KMITL Sci. J. Vol. 5 No. 1 Feb. 2005

On the statistical data management by the multiple regression analysis

Katsumi UJIIE[*1] and Eiji ITOH[*2]

[*1] Research Institute of Education, Tokai University,
Hiratsuka, Kanagawa, 259-1292, Japan
E-mail : ujiie@keyaki.cc.u-tokai.ac.jp

[*2] Department of Sports leisure management,
School of Physical Education, Tokai University,
Hiratsuka, Kanagawa, 259-1292, Japan
E-mail : eiji@keyaki.cc.u-tokai.ac.jp

Abstract

Regression analysis is one of the analyses to predict some variable y which is influenced by the data based on other variables $x_1, x_2, \cdots, x_p$ are as

$$a_0 + a_1 x_1 + \cdots + a_p x_p \xrightarrow{\text{prediction}} y.$$

This analysis is used as the factorial experiment with regard to examine the variables' contributing the prediction. In this study, we construct the calculation method of the linear multiple regression equation and the multiple correlation coefficient needs this prediction. Furthermore we carry out the prediction and the investigation based on the data of 30 students players of Baseball club in Tokai University as an application example.

Key Words and Phrases : prediction; factorial experiment; linear multiple regression equation; multiple correlation coefficient

1. Introduction

When we derive the relation equation between some variable y(criterion variable) and the variables $x_1, x_2, \cdots, x_p$(explanatory variables) influenced y, we predict the value of y from the values of $x_1, x_2, \cdots, x_p$ or evaluate the influence of each x. We call a simple regression analysis and a multiple regression analysis in case of the number of y is more than 2. Here we describe the multiple regression analysis in case of the number of explanatory variable is p.

In order to predict the value of the criterion variable y from the explanatory variables $x_1, x_2, \cdots, x_p$, we suppose the following relation between y and $x_1, x_2, \cdots, x_p$ by using a function f as

$$y_i = f(x_{1i}, x_{2i}, \cdots, x_{pi}) + e_i, \quad (i = 1, 2, \cdots, n),$$

where e_i is the error term that can not explain by the values $x_{1i}, x_{2i}, \cdots, x_{pi}$ of the explanatory variables. The form of the function f, e.g. the relation between y and $x_1, x_2, \cdots, x_p$ is generally unknown.

In the multiple regression analysis, we consider the linear equation of $x_{1i}, x_{2i}, \cdots, x_{pi}$ as $f(x_{1i}, x_{2i}, \cdots, x_{pi})$ and suppose the following linear multiple regression model

$$y_i = a_0 + a_1 x_{1i} + \cdots + a_p x_{pi} + e_i, \quad (i = 1, 2, \cdots, n),$$

where $a_1, a_2, \cdots, a_p$ are the regression coefficients and a_0 is a constant term.

In this study, we derive the multivariate regression equation in Chapter 2, the multiple correlation coefficient in Chapter 3. Furthermore we apply this analysis to two examples in Chapter 4.

2. Multiple regression equation

The multiple regression equation is used to predict the criterion variable from the explanatory variable or examine the relation(correlation) between the criterion variable and explanatory variable. Here we detail about linear regression from linear regression and nonlinear regression in the Multiple regression equation.

Now we give an criterion variable y and the n data(observations) of p explanatory variables $x_1, x_2, \cdots, x_p$ in Table 2.1.

Table 2.1 Criterion variable and explanatory variable

Data.No.	Criterion variable	Explanatory variable			
	y	x_1	x_2	$\cdots$	x_p
1	y_1	x_{11}	x_{21}	$\cdots$	x_{p1}
2	y_2	x_{12}	x_{22}	$\cdots$	x_{p2}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
i	y_i	x_{1i}	x_{2i}	$\cdots$	x_{pi}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
n	y_n	x_{1n}	x_{2n}	$\cdots$	x_{pn}

Then we construct an equation(prediction equation) to express the relation between $x_1, x_2, \cdots, x_p$ and y to predict the value of y from $x_1, x_2, \cdots, x_p$. Usually we suppose the prediction equation as

$$y_i = a_0 + a_1 x_{1i} + a_2 x_{2i} + \cdots + a_p x_{pi} + e_i \quad (i = 1, 2, \cdots, n) \tag{2.1}$$

Here we explain how to derive the prediction equation from the given data in Table 2.1. In other words, we show how to calculate the coefficient of variable $a_1, a_2, \cdots, a_p$ and the constant a_0 from the data in Table 2.1. Table 2.1 is rewritten by

