V-3-4. Optimization and pseudo-inverse matrix

In multiple regression analysis, we estimate optimal coefficient of variables to explain explained variables from data sets. When we express explained variable as $\boldsymbol{y}$ and explanatory variables as $\boldsymbol{x}$, example of data set is as follow.

$$\begin{gathered} y_1, x_{11}, x_{21} \cdots x_{p1} \\ y_2, x_{12}, x_{22} \cdots x_{p2} \\ y_3, x_{11}, x_{23} \cdots x_{p3} \\ \vdots \\ y_m, x_{1m}, x_{2m} \cdots x_{pm} \end{gathered}$$

We are expecting to write the relation as follow.

$$\begin{gathered} y_1 = a_1 x_{11} + a_2 x_{21} + \cdots + a_p x_{p1} \\ y_2 = a_1 x_{12} + a_2 x_{22} + \cdots + a_p x_{p2} \\ y_3 = a_1 x_{13} + a_2 x_{23} + \cdots + a_p x_{p3} \\ \vdots \\ y_m = a_1 x_{1m} + a_2 x_{2m} + \cdots + a_p x_{pm} \end{gathered}$$

However, it is generally impossible because explained variable data includes impacts of unknown variables, so we have to add error term as follow.

$$\begin{gathered} y_1 = a_1 x_{11} + a_2 x_{21} + \cdots + a_p x_{p1} + e_1 \\ y_2 = a_1 x_{12} + a_2 x_{22} + \cdots + a_p x_{p2} + e_2 \\ y_3 = a_1 x_{13} + a_2 x_{23} + \cdots + a_p x_{p3} + e_3 \\ \vdots \\ y_m = a_1 x_{1m} + a_2 x_{2m} + \cdots + a_p x_{pm} + e_m \end{gathered}$$

Multiple regression analysis is optimization of error term. There several ideas of the optimization, Least square method is one of them. The idea of least square method is minimizing of error term.

When we express the relation using matrix.

$$\boldsymbol{E} = \boldsymbol{Y} - \boldsymbol{XA}$$

$$\boldsymbol{E} = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_m \end{pmatrix}, \boldsymbol{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix}, \boldsymbol{X} = \begin{pmatrix} x_{11} & x_{21} & \cdots & x_{p1} \\ x_{12} & x_{22} & \cdots & x_{p2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1m} & x_{2m} & \cdots & x_{pm} \end{pmatrix}, \quad \boldsymbol{A} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{pmatrix}$$

we consider minimizing of magnitude of $\boldsymbol{E}$. In this case, magnitude of $\boldsymbol{E}$ is not the sum of error term $\sum_{i=1}^{m} e_i$. The magnitude of vector and matrix is norm.

The definition of norm is as follow

$\boldsymbol{X} = (x_1, x_2, \cdots x_n)$、$1 \leq q \leq \infty$

$\sqrt[q]{|x_1|^q + |x_2|^q + \cdots + |x_n|^q}$ is L^q norm and is expressed as $\|\boldsymbol{X}\|_q$

$$\|\boldsymbol{X}\|_1 = |x_1| + |x_2| + \cdots + |x_n|$$

$$\|\boldsymbol{X}\|_2 = \sqrt{|x_1|^2 + |x_2|^2 + \cdots + |x_n|^2}$$

L^2norm is Euclid distance. Least square method is minimization of L^2 norm.

$$\|\boldsymbol{E}\|_2 = \sqrt{|e_1|^2 + |e_2|^2 + \cdots + |e_m|^2}$$

$$= \sqrt{{e_1}^2 + {e_2}^2 + \cdots + {e_m}^2}$$

$$\|\boldsymbol{E}\|_2{}^2 = {e_1}^2 + {e_2}^2 + \cdots + {e_m}^2$$

$$= (e_1, e_2, \cdots . e_m)\begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_m \end{pmatrix}$$

$$\boldsymbol{E}^T = (e_1, e_2, \cdots . e_m)$$

◦

$$\boldsymbol{E} = \boldsymbol{Y} - \boldsymbol{XA}$$

$$\boldsymbol{E}^T = (\boldsymbol{Y} - \boldsymbol{XA})^T$$

$$\|\boldsymbol{E}\|_2{}^2 = (\boldsymbol{Y} - \boldsymbol{XA})^T(\boldsymbol{Y} - \boldsymbol{XA})$$

$$= \boldsymbol{Y}^T\boldsymbol{Y} - \boldsymbol{A}^T\boldsymbol{X}^T\boldsymbol{Y} - \boldsymbol{Y}^T\boldsymbol{XA} + \boldsymbol{X}^T\boldsymbol{A}^T\boldsymbol{XA}$$

$$= \boldsymbol{Y}^T\boldsymbol{Y} - \boldsymbol{2Y}^T\boldsymbol{XA} + \boldsymbol{X}^T\boldsymbol{A}^T\boldsymbol{XA}$$

