VI. Multivariable analysis

VI-1. Optimization

VI-1-1. Multiple linear regression analysis

VI-1-1-1. Multiple linear regression by pseudo-inverse matrix

When the explanatory variable in only one, objective variables can be explained a coefficient of the explanatory variable, explanatory variable, a constant and error as follow

$$y = ax + b + e$$

y: Objective variable

x: explanatory variable

a: coefficient

b: constant

e: error, ionexplicable residual

This regression is called simple linear regression. When there exist more than one explanatory variables, the relation is expressed as follow and we call the operation as multiple linear regression.

$$y = a_0 + a_1x_1 + a_2x_2 + \cdots + a_nx_n + e$$

y: Objective variable

x_i: explanatory variable

a_i: coefficient

a_0: constant

e: error, ionexplicable residual

There are various ideas, models and methods for optimization. Least square method and most likelihood method are commonly used method for optimization. It is known that the result regression by most likelihood method agrees to the result by least square method when the likelihood method hypothesizes normal distribution. The author explains multiple linear regression as least square method using pseudo-inverse matrix.

When there is following data set,

$$\mathrm{y}_1, x_{11}, x_{21} \cdots x_{p1}$$

$$\mathrm{y}_2, x_{12}, x_{22} \cdots x_{p2}$$

$$\mathrm{y}_3, x_{11}, x_{23} \cdots x_{p3}$$

$$\vdots$$

$$\mathrm{y}_m, x_{1m}, x_{2m} \cdots x_{pm}$$

The relation can be expressed as follow

$$\mathrm{y}_1 = a_0 + a_1x_{11} + a_2x_{21} + \cdots + a_px_{p1} + e_1$$
$$\mathrm{y}_2 = a_0 + a_1x_{12} + a_2x_{22} + \cdots + a_px_{p2} + e_2$$
$$\mathrm{y}_3 = a_0 + a_1x_{13} + a_2x_{23} + \cdots + a_px_{p3} + e_3$$
$$\vdots$$
$$\mathrm{y}_n = a_0 + a_1x_{1m} + a_2x_{2m} + \cdots + a_px_{pm} + e_m$$

The relation can be expressed as follow

$$\mathrm{y} = a_0 + a_1x_1 + a_2x_2 + \cdots + a_px_p + e$$

a_0: constant term

$$= \begin{pmatrix} 1 & x_1 & \cdots & x_p \end{pmatrix}\begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_p \end{pmatrix} + e$$

Putting

$$\boldsymbol{Y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad \boldsymbol{X_{+1}} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}, \quad \boldsymbol{A_{+1}} = \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_p \end{pmatrix}, \quad \boldsymbol{E} = \begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix}$$

$$\boldsymbol{Y} = \boldsymbol{X_{+1}A_{+1}} + \boldsymbol{E}$$

We learned that pseudo-inverse matrix $\left(\boldsymbol{X_{+1}}^{\#}\right)$ gives optimum solution of $\boldsymbol{A_{+1}}$ by least square method in paragraph “V-3-4. Optimization and pseudo-inverse matrix”.

$$\boldsymbol{A_{+1}} = \boldsymbol{X_{+1}}^{\#}\boldsymbol{Y}$$

Formula 73

Pseudo-inverse matrix $\left(\boldsymbol{X_{+1}}^{\#}\right)$ can be obtain by direct calculation of inverse matrix or inverse operation of singular value decomposition as follow.

$$\boldsymbol{X_{+1}}^{\#} = \left(\boldsymbol{X_{+1}}^{T}\boldsymbol{X_{+1}}\right)^{-1}\boldsymbol{X_{+1}}^{T}$$

or

$$\boldsymbol{X_{+1}}^{T}\boldsymbol{X_{+1}} = \begin{pmatrix} 1 & \cdots & 1 \\ x_{11} & \cdots & x_{n1} \\ \vdots & \ddots & \vdots \\ x_{1p} & \cdots & x_{np} \end{pmatrix}\begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix} = \begin{pmatrix} n & \sum_{\boldsymbol{j=1}}^{\boldsymbol{n}} x_{j1} & \cdots & \sum_{\boldsymbol{j=1}}^{\boldsymbol{n}} x_{jp} \\ \sum_{\boldsymbol{j=1}}^{\boldsymbol{n}} x_{j1} & & & \\ \vdots & & \boldsymbol{X}^{T}\boldsymbol{X} & \\ \sum_{\boldsymbol{j=1}}^{\boldsymbol{n}} x_{jp} & & & \end{pmatrix}$$

$$\boldsymbol{U^{T}X^{T}XU} = \boldsymbol{\Sigma} = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_p \end{pmatrix}$$

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0$$

In reality, the operation by upper equation is hard task. Inverse matrix can be calculated by inverse operation of singular value decomposition.

$$\boldsymbol{X}^{\#} = \boldsymbol{V\Sigma}^{\#}\boldsymbol{U}^T$$

$$\boldsymbol{X_{+1}X_{+1}}^T = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix} \begin{pmatrix} 1 & \cdots & 1 \\ x_{11} & \cdots & x_{n1} \\ \vdots & \ddots & \vdots \\ x_{1p} & \cdots & x_{np} \end{pmatrix}$$

$$\boldsymbol{V^T X_{+1} X_{+1}}^T \boldsymbol{V} = \boldsymbol{\Sigma_V}$$

$$\boldsymbol{\Sigma}^{\#} = \begin{pmatrix} \frac{1}{\gamma_1} & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & \frac{1}{\gamma_2} & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \cdots & \frac{1}{\gamma_r} & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \end{pmatrix}_{n \times p}$$

$$\gamma_i = \lambda_i \geq thresholod\ value$$

VI-1-1-2. Geometric meaning of multiple regression

In this paragraph, we consider geometrical meaning of multiple linear regression. At first, we consider the distance of a point and hyperplane, because meaning of the regression is to explain the distribution by a hyperplane of which average distance between the hyperplane and datapoints is minimum. Figure 62 shows equation of hyperplane including origin. Coordinate origin is O : $(0 \quad \cdots \quad 0)$.

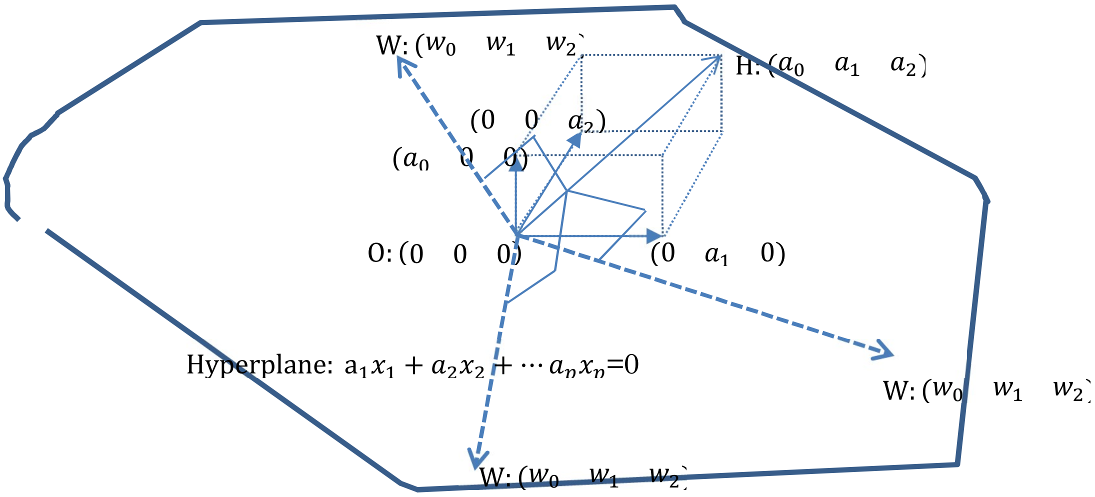
W: $(w_0 \quad w_1 \quad w_2)$
H: $(a_0 \quad a_1 \quad a_2)$
$(0 \quad 0 \quad a_2)$
$(a_0 \quad 0 \quad 0)$
O: $(0 \quad 0 \quad 0)$
$(0 \quad a_1 \quad 0)$
Hyperplane: $a_1x_1 + a_2x_2 + \cdots a_nx_n$=0
W: $(w_0 \quad w_1 \quad w_2)$
W: $(w_0 \quad w_1 \quad w_2)$

Fig 62. Formula of hyperplane including origin.