Table 2.2 Prediction errors

Data.No.	e
1	$e_1 = y_1 - (a_0 + a_1 x_{11} + a_2 x_{21} + \cdots + a_p x_{p1})$
2	$e_2 = y_2 - (a_0 + a_1 x_{12} + a_2 x_{22} + \cdots + a_p x_{p2})$
$\vdots$	$\vdots$
i	$e_i = y_i - (a_0 + a_1 x_{1i} + a_2 x_{2i} + \cdots + a_p x_{pi})$
$\vdots$	$\vdots$
n	$e_n = y_n - (a_0 + a_1 x_{1n} + a_2 x_{2n} + \cdots + a_p x_{pn})$

in terms of the prediction error. Based on (2.1) and Table 2.1, We calculate the values $\hat{a}_0, \hat{a}_1, \cdots, \hat{a}_p$ of $a_0, a_1, \cdots, a_p$ to minimize the sum squares of prediction error(the least squares method)

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \{y_i - (a_0 + a_1 x_{1i} + a_2 x_{2i} + \cdots + a_p x_{pi})\}^2,$$

then

$$\hat{a}_j = \frac{\begin{vmatrix} s_{11} & \cdots & s_{y1} & \cdots & s_{1p} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ s_{j1} & \cdots & s_{yj} & \cdots & s_{jp} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ s_{p1} & \cdots & s_{yp} & \cdots & s_{pp} \end{vmatrix}}{\begin{vmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{vmatrix}}, \qquad (j = 1, 2, \cdots, p), \tag{2.2}$$

$$\hat{a}_0 = \bar{y} - (\hat{a}_1 \bar{x}_1 + \cdots + \hat{a}_p \bar{x}_p), \tag{2.3}$$

where the covariance matrix of $x_1, x_2, \cdots, x_p$ is

$$V = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1l} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2l} & \cdots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ s_{j1} & s_{j2} & \cdots & s_{jl} & \cdots & s_{jp} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_{pl} & \cdots & s_{pp} \end{bmatrix}, \tag{2.4}$$

where

$$s_{jl} = \frac{1}{n}\sum_{i=1}^{n}(x_{ji} - \bar{x}_j)(x_{li} - \bar{x}_l), \quad (j, l = 1, 2, \cdots, p)$$

and the covariance of y and $x_1, x_2, \cdots, x_p$ are

$$\begin{cases} s_{y1} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})(x_{1i} - \bar{x}_1), \\ s_{y2} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})(x_{2i} - \bar{x}_2), \\ \qquad \vdots \\ s_{yp} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})(x_{pi} - \bar{x}_p). \end{cases} \tag{2.5}$$

Thus we can express the prediction equation as

$$y = \hat{a}_0 + \hat{a}_1 x_1 + \cdots + \hat{a}_p x_p \tag{2.6}$$

by using $\hat{a}_1, \hat{a}_2, \cdots, \hat{a}_p, \hat{a}_0$ in (2.2),(2.3). We call this equation as the (linear) multiple regression equation and $\hat{a}_1, \hat{a}_2, \cdots, \hat{a}_p$ as the regression coefficient for the explanatory variables $x_1, x_2, \cdots, x_p$ of the criterion variable y.

This equation applies the straight line to the points of n data inlaid the two-dimensional orthogonal coordinates of x-axis and y-axis in the case of p=1, while this equation applies the p-dimensional hyperplane to the points of n data inlaid the orthogonal coordinates of (p+1)-dimensional space consisted of y-axis, x_1-axis, x_2-axis, $\cdots$, x_p-axis in the case of general p.

3. Multiple correlation coefficient

When we let the prediction value of the criterion variable y_i to obtain the multiple regression equation(2.6),

$$Y_i = \hat{a}_0 + \hat{a}_1 x_{1i} + \hat{a}_2 x_{2i} + \cdots + \hat{a}_p x_{pi} \quad (i = 1, 2, \cdots, n).$$

We call the correlation coefficient(we write r_{yY}) between y_i and Y_i as the multiple correlation coefficient between y and $x_1, x_2, \cdots, x_p$ and write $r_{y\cdot 12\cdots p}$, that is,

Table 3.1 Criterion value, prediction value and prediction error

Criterion y	Prediction value Y	Prediction error e
y_1	Y_1	$e_1 = y_1 - Y_1$
y_2	Y_2	$e_2 = y_2 - Y_2$
$\vdots$	$\vdots$	$\vdots$
y_n	Y_n	$e_n = y_n - Y_n$

$$r_{yY} = \frac{s_{yY}}{\sqrt{s_{yy}s_{YY}}} = r_{y\cdot 12\cdots p}, \tag{3.1}$$

where

$$\begin{cases} s_{yy} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(y_i - \bar{y})^2, & \text{(variance of y)},\\ s_{YY} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(Y_i - \bar{Y})^2, & \text{(variance of Y)},\\ s_{yY} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(y_i - \bar{y})(Y_i - \bar{Y}), & \text{(covariance of y and Y)}. \end{cases}$$

