Deriving Least Square In Details

"Least Square Method to Statistics is Calculus to Mathematics."

The method of least squares is a standard approach in regression analysis to approximate the solution of overdetermined systems, i.e., sets of equations in which there are more equations than unknowns. "Least squares" means that the overall solution minimizes the sum of the squares of the residuals made in the results of every single equation.

The most important application is in data fitting. The best fit in the least-squares sense minimizes the sum of squared residuals (a residual being: the difference between an observed value, and the fitted value provided by a model). When the problem has substantial uncertainties in the independent variable (the x variable), then simple regression and least-squares methods have problems; in such cases, the methodology required for fitting errors-in-variables models may be considered instead of that for least squares.

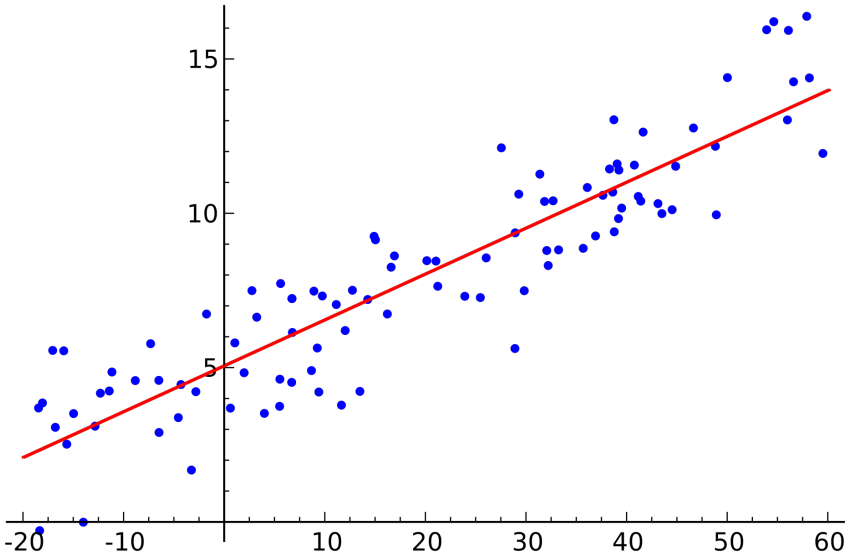

Fig. 1: A red line is to approximate all the data points such that the squared loss is as small as possible. In fact, there are 2 parameters to determine in this case – intercept and slope, where intercept here is 5.

Assume we can use a linear model to fit all the data points $\{x_i, y_i\}_{i=1,\ldots,n}$:

$$\tilde{y} = \beta_0 + \beta_1 x \tag{1}$$

where $\tilde{y}$ is our prediction, with the loss function being:

$$\mathcal{J} = \sum_{i=1}^{n} (y_i - \tilde{y}_i)^2 \tag{2}$$

our goal is to minimize $\mathcal{J}$, such that our prediction is as close to the ground-truth as possible. Now we assume that there are several factors x_i will influence the outcome ***independently***, then we can formulate the objective as:

$$\tilde{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \tag{3}$$

where our target is:

$$\min \mathcal{J} \ = \ \sum_{i=1}^{n} (y_i - \tilde{y}_i)^2 \tag{4}$$

Assume we can approximate by multiple linear regression by:

$$\begin{aligned} y_1 &\approx \beta_0 + \beta_1 X_{11} + \beta_2 X_{12} + \cdots + \beta_p X_{1p} := \hat{y}_1 \\ y_2 &\approx \beta_0 + \beta_1 X_{21} + \beta_2 X_{22} + \cdots + \beta_p X_{2p} := \hat{y}_2 \\ &\cdots \\ y_n &\approx \beta_0 + \beta_1 X_{n1} + \beta_2 X_{n2} + \cdots + \beta_p X_{np} := \hat{y}_n \end{aligned} \tag{5}$$

where $\boldsymbol{X}$ and $\boldsymbol{y}$ is known, and $\boldsymbol{\beta}$ is the coefficients to optimize.
Now let's define:

$$\begin{aligned} \boldsymbol{A} &= \begin{bmatrix} 1 & X_{11} & X_{12} & \dots & X_{1p} \\ 1 & X_{21} & X_{22} & \dots & X_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & X_{n2} & \dots & X_{np} \end{bmatrix} \\ \boldsymbol{y} &= [y_1, y_2, \dots, y_n]^T \\ \boldsymbol{\beta} &= [\beta_0, \beta_1, \beta_2, \dots, \beta_p]^T \end{aligned} \tag{6}$$

then we can formulate our objective function as:

$$\boldsymbol{J} = \min_{\boldsymbol{\beta}} \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{\beta}\|_2^2 \tag{7}$$