The hyperplane is characterized by normal vector $\overrightarrow{\mathrm{OH}}$ which is orthogonal to the hyper plane.

$$\overrightarrow{\mathrm{OH}} = (a_0 \quad a_1 \quad \cdots \quad a_p)$$

$\overrightarrow{\mathrm{OW}}$ is vector on the hyperplane.

$$\overrightarrow{\mathrm{OW}} = (w_0 \quad w_1 \quad \cdots \quad w_0)$$

$$\overrightarrow{\mathrm{OH}} \perp \overrightarrow{\mathrm{OW}}$$

$$(w_0 \quad w_1 \quad \cdots \quad w_p)\begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_p \end{pmatrix} = 0$$

$$\mathrm{a}_0 w_0 + a_1 w_1 + \cdots a_p w_p = 0$$

For simplification, we take unit vector of normal vector (***U***).

$$a_0{}^2 + a_1{}^2 + \cdots + a_p{}^2 = 1$$

We consider a parallel hyperplane to the hyperplane including origin as a general hyperplane. The point on the parallel hyperplane is $\mathrm{X}(x_0 \quad \cdots \quad x_p)$. The distance from the hyperplane including origin is t. The length of t is constant and normal vector from the hyper plane including origin to X is $\overrightarrow{\mathrm{WX}}$.

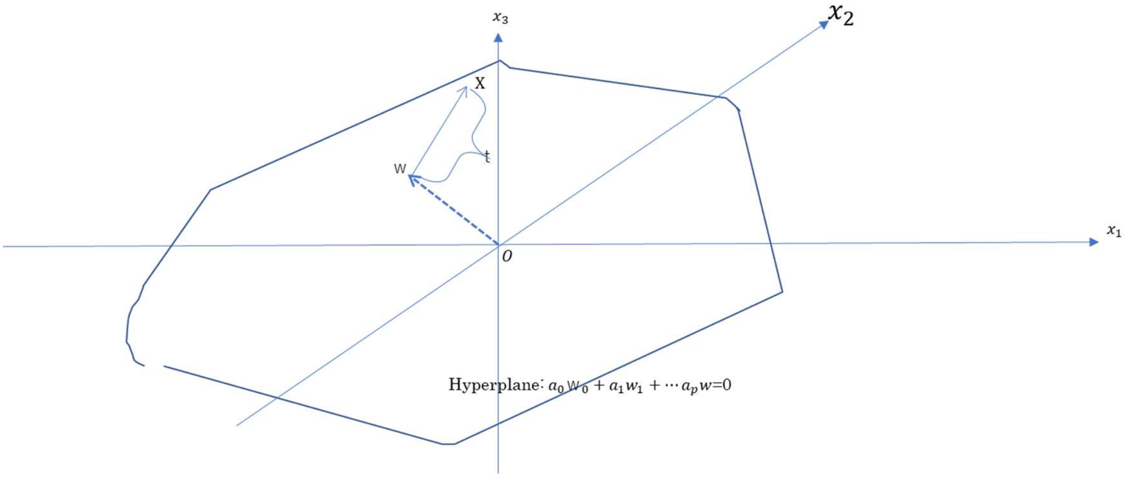

Fig 63. Points in a hyperplane

$$\overrightarrow{\mathrm{WX}} = \overrightarrow{\mathrm{OX}} - \overrightarrow{\mathrm{OW}} = t(a_0 \quad \cdots \quad a_p)$$

$$(x_0 - w_0 \quad \cdots \quad x_p - w_p) = t(a_0 \quad \cdots \quad a_p)$$

We take inner products of both sides.

$$(x_0 - w_0 \quad \cdots \quad x_p - w_p)\begin{pmatrix} a_0 \\ \vdots \\ a_p \end{pmatrix} = t(a_0 \quad \cdots \quad a_p)\begin{pmatrix} a_0 \\ \vdots \\ a_p \end{pmatrix}$$

$$\mathrm{a}_0 x_0 + a_1 x_1 + \cdots a_p x_p - (\mathrm{a}_0 w_0 + a_1 w_1 + \cdots a_p w_p) = t(a_0{}^2 + a_1{}^2 + \cdots + a_p{}^2)$$

$$\mathrm{a}_0 x_0 + a_1 x_1 + \cdots a_p x_p = t$$

$$\because\ \mathrm{a}_0 w_0 + a_1 w_1 + \cdots a_p w_p = 0$$

$$a_0{}^2 + a_1{}^2 + \cdots + a_p{}^2 = 1$$

From this we express hyperplane by following equation

$$\mathrm{a}_0 x_0 + a_1 x_1 + \cdots a_p x_p = t$$

In case of $a_0{}^2 + a_1{}^2 + \cdots + a_p{}^2 = 1,$ t is distance from the hyperplane to origin. The author explains an optimization, which is a kind of regression but is not general multiple regression.

$$a_0{}' x_0 + a_1{}' x_1 + \cdots + a_p{}' x_p = c$$

c: constant

When all data exist on the hyperplane

$$\boldsymbol{XA}' = \mathrm{c}\begin{pmatrix}1\\ \vdots\\ 1\end{pmatrix}_{n\times 1}$$

$$a_i = \frac{a_i{}'}{c}$$

$$\boldsymbol{X} = \begin{pmatrix}x_{10} & \cdots & x_{1p}\\ \vdots & \ddots & \vdots\\ x_{n0} & \cdots & x_{np}\end{pmatrix}_{n\times p} \quad , \quad \boldsymbol{A} = \begin{pmatrix}a_0\\ \vdots\\ a_p\end{pmatrix}$$

$$a_0 x_0 + a_1 x_1 + \cdots + a_p x_p = 1$$

$$\boldsymbol{XA} = \begin{pmatrix}1\\ \vdots\\ 1\end{pmatrix}_{n\times 1}$$

Generally, all data do not exist on the hyperplane and matrix $\boldsymbol{X}$ is not regular.

$$\boldsymbol{XA} \neq \begin{pmatrix}1\\ \vdots\\ 1\end{pmatrix}$$

We consider error term.

$$\boldsymbol{XA} - \begin{pmatrix}1\\ \vdots\\ 1\end{pmatrix} = \boldsymbol{E}$$

$$\boldsymbol{E} = \begin{pmatrix}e_1\\ \vdots\\ e_n\end{pmatrix}$$

e_j: error

We minimize $\boldsymbol{E}^{\boldsymbol{T}}\boldsymbol{E}$ by pseud-invers matrix $\boldsymbol{X}^{\#}$

$$\boldsymbol{XA} - \begin{pmatrix}1\\ \vdots\\ 1\end{pmatrix} = \boldsymbol{E}$$

$$\boldsymbol{XA} = \begin{pmatrix}1\\ \vdots\\ 1\end{pmatrix} + \boldsymbol{E}$$

$$\boldsymbol{X}^{\#}\boldsymbol{XA} = \boldsymbol{X}^{\#}\begin{pmatrix}1\\ \vdots\\ 1\end{pmatrix}$$

$$\boldsymbol{A} = \boldsymbol{X}^{\#}\begin{pmatrix}1\\ \vdots \\ 1\end{pmatrix}$$

This is not generally used multiple regression. Multiple regression is used for expression of relation between an objective variable and other explanatory variables, though this equation expresses the relation among all explaining variables. However, we can say that this is a kind of multiple regression. generally multiple regression is used for

We have to fix a variable as an objective variable and the coefficient of objective variables should be 1. Generally objective variance is denoted as y. Some readers may think indiscreetly that we could get solution as follow.

$$a_0x_0 + a_1x_1 + \cdots + a_px_p = 0$$

Divide both side by a_0

$$x_0 + \frac{a_1}{a_0}x_1 + \cdots + \frac{a_p}{a_0}x_p = \frac{c}{a_0}$$

$$x_0 = -\left(\frac{a_1}{a_0}x_1 + \cdots + \frac{a_p}{a_0}x_p\right)$$

We rewrite $-\frac{a_i}{a_0}$ as b_i, x_0 as y.

$$y = b_1x_1 + \cdots + b_px_p$$

When we add intercept,

$$y = b_1x_1 + \cdots + b_px_p\text{+c}$$

We do not need to consider c, because we can transform $\mathrm{y}' = \mathrm{y} - \mathrm{c}$.

However, this is not optimum solution for explanation of y, because coefficients of the formula is selected as optimum coefficient for explanation of total relation among variables not for explanation a particular variable.

The reader can understand reading following trial.