If we use the determination

$$S = \begin{vmatrix} s_{yy} & s_{y1} & s_{y2} & \cdots & s_{yp} \\ s_{1y} & s_{11} & s_{12} & \cdots & s_{1p} \\ s_{2y} & s_{21} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ s_{py} & s_{p1} & s_{p2} & \cdots & s_{pp} \end{vmatrix}, \tag{3.2}$$

we can represent(3.1) as

$$r_{y\cdot 12\cdots p} = \sqrt{1 - \frac{S}{s_{yy}s_{11}}}, \tag{3.3}$$

where s_{11} is the cofactor of 1×1 component of determinant S,that is,

$$S_{11} = \begin{vmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{vmatrix}$$

Then $r_{y\cdot 12\cdots p}$ satisfies

$$0 \leqq r_{y\cdot 12\cdots p} \leqq 1, \tag{3.4}$$

while r_{y1} satisfies $-1 \leqq r_{y1} \leqq 1$.

Here we consider the reason why (3.4) is correct and why the multiple correlation coefficient ($r_{y\cdot 12\cdots p}$) is the scale to represent the strength of the relation between the variable y and the variables $x_1, x_2, \cdots, x_p$.

(Give a deep significance of multiple correlation coefficient)

we realize some things as follows;

When $r_{y\cdot 12\cdots p} = \pm 1, \sum_{i=1}^{n} \tilde{e}_i^2 = 0$. Thus the prediction error is all 0 at $\tilde{e}_1 = \tilde{e}_2 = \cdots = \tilde{e}_p = 0$ and all the point of data are on the regression surface of the $(p+1)$-dimensional space. The value of $\sum_{i=1}^{n} \tilde{e}_i^2$ takes a larger value with $r_{y\cdot 12\cdots p}$ approaches 0 from 1 and the points of the data leave gradually from the regression surface. We call that the multiple correlation between the variable y and the variables $x_1, x_2, \cdots, x_p$ is powerful if $r_{y\cdot 12\cdots p}$ approaches 1 in the meaning of the points of the data line up along the regression surface, while we call that the multiple correlation between the variable y and the variables $x_1, x_2, \cdots ; x_p$ is weak if $r_{y\cdot 12\cdots p}$ approaches 0 in the meaning of the points of the data leave and disperse from the regression surface.

Similarly to the correlation coefficient, we notice that the what the multiple correlation is powerful means the points of the data line up along the regression surface. Usually the value of $r_{y\cdot 12\cdots p}$ dose not approach 1, while $r_{y\cdot 12\cdots p}$ has strongly the curved surface correlation in the case of points of the data do not line up along the surface but line up the curved surface. In this case, we must notice that we can not guess the strength of the relation by the value of the multiple correlation coefficient.

Notice We call the squares of the multiple correlation coefficient;

$$r^2_{y\cdot 12\cdots p} = \frac{s^2_{yY}}{s_{yy}s_{YY}} = \frac{s^2_{YY}}{s_{yy}s_{YY}} = \frac{s_{YY}}{s_{yy}}$$

$$= \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{\text{varience of prediction value}}{\text{varience of value of criterion variable}}$$

as the coefficient of determination or the contribution and we use this as the scale to represent the virtue to apply the regression surface to the data. We consider the magnitude of the sum of the squares(residual sum of the squares) of the prediction error as the scale to represent the virtue to apply the regression surface to the data.

By the sum of the squares of

$$\frac{1}{n}\sum_{i=1}^{n}\tilde{e}_i^2$$

is $s_{yy}(1 - r_{y\cdot 12\cdots p}), r^2_{y\cdot 12\cdots p}$ is used as the scale.

Also we can consider $r_{y\cdot 12\cdots p}$, square root of the coefficient of determination as a scale to represent the virtue to apply the regression surface to the data.

4. The example and a point of view of the analysis results

Example 1

Table 4.1 shows the data of the record for the throwing a ball y(m), grasping power x_1(kg), back muscle power x_2(kg), height x_3(cm), weight x_4(kg), of 30 players(university students) of the baseball club. Then derive the multiple regression equation for x_1, x_2, x_3, x_4, of y.