$\|E\|_2{}^2$is positive definite and we can obtain minimum value by following differential.

$$\frac{\partial\|\boldsymbol{E}\|_2{}^2}{\partial\boldsymbol{A}} = 0$$

$$\frac{\partial\|\boldsymbol{E}\|_2{}^2}{\partial\boldsymbol{A}} = \frac{\partial(\boldsymbol{Y}^T\boldsymbol{Y} - \boldsymbol{2Y}^T\boldsymbol{XA} + \boldsymbol{X}^T\boldsymbol{A}^T\boldsymbol{XA})}{\partial\boldsymbol{A}}$$

$$= \frac{\partial(\boldsymbol{Y}^T\boldsymbol{Y})}{\partial\boldsymbol{A}} - \boldsymbol{2}\frac{\partial(\boldsymbol{Y}^T\boldsymbol{XA})}{\partial\boldsymbol{A}} + \frac{\partial(\boldsymbol{X}^T\boldsymbol{A}^T\boldsymbol{XA})}{\partial\boldsymbol{A}} = \boldsymbol{0}$$

$$\frac{\partial(\boldsymbol{Y}^T\boldsymbol{Y})}{\partial\boldsymbol{A}} = 0$$

$$\frac{\partial(\boldsymbol{Y}^T\boldsymbol{XA})}{\partial\boldsymbol{A}} = \boldsymbol{Y}^T\boldsymbol{X}$$

$$\frac{\partial(\boldsymbol{X}^T\boldsymbol{A}^T\boldsymbol{XA})}{\partial\boldsymbol{A}} = \boldsymbol{2X}^T\boldsymbol{XA}$$

$$\frac{\partial\|\boldsymbol{E}\|_2{}^2}{\partial\boldsymbol{A}} = -\boldsymbol{2Y}^T\boldsymbol{X} + \boldsymbol{2X}^T\boldsymbol{XA} = \boldsymbol{0}$$

$$\boldsymbol{Y}^T\boldsymbol{X} = \boldsymbol{X}^T\boldsymbol{XA}$$

$$\boldsymbol{A} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{Y}^T\boldsymbol{X}$$

$$\boldsymbol{A} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$$

We consider $(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ is a matrix

$$(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T = \boldsymbol{X}^{\#}$$

In case of $\boldsymbol{Y} = \boldsymbol{XA}$, We can obtain $\boldsymbol{A}$ by multiplying $\boldsymbol{X}^{-1}$ from left

$$\boldsymbol{X}^{-1}\boldsymbol{Y} = \boldsymbol{X}^{-1}\boldsymbol{XA} = \boldsymbol{A}$$

Similarly, we can obtain $\boldsymbol{A}$ by multiplying $\boldsymbol{X}^{\#}$

$$\boldsymbol{X}^{\#}\boldsymbol{Y} = \boldsymbol{A}$$

$$\boldsymbol{X}^{\#} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$$

Formula 71

$$\boldsymbol{X}^{\#}\boldsymbol{X} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}(\boldsymbol{X}^T\boldsymbol{X}) = \boldsymbol{I}$$

The function of $\boldsymbol{X}^{\#}$ is analogous to inverse matrix, so we call this matrix pseudo-inverse matrix. Generalized inverse matrix is the other name of this matrix. However, $\boldsymbol{X}^{\#}$ gives least squares solution, when $\boldsymbol{X}$ is non-regular matrix. When $\boldsymbol{X}$ is regular matrix, $\boldsymbol{X}^{\#} = \boldsymbol{X}^{-1}$, and $\boldsymbol{X}^{\#}$ gives solution of simultaneous equation.