When we denote $a_0 = -1\ x_0 = y$

The formula of the hyperplane is as follow

$$y - (a_1x_1 + \cdots + a_px_p) = 0$$

The distance from a point, $\mathrm{D}_j\colon \begin{pmatrix} d_{yj} & d_{x1j} & \cdots & d_{xpj}\end{pmatrix}$ in hyperspace to the hyperplane is as follow

$$d_j = \frac{\left|d_{yj} - (a_1d_{x1j} + \cdots + a_pd_{xpj})\right|}{\sqrt{1^2 + {a_1}^2 + \cdots + {a_p}^2}} = \frac{|e_j|}{\sqrt{1^2 + {a_1}^2 + \cdots + {a_p}^2}}$$

We should minimize $\sum_{j=1}^{n} {d_j}^2$ not $\sum_{j=1}^{n} {e_j}^2$. We could not remove dominator, because

$$\sqrt{1^2 + {a_1}^2 + \cdots + {a_p}^2} \neq 1$$

$$d_j \neq e_j$$

We have to consider another approach.

When we denote objective variables as y, we need to find optimum hyperplane which minimize the distance between observed objective value and the hyperplane along with y axis as shown in figure 64.

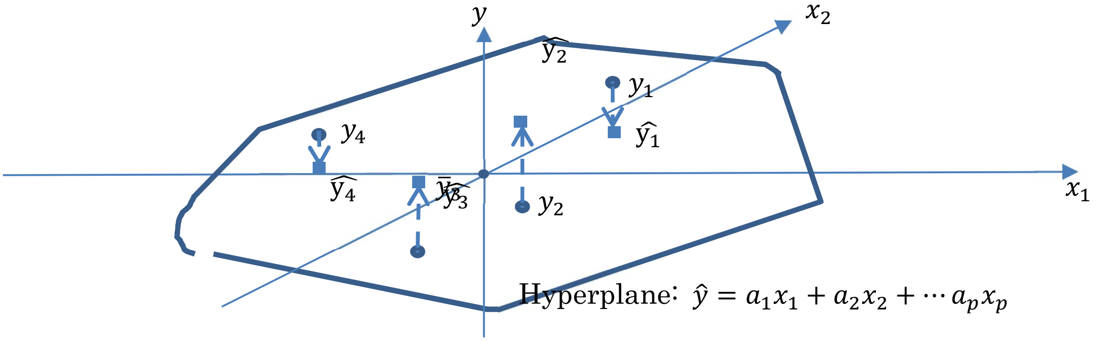

Fig. 64. Distance in y axis from data point and hyperplane.

For simplification, we consider the case when y intercept is 0. Such distribution is simply obtainable by subtracting average value of variables, and we can simplify the operation by this transformation.

When the matrix $\boldsymbol{X}$ is regular, we can solve following equation

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

, and express the relation as follow

$$y = a_1x_1 + a_2x_2 + \cdots a_px_p$$

However actually, observed y include error. The equation is expressed as follow.

$$y = a_1x_1 + a_2x_2 + \cdots a_px_p + e$$

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

$$\boldsymbol{E} = \begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix}$$

$$e_i\text{: error}$$

We need minimize $\boldsymbol{E^TE}$.

We calculate approximate solution of following equation.

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

$$\boldsymbol{XA} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

$$\boldsymbol{X}^{\#}\boldsymbol{X}\boldsymbol{A} = \boldsymbol{X}^{\#}\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

$$\boldsymbol{A} = \boldsymbol{X}^{\#}\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

Intercept

Upper explanation is a practical method to simplify the matrix and calculation. However, we need to remake the converted dataset from row dataset as follow.

$$y = d_y - \overline{d_y}$$

$$x_{ij} = d_{ij} - \overline{d_J}$$

d: original data

$\overline{d_y}$ and $\overline{d_J}$ is average

$$\overline{d_J} = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$$

$$\overline{d_y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

Then we put this in following equation

$$y = a_1x_1 + a_2x_2 + \cdots a_px_p$$

$$d_y - \overline{d_y} = a_1\left(d_1 - \overline{d_1}\right) + a_2\left(d_2 - \overline{d_2}\right) + \cdots + a_p\left(d_p - \overline{d_p}\right)$$

$$d_y = a_1d_1 + a_2d_2 + \cdots + a_pd_p + \overline{d_y} - \left(a_1\overline{d_1} + a_2\overline{d_2} + \cdots + a_p\overline{d_p}\right)$$

Then we can obtain intercept as follow.

$$\overline{d_y} - \left(a_1\overline{d_1} + a_2\overline{d_2} + \cdots + a_p\overline{d_p}\right) = a_0$$

If we feel this process is tangled, one possible idea for estimation of y intercept directly from the original dataset is to make constant term in matrix $\boldsymbol{X}$ and $\boldsymbol{A}$ as shown in the introduction of operation.

$$\boldsymbol{Y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad \boldsymbol{X_{+1}} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}, \quad \boldsymbol{A_{+1}} = \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_p \end{pmatrix}, \quad \boldsymbol{E} = \begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix}$$

$$\boldsymbol{Y} = \boldsymbol{X_{+1}}\boldsymbol{A_{+1}} + \boldsymbol{E}$$

$$\boldsymbol{A_{+1}} = \boldsymbol{X_{+1}}^{\#}\boldsymbol{Y}$$

a_0: constant term

More practical solution is assignment of equation of transformation to the equation of y and x after getting t $\boldsymbol{A}$.

$$y = (x_1 \quad \cdots \quad x_p)\begin{pmatrix} a_1 \\ \vdots \\ a_p \end{pmatrix} = a_1x_1 + a_2x_2 + \cdots a_px_p$$

$$y = d_y - \overline{d_y}$$

$$x_{ij} = d_{ij} - \overline{d_j}$$

d: original data

$\overline{d_y}$ and $\overline{d_j}$ is average

$$\overline{d_j} = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$$

$$\overline{d_y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

$$y = a_1x_1 + a_2x_2 + \cdots a_px_p$$

$$d_y - \overline{d_y} = a_1\left(d_1 - \overline{d_1}\right) + a_2\left(d_2 - \overline{d_2}\right) + \cdots + a_p\left(d_p - \overline{d_p}\right)$$

$$d_y = a_1d_1 + a_2d_2 + \cdots + a_pd_p + \overline{d_y} - \left(a_1\overline{d_1} + a_2\overline{d_2} + \cdots + a_p\overline{d_p}\right)$$

$$\overline{d_y} - \left(a_1\overline{d_1} + a_2\overline{d_2} + \cdots + a_p\overline{d_p}\right) = a_0$$

This is excursus. What the author wants to say in this explanation is that axis of objective variable is normal vector of the hyperplane of explaining the relation among explanatory variables. The equation of the multiple regression is obtainable by moving the hyperplane to include the point of mean of all variables.

VI-1-1-3. Exposition by differentiation and simultaneous equation

Using pseudo-inverse matrix, the author can easily introduce operation of multiple linear regression. However, pseudo-inverse matrix is a kind of black box for readers who do not have knowledge of linear algebra. The author adds detailed exposition using only differentiation and simultaneous equation for such readers.

Example of data set

	Explanatory variables			Objective variable
Sample no.	0	1	2 ⋯ i ⋯ p	y
1	a_0	d_{11}	$d_{12} \cdots d_{1i} \cdots d_{1p}$	d_{1y}
2	a_0	d_{21}	$d_{22} \cdots d_{2i} \cdots d_{2p}$	d_2y
⋮	⋮			⋮
k	a_0	d_{k1}	$d_{k2} \cdots d_{ki} \cdots dx_{kp}$	d_{ky}
⋮	⋮			⋮
n	a_0	d_{n1}	$d_{n2} \cdots d_{ni} \cdots d_{np}$	d_{ny}

$$d_{ky} = \ a_0 + \ a_1 d_{k1} + \ a_2 d_{k2} + \cdots + a_i d_{ki} + \cdots + a_p d_{kp} + e_k$$

$$e_k = d_{ky} - (\ a_0 + a_1 d_{k1} + \ a_2 d_{k2} + \cdots + a_i d_{ki} + \cdots + a_p d_{kp})$$

$${e_k}^2 = \left(d_{ky} - (\ a_0 + \ a_1 d_{k1} + \ a_2 d_{k2} + \cdots + a_i d_{ki} + \cdots + a_p d_{kp})\right)^2$$

$$E = \sum_{k=1}^{n} {e_k}^2 = \sum_{k=1}^{n} \left(d_{ky} - (\ a_0 + \ a_1 d_{k1} + \ a_2 d_{k2} + \cdots + a_i d_{ki} + \cdots + a_p d_{kp})\right)^2$$