A few Lemmas are now in order to give:

- $\|\boldsymbol{z}\|^2 = trace(\boldsymbol{z}^T\boldsymbol{z}) = \langle \boldsymbol{z}, \boldsymbol{z} \rangle$, when we denote $\langle\ \rangle$ as the sum of inner product between two factors. For example, assume

$$\boldsymbol{z} = \begin{bmatrix} z_{11} & z_{12} & z_{13} \\ z_{21} & z_{22} & z_{23} \\ z_{31} & z_{32} & z_{33} \end{bmatrix}$$

  then we have $trace(\boldsymbol{z}^T\boldsymbol{z}) = z_{11}^2 + z_{12}^2 + z_{13}^2 + z_{21}^2 + z_{22}^2 + z_{23}^2 + z_{31}^2 + z_{32}^2 + z_{33}^2 = \|\boldsymbol{z}\|^2 = \langle \boldsymbol{z}, \boldsymbol{z} \rangle$
- $trace(\boldsymbol{a}^T\boldsymbol{b}) = \langle \boldsymbol{a}, \boldsymbol{b} \rangle = a_{11}b_{11} + a_{12}b_{12} + \cdots + a_{32}b_{32} + a_{33}b_{33}$
- $\frac{\partial trace(\boldsymbol{a}^T\boldsymbol{b})}{\partial \boldsymbol{a}} = \frac{\partial \langle \boldsymbol{a}, \boldsymbol{b} \rangle}{\partial \boldsymbol{a}}$ that is:

$$\begin{bmatrix} \frac{\partial \langle \boldsymbol{a}, \boldsymbol{b} \rangle}{\partial a_{11}} & \frac{\partial \langle \boldsymbol{a}, \boldsymbol{b} \rangle}{\partial a_{12}} & \frac{\partial \langle \boldsymbol{a}, \boldsymbol{b} \rangle}{\partial a_{13}} \\ \frac{\partial \langle \boldsymbol{a}, \boldsymbol{b} \rangle}{\partial a_{21}} & \frac{\partial \langle \boldsymbol{a}, \boldsymbol{b} \rangle}{\partial a_{22}} & \frac{\partial \langle \boldsymbol{a}, \boldsymbol{b} \rangle}{\partial a_{23}} \\ \frac{\partial \langle \boldsymbol{a}, \boldsymbol{b} \rangle}{\partial a_{31}} & \frac{\partial \langle \boldsymbol{a}, \boldsymbol{b} \rangle}{\partial a_{32}} & \frac{\partial \langle \boldsymbol{a}, \boldsymbol{b} \rangle}{\partial a_{33}} \end{bmatrix} = \begin{bmatrix} \frac{\partial a_{11}b_{11} + a_{12}b_{12} + \cdots + a_{32}b_{32} + a_{33}b_{33}}{\partial a_{11}} & \cdots & \frac{\partial a_{11}b_{11} + a_{12}b_{12} + \cdots + a_{32}b_{32} + a_{33}b_{33}}{\partial a_{13}} \\ \frac{\partial a_{11}b_{11} + a_{12}b_{12} + \cdots + a_{32}b_{32} + a_{33}b_{33}}{\partial a_{21}} & \cdots & \frac{\partial a_{11}b_{11} + a_{12}b_{12} + \cdots + a_{32}b_{32} + a_{33}b_{33}}{\partial a_{23}} \\ \frac{\partial a_{11}b_{11} + a_{12}b_{12} + \cdots + a_{32}b_{32} + a_{33}b_{33}}{\partial a_{31}} & \cdots & \frac{\partial a_{11}b_{11} + a_{12}b_{12} + \cdots + a_{32}b_{32} + a_{33}b_{33}}{\partial a_{33}} \end{bmatrix} = \begin{bmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \\ b_{31} & b_{32} & b_{33} \end{bmatrix} = \boldsymbol{b} \tag{8}$$

