{\partial \phi}{\partial x_j} = -2 \sum_{i=1}^{m} (b_i - x_0 - x_1 a_{i1} - x_2 a_{i2} - \cdots - x_n a_{in}) a_{ij}, \quad j = 0, 1, \ldots, n.$$

Moving the -2 out of the way, we see that the critical point(s) satisfy

$$\sum_{i=1}^{m} \left(b_i - x_0 - x_1 a_{i1} - x_2 a_{i2} - \cdots - x_n a_{in}\right) a_{ij} = 0, \quad j = 0, 1, \ldots, n.$$

Guess what? Now, we have exactly $n+1$ equations for the $n+1$ unknowns. This set of equations is called the **normal equations**.

Denote $y_i = b_i - x_0 - x_1 a_{i1} - x_2 a_{i2} - \cdots - x_n a_{in}$ and $\boldsymbol{y} = (y_1, y_2, \ldots, y_m)^{\mathrm{T}}$. Then, the equation reads

$$\sum_{i=1}^{m} a_{ij} y_i = 0, \quad j = 0, 1, \ldots, n.$$

which now requires you to recall the definition of matrix vector multiplication –

$$(\boldsymbol{A}\boldsymbol{x})_i = (a_{i1}, a_{i2}, \ldots, a_{in}) \cdot (x_1, x_2 \ldots, x_n) \implies i^{th} \text{ row of } \boldsymbol{A} \text{ dotted with } \boldsymbol{x}.$$

Let's visualize what $\sum_{i=1}^{m} a_{ij} y_i$ really is:

$$(a_{1j}, a_{2j}, a_{3j}, \ldots, a_{mj}) \cdot (y_1, \ldots, y_m)$$

where we realize that $(a_{1j}, a_{2j}, a_{3j}, \ldots, a_{mj})$ is the j^{th} **column** of $\boldsymbol{A}$, which means it is the j^{th} **row** of $\boldsymbol{A}^{\mathrm{T}}$. Thus,

$$\sum_{i=1}^{m} a_{ij} y_i = \left(\boldsymbol{A}^{\mathrm{T}} \boldsymbol{y}\right)_j .$$

Enumerating over all $j = 1, 2, \ldots, m$, we find that the **normal equations** can be written in matrix form,

$$\boldsymbol{A}^{\mathrm{T}} \boldsymbol{y} = 0.$$

Now, looking at the definition of $\boldsymbol{y}$, we have

$$\begin{aligned} y_1 &= b_1 - x_0 - x_1 a_{11} - x_2 a_{12} - \cdots - x_n a_{1n} \\ &. \\ &. \\ y_m &= b_m - x_0 - x_1 a_{m1} - x_2 a_{m2} - \cdots - x_n a_{mn} \end{aligned}$$

which is

$$\boldsymbol{y} = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}.$$

Inserting this back into the normal equation $\boldsymbol{A}^{\mathrm{T}} \boldsymbol{y} = 0$, we have

$$\boldsymbol{A}^{\mathrm{T}} (\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}) = 0 \implies \boldsymbol{A}^{\mathrm{T}} \boldsymbol{A} \boldsymbol{x} = \boldsymbol{A}^{T} \boldsymbol{b},$$

the celebrated final form of the **normal equations**. All we need is the observation data: the dependent variable $\boldsymbol{b}$, and the independent variables $\boldsymbol{A}$.

We put the problem in full form: the (minimizer) solution to the **least square problem**

$$\widetilde{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}} \|\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}\|_2^2$$

is the solution to the set of **normal equations**

$$\boldsymbol{A}^{\mathrm{T}} \boldsymbol{A} \widetilde{\boldsymbol{x}} = \boldsymbol{A}^{T} \boldsymbol{b}.$$

Now, after finding where the critical point is, we still need to confirm that this critical point indeed gives me the minimum, not the maximum.

Theorem. *Let $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$. Every solution $\widetilde{\boldsymbol{x}}$ to $\boldsymbol{A}^{\mathrm{T}} \boldsymbol{A} \widetilde{\boldsymbol{x}} = \boldsymbol{A}^{T} \boldsymbol{b}$ satisfies*

$$\|\boldsymbol{b} - \boldsymbol{A}\widetilde{\boldsymbol{x}}\|_2 \leq \|\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}\|_2 \quad \forall \boldsymbol{x} \in \mathbb{R}^n,$$

that is, $\widetilde{\boldsymbol{x}}$, if exists, is the global minimizer of $\|\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}\|_2$.