It is obvious that E has minimum value. We calculate $a_0, \cdots a_p \ and \ a_{p+1}$ which gives extreme value of E by differentiation of E by $a_0, \cdots a_p$

$$\frac{dE}{da_i} = 0$$

When we denote

$$F_k = d_{ky} - (\ a_0 + \ a_1 d_{ki} + \ a_2 d_{k2} + \cdots + a_i d_{kp})$$

$$E = \sum_{k=1}^{n} {F_k}^2$$

$$\frac{dE}{da_i} = \frac{dE}{dF_k}\frac{dF_k}{da_i}$$

$$\frac{dE}{dF} = 2\sum_{k=1}^{n} F_k = 2\sum_{k=1}^{n} \left(d_{ky} - (\ \ a_0 + a_1 d_{k1} + \ a_2 d_{k2} + \cdots + a_i d_{ki} + \cdots + a_p d_{kp})\right)$$

$$\frac{dF_k}{da_i} = 2d_{ki}$$

$$\frac{dF_k}{da_0} = 2n$$

$$\frac{dE}{da_i} = 2\sum_{k=1}^{n} \left(d_{ky} - (\ a_0 + \ a_1 d_{k1} + \ a_2 d_{k2} + \cdots + a_i d_{ki} + \cdots + a_p d_{kp})\right) d_{ki} = 0$$

$$d_{oi} = 1$$

$$\frac{dE}{2da_i} = \sum_{k-1}^{n} d_{ky} d_{ki} - \left(a_0 + \sum_{k=1}^{n} d_{ki} \ a_1 \sum_{k-1}^{n} d_{k1} d_{ki} + \ x_2 \sum_{k=1}^{n} d_{k2} d_{ki} + \cdots + x_p \sum_{k=1}^{n} d_{kp} d_{ki}\right) = 0$$

$$a_1 \sum_{k-1}^{n} d_{k1} d_{ki} + \ a_2 \sum_{k=1}^{n} d_{k2} d_{ki} + \cdots + a_p \sum_{k=1}^{n} d_{kp} d_{ki} + a_{p+1} \sum_{k=1}^{n} d_{ki} = \sum_{k-1}^{n} d_{ky} d_{ki}$$

When we rewrite this equation by form of simultaneous equation

$$na_0 + a_1 \sum_{k-1}^{n} d_{k1} + \ a_2 \sum_{k=1}^{n} d_{k2} + \cdots + a_p \sum_{k=1}^{n} d_{kp} = \sum_{k-1}^{n} d_{ky}$$

$$a_0 \sum_{k=1}^{n} d_{k1} + a_1 + \sum_{k-1}^{n} d_{k1} d_{k1} + \ a_2 \sum_{k=1}^{n} d_{k2} d_{k1} + \cdots + a_p \sum_{k=1}^{n} d_{kp} d_{k1} = \sum_{k-1}^{n} d_{ky} d_{k1}$$

$$+a_0 \sum_{k=1}^{n} d_{k2} + a_1 \sum_{k-1}^{n} d_{k1} d_{k2} + \ a_2 \sum_{k=1}^{n} d_{k2} d_{k2} + \cdots + a_p \sum_{k=1}^{n} d_{kp} d_{k2} = \sum_{k-1}^{n} d_{ky} d_{k2}$$

$$\vdots$$

$$+a_0 \sum_{k=1}^{n} d_{kp} + a_1 \sum_{k-1}^{n} d_{k1} d_{kp} + \ a_2 \sum_{k=1}^{n} d_{k2} d_{kp} + \cdots + a_p \sum_{k=1}^{n} d_{kp} d_{kp} = \sum_{k-1}^{n} d_{ky} d_{kp}$$

We can get $a_0, a_1, \cdots, a_p$ as solution of upper simultaneous equation. For clear understanding of the relation between solution by pseudo inverse matrix and solution by differentiation and simultaneous equation, we express the simultaneous equation in form of matrix

$$\begin{pmatrix} n & \sum_{k-1}^{n} d_{k1} & \sum_{k=1}^{n} d_{k2} & \cdots & \sum_{k=1}^{n} d_{kp} \\ \sum_{k=1}^{n} d_{k1} & \sum_{k-1}^{n} {d_{k1}}^2 & \sum_{k=1}^{n} d_{k2} d_{k1} & \cdots & \sum_{k=1}^{n} d_{kp} d_{k1} \\ \sum_{k=1}^{n} d_{k2} & \sum_{k=1}^{n} d_{k1} d_{k2} & \sum_{k=1}^{n} {d_{k2}}^2 & \cdots & \sum_{k=1}^{n} d_{kp} d_{k2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum_{k=1}^{n} d_{kp} & \sum_{k-1}^{n} d_{k1} d_{kp} & \sum_{k=1}^{n} d_{k2} d_{kp} & \cdots & \sum_{k=1}^{n} {d_{kp}}^2 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_p \end{pmatrix}$$

$$= \begin{pmatrix} 1 & \cdots & 1 \\ d_{11} & \cdots & d_{n1} \\ \vdots & \ddots & \vdots \\ d_{1p} & \cdots & d_{np} 1 \end{pmatrix} \begin{pmatrix} d_{1y} \\ \vdots \\ d_{ny} \end{pmatrix}$$

Equation i

We denote matrixes as follows.

$$\begin{pmatrix} 1 & d_{11} & \cdots & d_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & d_{n1} & \cdots & d_{np} \end{pmatrix}_{n\times(p+1)} = \boldsymbol{D_{+1}}$$

$$\begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_p \end{pmatrix} = \boldsymbol{A_{+1}}$$

$$\begin{pmatrix} d_{1y} \\ \vdots \\ d_{ny} \end{pmatrix} = \boldsymbol{Y}$$

,

$$\boldsymbol{D_{+1}}^{\boldsymbol{T}} = \begin{pmatrix} 1 & \cdots & 1 \\ d_{11} & \cdots & d_{n1} \\ \vdots & \ddots & \vdots \\ d_{1p} & \cdots & d_{np} 1 \end{pmatrix}$$

$$\boldsymbol{D_{+1}}^{\boldsymbol{T}}\boldsymbol{D_{+1}} = \begin{pmatrix} 1 & \cdots & 1 \\ d_{11} & \cdots & d_{n1} \\ \vdots & \ddots & \vdots \\ d_{1p} & \cdots & d_{np}1 \end{pmatrix} \begin{pmatrix} 1 & d_{11} & \cdots & d_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & d_{n1} & \cdots & d_{np} \end{pmatrix} = \begin{pmatrix} n & \sum_{k-1}^{n} d_{k1} & \sum_{k=1}^{n} d_{k2} & \cdots & \sum_{k=1}^{n} d_{kp} \\ \sum_{k=1}^{n} d_{k1} & \sum_{k-1}^{n} {d_{k1}}^2 & \sum_{k=1}^{n} d_{k2}d_{k1} & \cdots & \sum_{k=1}^{n} d_{kp}d_{k1} \\ \sum_{k=1}^{n} d_{k2} & \sum_{k=1}^{n} d_{k1}d_{k2} & \sum_{k=1}^{n} {d_{k2}}^2 & \cdots & \sum_{k=1}^{n} d_{kp}d_{k2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum_{k=1}^{n} d_{kp} & \sum_{k-1}^{n} d_{k1}d_{kp} & \sum_{k=1}^{n} d_{k2}d_{kp} & \cdots & \sum_{k=1}^{n} {d_{kp}}^2 \end{pmatrix}$$

equation i can be expressed as follow.

$$\boldsymbol{D_{+1}}^{\boldsymbol{T}}\boldsymbol{D_{+1}}\boldsymbol{A_{+1}} = \boldsymbol{D_{+1}}^{\boldsymbol{T}}\boldsymbol{Y}$$

Multiply$\left(\boldsymbol{D_{+1}}^{\boldsymbol{T}}\boldsymbol{D_{+1}}\right)^{-1}$ to both sides.

$$\left(\boldsymbol{D_{+1}}^{\boldsymbol{T}}\boldsymbol{D_{+1}}\right)^{-1}\left(\boldsymbol{D_{+1}}^{\boldsymbol{T}}\boldsymbol{D_{+1}}\right)\boldsymbol{A_{+1}} = \left(\boldsymbol{D_{+1}}^{\boldsymbol{T}}\boldsymbol{D_{+1}}\right)^{-1}\boldsymbol{D_{+1}}^{\boldsymbol{T}}\boldsymbol{Y}$$

$$\boldsymbol{A_{+1}} = \left(\boldsymbol{D_{+1}}^{\boldsymbol{T}}\boldsymbol{D_{+1}}\right)^{-1}\boldsymbol{D_{+1}}^{\boldsymbol{T}}\boldsymbol{Y}$$

$$\left(\boldsymbol{D_{+1}}^{\boldsymbol{T}}\boldsymbol{D_{+1}}\right)^{-1}\boldsymbol{D_{+1}}^{\boldsymbol{T}} = \boldsymbol{D_{+1}}^{\#}$$

$\boldsymbol{D_{+1}}^{\#}$: pseudo inverse matrix of$\boldsymbol{D_{+1}}$

We could confirm identity of solution by pseud inverse matrix and solution by differentiation and simultaneous equation. However, this is excursus. Going back to mainstream.