– $trace(\boldsymbol{ab}) = trace(\boldsymbol{ba})$. Assume that

$$\boldsymbol{b} = \begin{bmatrix} b_{11}\ b_{12}\ b_{13} \\ b_{21}\ b_{22}\ b_{23} \end{bmatrix}$$

$$\boldsymbol{a} = \begin{bmatrix} a_{11}\ a_{12} \\ a_{21}\ a_{22} \\ a_{31}\ a_{32} \end{bmatrix}$$

then by definition we have: $trace(\boldsymbol{ab}) = a_{11}b_{11} + a_{12}b_{21} + a_{21}b_{12} + a_{22}b_{22} + a_{31}b_{13} + a_{32}b_{23}$, while $trace(\boldsymbol{ba}) = b_{11}a_{11} + b_{12}a_{21} + b_{13}a_{31} + b_{21}a_{12} + b_{22}a_{22} + b_{23}a_{32}$, which is equivalent.

– $trace(\boldsymbol{a}) = trace(\boldsymbol{a}^T)$

Now let's turn to Eq. (7):

$$\begin{aligned} \boldsymbol{J} &= \|\boldsymbol{y} - \boldsymbol{A\beta}\|_2^2 \\ &= trace[(\boldsymbol{y} - \boldsymbol{A\beta})^T(\boldsymbol{y} - \boldsymbol{A\beta})] \\ &= trace(\boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{\beta}^T\boldsymbol{A}^T\boldsymbol{y} - \boldsymbol{y}^T\boldsymbol{A\beta} + \boldsymbol{\beta}^T\boldsymbol{A}^T\boldsymbol{A\beta}) \end{aligned} \tag{9}$$

Recall that a convex function will achieve the minimum *iff* the derivative is *zero*, then we have:

$$\frac{\partial \boldsymbol{J}}{\partial \boldsymbol{\beta}} = \boldsymbol{0}$$

which is equivalent to:

$$\begin{aligned} \frac{\partial \boldsymbol{J}}{\partial \boldsymbol{\beta}} &= \frac{\partial\ trace(\boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{\beta}^T\boldsymbol{A}^T\boldsymbol{y} - \boldsymbol{y}^T\boldsymbol{A\beta} + \boldsymbol{\beta}^T\boldsymbol{A}^T\boldsymbol{A\beta})}{\partial \boldsymbol{\beta}} \\ &= 2(\boldsymbol{A}^T\boldsymbol{A\beta} - \boldsymbol{A}^T\boldsymbol{y}) \\ &= \boldsymbol{0} \end{aligned} \tag{10}$$

then we have:

$$\boldsymbol{\beta}^* = (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{y}$$

It follows from the following:

$$trace(\boldsymbol{\beta}^T\boldsymbol{A}^T\boldsymbol{y}) = trace(\boldsymbol{y}^T\boldsymbol{A\beta})$$

by making use of Lemma 5, while:

$$\frac{\partial\ trace(\boldsymbol{\beta}^T\boldsymbol{A}^T\boldsymbol{y})}{\partial \boldsymbol{\beta}} = \boldsymbol{A}^T\boldsymbol{y} \tag{11}$$

by making use of Lemma 3. However, the following one is a lit-bit tricky since we have two $\boldsymbol{\beta}$, thus we could first fix the second one and take the derivative with respect to the first, vice versa.

$$\begin{aligned} \frac{\partial\ trace(\boldsymbol{\beta}^T\boldsymbol{A}^T\boldsymbol{A\beta})}{\partial \boldsymbol{\beta}} &= \frac{\partial\ trace[\boldsymbol{\beta}^T(\boldsymbol{A}^T\boldsymbol{A\beta})]}{\partial \boldsymbol{\beta}} + \frac{\partial\ trace[(\boldsymbol{\beta}^T\boldsymbol{A}^T\boldsymbol{A})\boldsymbol{\beta}]}{\partial \boldsymbol{\beta}} \\ &= \frac{\partial\ trace[\boldsymbol{\beta}^T(\boldsymbol{A}^T\boldsymbol{A\beta})]}{\partial \boldsymbol{\beta}} + \frac{\partial\ trace\{[(\boldsymbol{\beta}^T\boldsymbol{A}^T\boldsymbol{A})\boldsymbol{\beta}]^T\}}{\partial \boldsymbol{\beta}} \\ &= 2\frac{\partial\ trace[\boldsymbol{\beta}^T(\boldsymbol{A}^T\boldsymbol{A\beta})]}{\partial \boldsymbol{\beta}} \\ &= 2\boldsymbol{A}^T\boldsymbol{A\beta} \end{aligned} \tag{12}$$