Proof. Given $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^m$, we have

$$\|\boldsymbol{u}+\boldsymbol{v}\|_2^2 = (\boldsymbol{u}+\boldsymbol{v})^{\mathrm{T}}(\boldsymbol{u}+\boldsymbol{v}) = \|\boldsymbol{u}\|_2^2 + 2\boldsymbol{u}^{\mathrm{T}}\boldsymbol{v} + \|\boldsymbol{v}\|_2^2 .$$

Then,

$$\begin{aligned}
\|\boldsymbol{b}-\boldsymbol{A}\boldsymbol{x}\|_2^2 &= \|\boldsymbol{b}-\boldsymbol{A}\widetilde{\boldsymbol{x}}+\boldsymbol{A}\widetilde{\boldsymbol{x}}-\boldsymbol{A}\boldsymbol{x}\|_2^2 \\
&= \|\boldsymbol{b}-\boldsymbol{A}\widetilde{\boldsymbol{x}}\|_2^2 + 2\left(\boldsymbol{A}\left(\widetilde{\boldsymbol{x}}-\boldsymbol{x}\right)\right)^{\mathrm{T}}\left(\boldsymbol{b}-\boldsymbol{A}\widetilde{\boldsymbol{x}}\right) + \|\boldsymbol{A}\left(\widetilde{\boldsymbol{x}}-\boldsymbol{x}\right)\|_2^2 \\
&\geq \|\boldsymbol{b}-\boldsymbol{A}\widetilde{\boldsymbol{x}}\|_2^2 + 2\left(\widetilde{\boldsymbol{x}}-\boldsymbol{x}\right)^{\mathrm{T}} \cancelto{0}{\boldsymbol{A}^{\mathrm{T}}\left(\boldsymbol{b}-\boldsymbol{A}\widetilde{\boldsymbol{x}}\right)} \\
&= \|\boldsymbol{b}-\boldsymbol{A}\widetilde{\boldsymbol{x}}\|_2^2 .
\end{aligned}$$

□

Existence of a Solution

It remains to show that $\widetilde{\boldsymbol{x}}$ indeed exists, and under one more condition on $\boldsymbol{A}$, is also unique. Existence is not hard if we know a little bit of linear algebra. Note that $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{b}$ lies in the range of $\boldsymbol{A}^{\mathrm{T}}$. But we also can show that the range of $\boldsymbol{A}^{\mathrm{T}}$ and that of $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}$ are the same (a fundamental theorem in linear algebra), which means there exists $\boldsymbol{x}$ such that $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x} = \boldsymbol{A}^{\mathrm{T}}\boldsymbol{b}$ since both sides of the equation maps to the same subspace.

Uniqueness of the Solution

If $\det\left(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\right) \neq 0$, then we are all set because then $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}$ is invertible, and

$$\widetilde{\boldsymbol{x}} = \left(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\right)^{-1}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{b}.$$

But is this always the case? This should depend on $\boldsymbol{A}$ – but here $\boldsymbol{A}$ is not necessarily square. So the usual technique from linear algebra won't work.

The claim here is that $\boldsymbol{A}$ must have linearly independent columns iff $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}$ is invertible. For the forward direction, assume that $\boldsymbol{A}$ has linearly independent columns, we suppose, on the contrary, that $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}$ is not invertible. Then, $\det\left(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\right) = 0$, which means there exists nonzero $\boldsymbol{z} = \mathbf{0}$ such that

$$\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{z} = 0.$$

Now, multiplying $\boldsymbol{z}^{\mathrm{T}}$ on the left, we have

$$\boldsymbol{z}^{\mathrm{T}}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{z} = 0 \implies (\boldsymbol{A}\boldsymbol{z})^{\mathrm{T}}(\boldsymbol{A}\boldsymbol{z}) = 0 \implies \|\boldsymbol{A}\boldsymbol{z}\|_2^2 = 0 \implies \boldsymbol{A}\boldsymbol{z} = 0, \quad \text{where } \boldsymbol{z} \neq \mathbf{0}.$$

But this is impossible because $\boldsymbol{A}$ has linearly independent columns, i.e., the only solution to $\boldsymbol{A}\boldsymbol{z} = \mathbf{0}$ is $\boldsymbol{z} = \mathbf{0}$. Contradiction!