When we neglect constant term, the relation can be expresses as follow

$$\boldsymbol{DA} = \boldsymbol{Y}$$

$$\boldsymbol{A} = \boldsymbol{D^T Y}$$

$$(\boldsymbol{D^T D})^{-\boldsymbol{1}}\boldsymbol{D^T DA} = (\boldsymbol{D^T D})^{-\boldsymbol{1}}\boldsymbol{D^T Y}$$

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

$$\boldsymbol{S} = \boldsymbol{D^T D} = \begin{pmatrix} d_{11} & \cdots & d_{n1} \\ \vdots & \ddots & \vdots \\ d_{1p} & \cdots & d_{np} \end{pmatrix} \begin{pmatrix} d_{11} & \cdots & d_{1p} \\ \vdots & \ddots & \vdots \\ d_{n1} & \cdots & d_{np} \end{pmatrix} = \begin{pmatrix} \sum_{k-1}^{n} {d_{k1}}^2 & \sum_{k=1}^{n} d_{k2}d_{k1} & \cdots & \sum_{k=1}^{n} d_{kp}d_{k1} \\ \sum_{k=1}^{n} d_{k1}d_{k2} & \sum_{k=1}^{n} {d_{k2}}^2 & \cdots & \sum_{k=1}^{n} d_{kp}d_{k2} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{k-1}^{n} d_{k1}d_{kp} & \sum_{k=1}^{n} d_{k2}d_{kp} & \cdots & \sum_{k=1}^{n} {d_{kp}}^2 \end{pmatrix}$$

Denoting

$$SS_{ij} = \sum_{k=1}^{n} d_{ki}d_{kj}$$

$$\boldsymbol{S} = \boldsymbol{D^T D} = \begin{pmatrix} SS_{11} & SS_{12} & \cdots & SS_{1i} & \cdots & SS_{1n} \\ SS_{21} & SS_{12} & \cdots & SS_{2i} & \cdots & SS_{2n} \\ \vdots & \vdots & \ddots & \vdots & & \vdots \\ SS_{j1} & SS_{j2} & \cdots & SS_{ji} & \cdots & SS_{jn} \\ \vdots & \vdots & & \vdots & \ddots & \vdots \\ SS_{n1} & SS_{n2} & \cdots & SS_{ni} & \cdots & SS_{nn} \end{pmatrix}$$

Using this notation

$$\boldsymbol{D^T Y} = \begin{pmatrix} d_{11} & \cdots & d_{n1} \\ \vdots & \ddots & \vdots \\ d_{1p} & \cdots & d_{np} \end{pmatrix} \begin{pmatrix} d_{1y} \\ \vdots \\ d_{ny} \end{pmatrix} = \begin{pmatrix} \sum_{k=1}^{n} d_{k1} d_{ky} \\ \vdots \\ \sum_{k=1}^{n} d_{kp} d_{ky} \end{pmatrix}$$

$$\boldsymbol{D^T D A} = \boldsymbol{D^T Y}$$

$$\boldsymbol{SA} = \begin{pmatrix} SS_{1y} \\ \vdots \\ SS_{py} \end{pmatrix}$$

$$\boldsymbol{A} = \boldsymbol{S^{-1}} \begin{pmatrix} SS_{1y} \\ \vdots \\ SS_{py} \end{pmatrix}$$

This simplest expression of the solution. The author loves simple expression, though some readers feel frustration with simple expression, because we cannot understand operation of calculation from simple expression. Followings are example of operation.

Sample no.	1	2	3	y
1	d_{11}	d_{12}	d_{13}	d_{1y}
2	d_{21}	d_{22}	d_{23}	d_{2}y
⋮		⋮		⋮
k	d_{k1}	d_{k2}	d_{k3}	d_{ky}
⋮		⋮		⋮
n	d_{n1}	d_{n2}	d_{n3}	d_{ny}

$$\boldsymbol{S} = \boldsymbol{D^T D} = \begin{pmatrix} d_{11} & \cdots & d_{n1} \\ \vdots & \ddots & \vdots \\ d_{1p} & \cdots & d_{np} \end{pmatrix} \begin{pmatrix} d_{11} & \cdots & d_{1p} \\ \vdots & \ddots & \vdots \\ d_{n1} & \cdots & d_{np} \end{pmatrix} = \begin{pmatrix} SS_{11} & SS_{12} & SS_{13} \\ SS_{21} & SS_{22} & SS_{23} \\ SS_{31} & SS_{32} & SS_{33} \end{pmatrix}$$

$$\begin{pmatrix} d_{11} & \cdots & d_{n1} \\ \vdots & \ddots & \vdots \\ d_{1p} & \cdots & d_{np} \end{pmatrix} \begin{pmatrix} d_{1y} \\ \vdots \\ d_{ny} \end{pmatrix} = \begin{pmatrix} SS_{1y} \\ SS_{2y} \\ SS_{3y} \end{pmatrix}$$

Calculation of determinant

$$|\boldsymbol{S}| = \begin{vmatrix} SS_{11} & SS_{12} & SS_{13} \\ SS_{21} & SS_{22} & SS_{23} \\ SS_{31} & SS_{32} & SS_{33} \end{vmatrix}$$

$$= SS_{11}SS_{22}SS_{33} + SS_{12}SS_{23}SS_{31} + SS_{13}SS_{21}SS_{32} - SS_{13}SS_{22}SS_{31} - SS_{12}SS_{21}SS_{33} - SS_{11}SS_{23}SS_{32}$$

$$= SS_{11}SS_{22}SS_{33} + SS_{12}SS_{23}SS_{31} + SS_{13}SS_{21}SS_{32} - SS_{11}{SS_{23}}^2 - SS_{22}{SS_{13}}^2 - SS_{33}{SS_{12}}^2$$

Cofactor matrix $\tilde{\boldsymbol{S}}$

$$\tilde{\boldsymbol{S}} = \begin{pmatrix} \begin{vmatrix} SS_{22} & SS_{23} \\ SS_{32} & SS_{33} \end{vmatrix} & -\begin{vmatrix} SS_{21} & SS_{23} \\ SS_{31} & SS_{33} \end{vmatrix} & \begin{vmatrix} SS_{21} & SS_{22} \\ SS_{31} & SS_{32} \end{vmatrix} \\ -\begin{vmatrix} SS_{12} & SS_{13} \\ SS_{32} & SS_{33} \end{vmatrix} & \begin{vmatrix} SS_{11} & SS_{13} \\ SS_{31} & SS_{33} \end{vmatrix} & -\begin{vmatrix} SS_{11} & SS_{12} \\ SS_{31} & SS_{32} \end{vmatrix} \\ \begin{vmatrix} SS_{12} & SS_{13} \\ SS_{22} & SS_{23} \end{vmatrix} & -\begin{vmatrix} SS_{11} & SS_{13} \\ SS_{21} & SS_{23} \end{vmatrix} & \begin{vmatrix} SS_{11} & SS_{12} \\ SS_{21} & SS_{22} \end{vmatrix} \end{pmatrix}'$$