For an exercise, we us this method for simple linear regression

$$y = ax + b$$

$$\boldsymbol{Y} = \boldsymbol{XA}$$

$$\boldsymbol{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \boldsymbol{X} = \begin{pmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{pmatrix}, \qquad \boldsymbol{A} = \begin{pmatrix} a \\ b \end{pmatrix}$$

$$\boldsymbol{X}^T = \begin{pmatrix} x_1 & x_2 & \cdots & x_n \\ 1 & 1 & \cdots & 1 \end{pmatrix}$$

$$\boldsymbol{X}^T\boldsymbol{X} = \begin{pmatrix} x_1 & x_2 & \cdots & x_n \\ 1 & 1 & \cdots & 1 \end{pmatrix}\begin{pmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n} \boldsymbol{x_i}^2 & \sum_{i=1}^{n} \boldsymbol{x_i} \\ \sum_{i=1}^{n} \boldsymbol{x_i} & \boldsymbol{n} \end{pmatrix}$$

$$(\boldsymbol{X}^T\boldsymbol{X})^{-1} = \begin{pmatrix} \sum_{i=1}^{n} \boldsymbol{x_i}^2 & \sum_{i=1}^{n} \boldsymbol{x_i} \\ \sum_{i=1}^{n} \boldsymbol{x_i} & \boldsymbol{n} \end{pmatrix}^{-1} = \frac{1}{\boldsymbol{n}\sum_{i=1}^{n} \boldsymbol{x_i}^2 - (\sum_{i=1}^{n} \boldsymbol{x_i})^2} \begin{pmatrix} \boldsymbol{n} & -\sum_{i=1}^{n} \boldsymbol{x_i} \\ -\sum_{i=1}^{n} \boldsymbol{x_i} & \sum_{i=1}^{n} \boldsymbol{x_i}^2 \end{pmatrix}$$

$$\boldsymbol{X^{\#}} = (\boldsymbol{X^T X})^{-1}\boldsymbol{X^T} = \frac{1}{\boldsymbol{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}} \begin{pmatrix} \boldsymbol{n} & -\sum_{i=1}^{n} \boldsymbol{x_i} \\ -\sum_{i=1}^{n} \boldsymbol{x_i} & \sum_{i=1}^{n} \boldsymbol{x_i^2} \end{pmatrix} \begin{pmatrix} x_1 & x_2 & \cdots & x_n \\ 1 & 1 & \cdots & 1 \end{pmatrix}$$

$$= \frac{1}{\boldsymbol{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}} \begin{pmatrix} nx_1 - \sum_{i=1}^{n} \boldsymbol{x_i} & \boldsymbol{n}x_2 - \sum_{i=1}^{n} \boldsymbol{x_i} & \cdots & \boldsymbol{n}x_n - \sum_{i=1}^{n} \boldsymbol{x_i} \\ -x_1 \sum_{i=1}^{n} \boldsymbol{x_i} + \sum_{i=1}^{n} \boldsymbol{x_i^2} & -x_2 \sum_{i=1}^{n} \boldsymbol{x_i} + \sum_{i=1}^{n} \boldsymbol{x_i^2} & \cdots & x_n \sum_{i=1}^{n} \boldsymbol{x_i} + \sum_{i=1}^{n} \boldsymbol{x_i^2} \end{pmatrix}$$

$$\boldsymbol{X^{\#}Y} = \frac{1}{\boldsymbol{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}} \begin{pmatrix} nx_1 - \sum_{i=1}^{n} \boldsymbol{x_i} & \boldsymbol{n}x_2 - \sum_{i=1}^{n} \boldsymbol{x_i} & \cdots & \boldsymbol{n}x_n - \sum_{i=1}^{n} \boldsymbol{x_i} \\ -x_1 \sum_{i=1}^{n} \boldsymbol{x_i} + \sum_{i=1}^{n} \boldsymbol{x_i^2} & -x_2 \sum_{i=1}^{n} \boldsymbol{x_i} + \sum_{i=1}^{n} \boldsymbol{x_i^2} & \cdots & x_n \sum_{i=1}^{n} \boldsymbol{x_i} + \sum_{i=1}^{n} \boldsymbol{x_i^2} \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

$$= \frac{1}{\boldsymbol{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}} \begin{pmatrix} \boldsymbol{n\sum_{i=1}^{n} x_i y_i + \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i} \\ \boldsymbol{\sum_{i=1}^{n} x_i y_i \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i^2} \end{pmatrix} = \begin{pmatrix} a \\ b \end{pmatrix}$$

$$a = \frac{\boldsymbol{n\sum_{i=1}^{n} x_i y_i + \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}}{\boldsymbol{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}}$$

$$b = \frac{\boldsymbol{\sum_{i=1}^{n} x_i y_i \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i^2}}{\boldsymbol{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}}$$

a is regression coefficient