Matrix $\boldsymbol{S}$ is symmetric

$$\tilde{\boldsymbol{S}} = \tilde{\boldsymbol{S}}^{\boldsymbol{T}}$$

$$\boldsymbol{S}^{\boldsymbol{-1}} = \frac{\tilde{\boldsymbol{S}}}{\begin{vmatrix} SS_{11} & SS_{12} & SS_{13} \\ SS_{21} & SS_{22} & SS_{23} \\ SS_{31} & SS_{32} & SS_{33} \end{vmatrix}}$$

$$S^{-1} = \frac{\begin{pmatrix} \begin{vmatrix} SS_{22} & SS_{23} \\ SS_{32} & SS_{33} \end{vmatrix} & -\begin{vmatrix} SS_{12} & SS_{13} \\ SS_{32} & SS_{33} \end{vmatrix} & \begin{vmatrix} SS_{12} & SS_{13} \\ SS_{22} & SS_{23} \end{vmatrix} \\ -\begin{vmatrix} SS_{21} & SS_{23} \\ SS_{31} & SS_{33} \end{vmatrix} & \begin{vmatrix} SS_{11} & SS_{13} \\ SS_{31} & SS_{33} \end{vmatrix} & -\begin{vmatrix} SS_{11} & SS_{13} \\ SS_{21} & SS_{23} \end{vmatrix} \\ \begin{vmatrix} SS_{21} & SS_{22} \\ SS_{31} & SS_{32} \end{vmatrix} & -\begin{vmatrix} SS_{11} & SS_{12} \\ SS_{31} & SS_{32} \end{vmatrix} & \begin{vmatrix} SS_{11} & SS_{12} \\ SS_{21} & SS_{22} \end{vmatrix} \end{pmatrix}}{\begin{vmatrix} SS_{11} & SS_{12} & SS_{13} \\ SS_{21} & SS_{22} & SS_{23} \\ SS_{31} & SS_{32} & SS_{33} \end{vmatrix}}$$

$$\begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix} = \boldsymbol{S}^{\boldsymbol{-1}} \begin{pmatrix} SS_{1y} \\ SS_{2y} \\ SS_{3y} \end{pmatrix}$$

$$\begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix} = \frac{1}{\begin{vmatrix} SS_{11} & SS_{12} & SS_{13} \\ SS_{21} & SS_{22} & SS_{23} \\ SS_{31} & SS_{32} & SS_{33} \end{vmatrix}} \begin{pmatrix} \begin{vmatrix} SS_{22} & SS_{23} \\ SS_{32} & SS_{33} \end{vmatrix} & -\begin{vmatrix} SS_{12} & SS_{13} \\ SS_{32} & SS_{33} \end{vmatrix} & \begin{vmatrix} SS_{12} & SS_{13} \\ SS_{22} & SS_{23} \end{vmatrix} \\ -\begin{vmatrix} SS_{21} & SS_{23} \\ SS_{31} & SS_{33} \end{vmatrix} & \begin{vmatrix} SS_{11} & SS_{13} \\ SS_{31} & SS_{33} \end{vmatrix} & -\begin{vmatrix} SS_{11} & SS_{13} \\ SS_{21} & SS_{23} \end{vmatrix} \\ \begin{vmatrix} SS_{21} & SS_{22} \\ SS_{31} & SS_{32} \end{vmatrix} & -\begin{vmatrix} SS_{11} & SS_{12} \\ SS_{31} & SS_{32} \end{vmatrix} & \begin{vmatrix} SS_{11} & SS_{12} \\ SS_{21} & SS_{22} \end{vmatrix} \end{pmatrix} \begin{pmatrix} SS_{1y} \\ SS_{2y} \\ SS_{3y} \end{pmatrix}$$

$$a_1 = \frac{\begin{vmatrix} SS_{22} & SS_{23} \\ SS_{32} & SS_{33} \end{vmatrix} SS_{1y} - \begin{vmatrix} SS_{12} & SS_{13} \\ SS_{32} & SS_{33} \end{vmatrix} SS_{2y} + \begin{vmatrix} SS_{12} & SS_{13} \\ SS_{22} & SS_{23} \end{vmatrix} SS_{3y}}{SS_{11}SS_{22}SS_{33} + SS_{12}SS_{23}SS_{31} + SS_{13}SS_{21}SS_{32} - SS_{11}{SS_{23}}^2 - SS_{22}{SS_{13}}^2 - SS_{33}{SS_{12}}^2}$$

$$a_2 = \frac{-\begin{vmatrix} SS_{21} & SS_{23} \\ SS_{31} & SS_{33} \end{vmatrix} SS_{1y} + \begin{vmatrix} SS_{11} & SS_{13} \\ SS_{31} & SS_{33} \end{vmatrix} SS_{2y} - \begin{vmatrix} SS_{11} & SS_{13} \\ SS_{21} & SS_{23} \end{vmatrix} SS_{3y}}{SS_{11}SS_{22}SS_{33} + SS_{12}SS_{23}SS_{31} + SS_{13}SS_{21}SS_{32} - SS_{11}{SS_{23}}^2 - SS_{22}{SS_{13}}^2 - SS_{33}{SS_{12}}^2}$$

$$a_3 = \frac{\begin{vmatrix} SS_{21} & SS_{22} \\ SS_{31} & SS_{32} \end{vmatrix} SS_{1y} - \begin{vmatrix} SS_{11} & SS_{12} \\ SS_{31} & SS_{32} \end{vmatrix} SS_{2y} + \begin{vmatrix} SS_{11} & SS_{12} \\ SS_{21} & SS_{22} \end{vmatrix} SS_{3y}}{SS_{11}SS_{22}SS_{33} + SS_{12}SS_{23}SS_{31} + SS_{13}SS_{21}SS_{32} - SS_{11}{SS_{23}}^2 - SS_{22}{SS_{13}}^2 - SS_{33}{SS_{12}}^2}$$

The author proposes following notation system

$$S^{-1} = \frac{\begin{pmatrix} \begin{vmatrix} SS_{22} & SS_{23} \\ SS_{32} & SS_{33} \end{vmatrix} & -\begin{vmatrix} SS_{12} & SS_{13} \\ SS_{32} & SS_{33} \end{vmatrix} & \begin{vmatrix} SS_{12} & SS_{13} \\ SS_{22} & SS_{23} \end{vmatrix} \\ -\begin{vmatrix} SS_{21} & SS_{23} \\ SS_{31} & SS_{33} \end{vmatrix} & \begin{vmatrix} SS_{11} & SS_{13} \\ SS_{31} & SS_{33} \end{vmatrix} & -\begin{vmatrix} SS_{11} & SS_{13} \\ SS_{21} & SS_{23} \end{vmatrix} \\ \begin{vmatrix} SS_{21} & SS_{22} \\ SS_{31} & SS_{32} \end{vmatrix} & -\begin{vmatrix} SS_{11} & SS_{12} \\ SS_{31} & SS_{32} \end{vmatrix} & \begin{vmatrix} SS_{11} & SS_{12} \\ SS_{21} & SS_{22} \end{vmatrix} \end{pmatrix}}{\begin{vmatrix} SS_{11} & SS_{12} & SS_{13} \\ SS_{21} & SS_{22} & SS_{23} \\ SS_{31} & SS_{32} & SS_{33} \end{vmatrix}}$$

$$= \frac{\begin{pmatrix} \begin{vmatrix} SS_{22} & SS_{23} \\ SS_{32} & SS_{33} \end{vmatrix} & -\begin{vmatrix} SS_{12} & SS_{13} \\ SS_{32} & SS_{33} \end{vmatrix} & \begin{vmatrix} SS_{12} & SS_{13} \\ SS_{22} & SS_{23} \end{vmatrix} \\ -\begin{vmatrix} SS_{21} & SS_{23} \\ SS_{31} & SS_{33} \end{vmatrix} & \begin{vmatrix} SS_{11} & SS_{13} \\ SS_{31} & SS_{33} \end{vmatrix} & -\begin{vmatrix} SS_{11} & SS_{13} \\ SS_{21} & SS_{23} \end{vmatrix} \\ \begin{vmatrix} SS_{21} & SS_{22} \\ SS_{31} & SS_{32} \end{vmatrix} & -\begin{vmatrix} SS_{11} & SS_{12} \\ SS_{31} & SS_{32} \end{vmatrix} & \begin{vmatrix} SS_{11} & SS_{12} \\ SS_{21} & SS_{22} \end{vmatrix} \end{pmatrix}}{|\boldsymbol{S}|}$$

$$= \begin{pmatrix} \frac{\begin{vmatrix} SS_{22} & SS_{23} \\ SS_{32} & SS_{33} \end{vmatrix}}{|\boldsymbol{S}|} & -\frac{\begin{vmatrix} SS_{12} & SS_{13} \\ SS_{32} & SS_{33} \end{vmatrix}}{|\boldsymbol{S}|} & \frac{\begin{vmatrix} SS_{12} & SS_{13} \\ SS_{22} & SS_{23} \end{vmatrix}}{|\boldsymbol{S}|} \\ \frac{-\begin{vmatrix} SS_{21} & SS_{23} \\ SS_{31} & SS_{33} \end{vmatrix}}{|\boldsymbol{S}|} & \frac{\begin{vmatrix} SS_{11} & SS_{13} \\ SS_{31} & SS_{33} \end{vmatrix}}{|\boldsymbol{S}|} & -\frac{\begin{vmatrix} SS_{11} & SS_{13} \\ SS_{21} & SS_{23} \end{vmatrix}}{|\boldsymbol{S}|} \\ \frac{\begin{vmatrix} SS_{21} & SS_{22} \\ SS_{31} & SS_{32} \end{vmatrix}}{|\boldsymbol{S}|} & -\frac{\begin{vmatrix} SS_{11} & SS_{12} \\ SS_{31} & SS_{32} \end{vmatrix}}{|\boldsymbol{S}|} & \frac{\begin{vmatrix} SS_{11} & SS_{12} \\ SS_{21} & SS_{22} \end{vmatrix}}{|\boldsymbol{S}|} \end{pmatrix}$$

$$= \begin{pmatrix} SS^{11} & SS^{12} & SS^{13} \\ SS^{21} & SS^{22} & SS^{23} \\ SS^{31} & SS^{32} & SS^{33} \end{pmatrix}$$

$$SS^{ij} = \frac{i,j\ cofactor\ of\ S}{|\boldsymbol{S}|}$$

$$SS^{ij} = SS^{ji}$$

$$\begin{pmatrix} SS^{11} & \cdots & SS^{1p} \\ \vdots & \ddots & \vdots \\ SS^{p1} & \cdots & SS^{pp} \end{pmatrix} \begin{pmatrix} SS_{11} & \cdots & SS_{1p} \\ \vdots & \ddots & \vdots \\ SS_{p1} & \cdots & SS_{pp} \end{pmatrix} = \boldsymbol{I}$$

Using this notation system

$$a_1 = SS^{11}SS_{1y} + SS^{12}SS_{2y} + SS^{13}SS_{3y}$$
$$a_2 = SS^{21}SS_{1y} + SS^{22}SS_{2y} + SS^{23}SS_{3y}$$
$$a_3 = SS^{31}SS_{1y} + SS^{32}SS_{2y} + SS^{33}SS_{3y}$$

VI-1-1-4. Significance of regression

Estimation of optimum $\boldsymbol{A}$ is not so complicated, however we should consider the significance of the estimation and should discuss separation of variances for significance test.

The relation among observed y, estimated $\hat{y}$ and error is simple

$$y = \hat{y} + e$$

Here, we consider data standardized by mean. When we calculate square of y.

$$y^2 = (\hat{y} + e)^2 = \hat{y}^2 + 2\hat{y}e + e^2$$

Sum of square is as follow

$$\sum_{j=1}^{n} {y_j}^2 == \sum_{j=1}^{n} {\hat{y}_j}^2 + 2\sum_{j=1}^{n} \hat{y}_j e_j + \sum_{j=1}^{n} {e_j}^2$$

When second term or right side is 0, we separate Sums of square as follows

$$\text{Sum of square of objective value: } \mathrm{SS}_y = \sum_{j=1}^{n} {y_j}^2$$

$$\text{Sum of square of expactation value } \mathrm{SS}_{\hat{y}} = \sum_{j=1}^{n} {\hat{y}_j}^2$$

$$\text{Sum of square of error } \mathrm{SS}_e = \sum_{j=1}^{n} {e_j}^2$$

We expect $\sum_{j=1}^{n} \hat{y}_j e_j$ to be 0, though we cannot conclude $\sum_{j=1}^{n} \hat{y}_j e_j = 0$ from the formula

Proof $\sum_{j=1}^{n} \hat{y}_j e_j = 0$

Using pseudo inverse matrix

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

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

$$\boldsymbol{Y} = \boldsymbol{I}\,\boldsymbol{Y}$$

$\boldsymbol{I}$: unit vector

$$\boldsymbol{E} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}} = (\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\boldsymbol{T}})\boldsymbol{Y}$$

$$\boldsymbol{X^T E} = \boldsymbol{X^T}(\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X^T X})^{-1}\boldsymbol{X^T})\boldsymbol{Y} = \boldsymbol{0}$$

$$\because \boldsymbol{X^T}(\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X^T X})^{-1}\boldsymbol{X^T}) = \boldsymbol{X^T} - \boldsymbol{X^T X}(\boldsymbol{X^T X})^{-1}\boldsymbol{X^T} = \boldsymbol{X^T} - \boldsymbol{X^T} = \boldsymbol{0}$$

$$\boldsymbol{A^T X^T E} = \boldsymbol{0}$$

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

$$\boldsymbol{Y^T E} = \boldsymbol{0}$$

$$\sum_{j=1}^{n} \hat{y}_j e_j = \boldsymbol{Y^T E} = 0$$

$$\mathrm{Q.E.D}$$

We can conclude that

$$\mathrm{SS}_y = \mathrm{SS}_{\hat{y}} + \mathrm{SS}_e$$

$$\mathrm{SS}_y = \sum_{j=1}^{n} {y_j}^2, \quad \mathrm{SS}_{\hat{y}} = \sum_{j=1}^{n} {\hat{y}_j}^2, \quad \mathrm{SS}_e = \sum_{j=1}^{n} {e_j}^2$$

In the paragraphVI-1-1-2. Geometric meaning of multiple linear regression, we proved

$$\mathrm{SS}_y = \mathrm{SS}_{\hat{y}} + \mathrm{SS}_e$$

$$\mathrm{SS}_y = \sum_{j=1}^{n} {y_j}^2, \quad \mathrm{SS}_{\hat{y}} = \sum_{j=1}^{n} {\hat{y}_j}^2, \quad \mathrm{SS}_e = \sum_{j=1}^{n} {e_j}^2$$

Degree of freedom of total SS is $n-1$, degree of freedom of SS_e is $n-p-1$, and degree of freedom of regression is p.

From this, we can summarize the result of multiple regression as Table 43.

Table 43. Summary of result of multiple regression analysis

factor	SS	degree od freedom	variance (V)	ratio（F）
total	SS_{yy}	$n-1$	$V_t = \frac{SS_R}{n-1}$	
regression	SS_R	p	$V_R = \frac{SS_R}{p}$	$\mathrm{F}_0 = \frac{V_R}{V_e}$
residual	SS_e	$n-p-1$	$V_e = \frac{SS_R}{n-p-1}$	

We can apply F test to this result. Null hypothesis of this analysis is

$$a_1 = a_2 = \cdots = a_p = 0$$

Personally, the author is thinking that the analysis has little practical meaning, because we usually do not implement multiple variance analysis among factors which has unlikely possibility of no relation and we do not get any merit when the null hypothesis is rejected. Generally, we want to know power of explanation of the regression expression. For this purpose, we use coefficient of determination. Coefficient of determination is ration of explained variance in total variance.

$$\mathrm{R}^2 = \frac{\mathrm{SS}_{\hat{y}}}{\mathrm{SS}_y} = \frac{\sum_{j=1}^{n} {y_j}^2}{\sum_{j=1}^{n} {y_j}^2} = \frac{\mathrm{SS}_y - \mathrm{SS}_{\hat{e}}}{\mathrm{SS}_y} = 1 - \frac{\mathrm{SS}_{\hat{e}}}{\mathrm{SS}_y}$$

$$\mathrm{R} = \sqrt{\frac{\mathrm{SS}_{\hat{y}}}{\mathrm{SS}_y}} = \sqrt{\frac{\sum_{j=1}^{n} {y_j}^2}{\sum_{j=1}^{n} {y_j}^2}}$$

R^2: coefficent of determination

R: multiple correlation coefficent

Multiple correlation is intuitively understandable indicator of availability of regression expression. However, this indicator has mathematical weakness. The multiple correlation coefficient increases with increase of number of factors, and it reaches 1, when the number of factors comes into the number of data. Because, the data set matrix is regular, when $n = p$. As an example when we gather 10 people and ask them the money they have, then measure the length of ten fingers, the multiple regression analysis of the relation between amount of the money and length of ten fingers gives $\mathrm{R}^2 = 1$. Degree of freedom adjusted determinant coefficient of determination is recommendable as evaluation of availability of regression expression. Adjusted coefficient of determination uses ratio of variances of total and error.

$$v_e = \frac{\mathrm{SS}_{\hat{e}}}{(n-1-p)}$$

$$v_t = \frac{\mathrm{SS}_y}{(n-1)}$$

$$\mathrm{R}^2{}_{adj} = 1 - \frac{v_e}{v_t} = 1 - \frac{\mathrm{SS}_{\hat{e}}}{\mathrm{SS}_y} \frac{(n-1)}{(n-1-p)}$$

$\mathrm{R}^2{}_{adj}$: Degree of freedom adjusted determinant coefficient

p: In this, case p include constant term. number of explanatory variance is $p - 1$

In many cases, we want to know availability or importance of each explanatory variables. One important purpose of multiple variance analysis is simple explanation of phenomenon. It is better to cut off meaningless variables. Simply, we compare absolute values of regression coefficient. However, generally we use different unit of measurement among variables such as m cm mm, dollar, kg, ton number of pieces and so on. When we compare the value of coefficient between variables measured by mm and cm, the value of coefficient is 10 times higher in the variable measured by mm. Generally, we cannot compare values of coefficient directly. However, such comparison is possible when all variables have same variances. One possible ide of such analysis is to standardize the data by dividing all value by deviation of each variable. The author does not recommend this approach. Analytical method should be selected by analyst, because only analyst knows purpose of analysis. In the case when we analyze happiness of the people and items of family expenditure, all explanatory variables are expressed in amount of money. In this case several items such as expenditures for entertainment and culture has large variances though expenditure for food is stable. However, amount of expenditure has important meaning, in such case standardization by deviation is not recommendable.

T test of coefficients of variables can be used for selection of variables. As in explanation in analysis of variance, T test is comparison between observed t and stochastically calculated critical value of t. Null hypothesis is that 0 is included in error range of estimated coefficient.

So, t value of coefficient is obtained by dividing distance from 0 to coefficient a_i by standard error of the coefficient SE_i.

$$\mathrm{t} = \frac{a_i - 0}{\mathrm{SE}_i}$$

We consider SE_i

The total standard error is

$$SE_{tota}\sqrt{\frac{v_e}{n}}$$

$$v_e = \frac{\mathrm{SS}_{\hat{e}}}{(n-1-p)}$$

Here, we apply our local notation system.

$$\begin{pmatrix} SS^{11} & \cdots & SS^{1p} \\ \vdots & \ddots & \vdots \\ SS^{p1} & \cdots & SS^{pp} \end{pmatrix}\begin{pmatrix} SS_{11} & \cdots & SS_{13} \\ \vdots & \ddots & \vdots \\ SS_{31} & \cdots & SS_{33} \end{pmatrix} = \boldsymbol{I}$$

When we consider $\begin{pmatrix} SS^{11} & \cdots & SS^{1p} \\ \vdots & \ddots & \vdots \\ SS^{p1} & \cdots & SS^{pp} \end{pmatrix}$ is sharing of total standard error.

Among the factors in the matrix, $(SS^{i1} \;\cdots\; SS^{ip})$ is factors determining a_i . $SS^{i1} \;\cdots\; SS^{ip}$ are the ratio of determinant of $\boldsymbol{S}$ and cofactor. SS^{ii} means factor determining a_i but independent from SS_{11}. This is the ratio of variance of error of a_i in the total variance.

$$SE_i = \sqrt{SS^{ii}\frac{v_e}{n}}$$

$$\mathrm{t} = \frac{a_i}{\sqrt{SS^{ii}\frac{v_e}{n}}}$$

Table 44 is an example of summary of result of the t tests.

Variable	Coefficient	Standard error	t	P value
x_1	a_1	$\sqrt{SS^{11}\frac{v_e}{n}}$	$\frac{a_1}{\sqrt{SS^{11}\frac{v_e}{n}}}$	P_1
x_2	a_2	$\sqrt{SS^{22}\frac{v_e}{n}}$	$\frac{a_2}{\sqrt{SS^{22}\frac{v_e}{n}}}$	P_2
			$\vdots$	
x_p	a_2	$\sqrt{SS^{pp}\frac{v_e}{n}}$	$\frac{a_2}{\sqrt{SS^{pp}\frac{v_e}{n}}}$	P_p

P can be obtained from computer software or tables in statistic books.

Using results of t test, we can get useful information for the selection of significant variables. However, there are effects of combination of variables. When the purpose of multiple variance analysis is making mathematical model, we must test various combination of variables. There are several systems of searching method of combination of variances.

When the number of candidate variable Including combination effect is small, we can implement round robin test. However, it is not realistic, when the number of candidates is large. We rise and fall the number of variables in the combination of variables by fixed rule and indicators in such case.

Forward selection method: Starting from simple linear regression between objective variable and the explanatory variable which shows smallest P in t test. Then we add next explanatory variable which shows second smallest P observing degree of freedom adjusted determinant coefficient. There are several indicators of explanatory poser of regressed equation. Here we use degree of freedom adjusted determinant coefficient. When the coefficient increases, we accept the variables as explanatory variable in the model. Then repeat same operation inputting variables which shows next smallest P to reach at stable phase in degree of freedom adjusted determinant coefficient.

Backward selection method: Starting from multiple variance linear regression between objective variable and all candidate of variables. The remove variable which shows largest P in t test. Repeat same operation until reaching previously determined threshold of freedom adjusted determinant coefficient.

Stepwise selection method: First operation starts from multivariance analysis of fixed number of variables. We can use the result of diagonalization for determination of starting point. We can determine the rank from eigenvalues considering small eigenvalues are practically 0.

Then add the variable when the statistic value increase with addition of new explanatory variable. If there exists variable whose statistic value is decreased, we can kick out the variable from the model.

There are several statistical indicators including Akaike's information criteria, and there are several manual books of selection method. Readers who want to use multiple linear regression for modeling, please refer those manuals.

Another mathematical weakness of multiple linear regression is existence of multicollinearity. Multicollinearity is phenomenon in which several explanatory variables have correlation. When there exists multicollinearity, the result of regression is unstable, because variables which have correlation attract coefficient among them and we cannot fix the model by exploratory method. When there exists multicollinearity, we should select a representative variance from the variables which have multicollinearity or have to make a synthesis variable from the variables. In the operation of multiple linear regression, we make variance covariance matrix $(\boldsymbol{X}^T\boldsymbol{X})$. In this process we can check correlation among explanatory variances. If is not enough, we can make correlation matrix from variance and covariance matrix. The author recommends checking of correlation matrix among explanatory variables before multiple linear regression. Another approach is doing principle component analysis (PCA) among explanatory variables before multiple linear regression. We can check multicollinearity among variables, and we can use principle component score as synthesis variable in some cases. Mathematically this method is strong, However, we cannot understand scientific meaning of the principle component (in another word, we cannot give a name to the component) in many cases when the meaning of the component is abstract. The other merit of PCA is that we can find latent factors by PCA in some case.
When we collect record of 100-meter sprint and physical measurement of runners (body height, body weight, sitting height, bust measurement and age) from all ages. When we want to know relation between sprint speed and physical characteristics, multiple linear regression is a useful method. However, all data of physical measurement has multicollinearity, because all physical measurement is strongly related to body size. We cannot get any useful information from multiple linear regression directly using row data other than that sprint speed has strongly related with body size. We can expect to find latent component by PCA. Because, we can easily suppose body size will be the first principle component and other components are orthogonal (Independent) to body size. We can estimate factors practically relating component from eigenvalues. Using

knowledge and experiences in the field, we consider the meaning of following components. It may be obesity, relative length of legs, mass of muscle and so on. Then we make synthesis functions such as followings

$$\text{Body mass index: BMI} = \frac{body\ weigt}{(body\ height)^3}$$

$$\text{Relative length of legs: } 1 - \frac{sitting\ heigh}{body\ heig}$$

$$\text{Relative muscle ratio: } \frac{Bust\ measurement}{BMI}$$

Making two-dimensional scatter graphs of PCA, we can consider useful synthesis variable. Then we implement multiple linear regression using synthesis variable and body height as the representative variable of body size.

When number of variables are small, it is not recommendable to use computerized selection system of variables in automatic manner. The author has experience to read a draft of master thesis concerning modeling of relation between evaluated taste of cheese by organoleptic test and physical nature of cheese using multiple linear regression and stepwise selection method of variables. She concluded that most important physical nature is hardness of cheese. The author evaluates her conclusion is not correct or insufficient, because hardness and amount of amino acids in the cheese are result of aging. We should conclude that people joined in organoleptic test could detect difference in amount or combination of amino acids.

Implication of this episode is risk of automatous use of multiple linear regression and risk of blind acceptance of computerized system. We have to interpret the result of multivariable liner regression by our knowledge and experiences and evaluate applicability of selected variables in various aspects by ourselves.