Table 4.1 Data for 30 players(university students) of the baseball club

Data	Criterion variable	Explanatory variable			
No.	Throwing a ball	Grasping power	Back muscle power	Height	Weight
1	82.0	52.0	170.0	180.0	74.9
2	90.0	59.0	196.0	170.0	68.0
3	89.0	57.0	219.0	164.2	68.0
4	85.0	53.0	192.0	180.0	71.5
5	90.0	61.0	200.0	177.5	75.4
6	100.0	59.0	240.0	176.5	74.5
7	100.0	48.0	205.0	173.0	76.4
8	89.0	47.0	150.0	172.0	66.6
9	90.0	54.0	185.0	180.0	76.4
10	90.0	57.0	190.0	172.0	69.1
11	91.0	62.0	200.0	180.0	80.5
12	95.0	62.0	247.0	185.0	84.2
13	82.0	53.0	240.0	172.0	73.4
14	95.0	56.0	190.0	182.5	82.2
15	86.0	58.0	250.0	183.0	84.3
16	69.0	74.0	180.0	178.0	84.5
17	90.0	44.0	147.0	164.0	55.9
18	95.0	54.0	125.5	175.0	82.4
19	93.0	48.0	155.0	177.0	78.0
20	100.0	56.0	166.0	172.0	68.3
21	92.0	52.0	145.0	171.0	65.0
22	101.0	52.0	150.0	174.0	75.3
23	90.0	50.0	166.5	179.0	82.1
24	95.0	59.0	180.0	184.0	81.7
25	83.0	50.0	176.0	176.0	74.2
26	86.0	62.0	250.0	177.0	82.5
27	81.0	52.0	298.0	181.0	77.5
28	75.0	51.0	178.0	176.0	68.9
29	108.0	62.0	196.0	184.0	78.4
30	105.0	55.0	180.0	184.0	78.4
Tatal	2717.0	1659.0	5767.0	5299.7	2258.5
Mean	90.6	55.3	192.2	176.7	75.3

(Answer)

We use [2] in the References for the calculation and suppose the significant figure as three place. The value of varience-covarience matrix and s_{y1}, s_{y2}, s_{y3}, s_{y4}, on the Table 4.1 are

$$V = \begin{pmatrix} 35.076 & 76.796 & 11.699 & 20.721 \\ 76.796 & 1431.928 & 58.593 & 80.572 \\ 11.699 & 58.593 & 29.335 & 28.532 \\ 20.721 & 80.572 & 28.532 & 45.196 \end{pmatrix},$$

$$s_{y1} = -5.3, \quad s_{y2} = -52.8, \quad s_{y3} = 5.1, \quad s_{y4} = 1.8.$$

We obtain the following simultaneous equations.

$$\begin{cases} 35.076\hat{a}_1 + 76.796\hat{a}_2 + 11.699\hat{a}_3 + 20.721\hat{a}_4 = -5.3, \\ 76.796\hat{a}_1 + 1431.928\hat{a}_2 + 58.593\hat{a}_3 + 80.572\hat{a}_4 = -52.8, \\ 11.699\hat{a}_1 + 58.593\hat{a}_2 + 29.335\hat{a}_3 + 28.532\hat{a}_4 = 5.1, \\ 20.721\hat{a}_1 + 80.572\hat{a}_2 + 28.532\hat{a}_3 + 45.196\hat{a}_4 = 1.8. \end{cases}$$

If we solve this simultaneous equations, we obtain the regression coefficient.

$$\hat{a}_1 = -0.155, \quad \hat{a}_2 = -0.041, \quad \hat{a}_3 = 0.356, \quad \hat{a}_4 = -0.041, \quad \hat{a}_0 = 47.140.$$

Thus we can express the linear multiple regression equations and the multiple correlation coefficient as

$$y = 47.140 - 0.155x_1 - 0.041x_2 + 0.356x_3 - 0.041x_4, \quad r_{y \cdot 1234} = 0.2621$$

respectively.

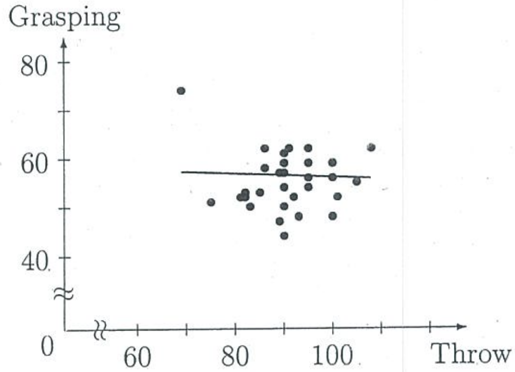

Fig.4.1 Relation between throwing a ball and grasping power

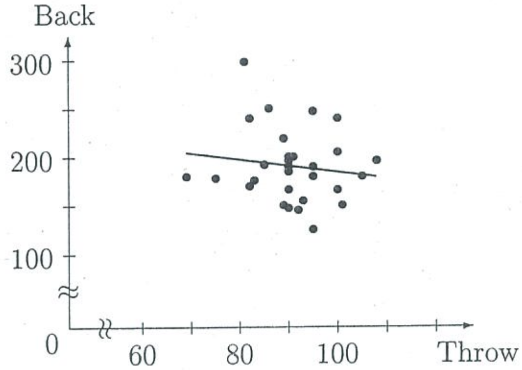

Fig.4.2 Relation between throwing a ball and back muscle power

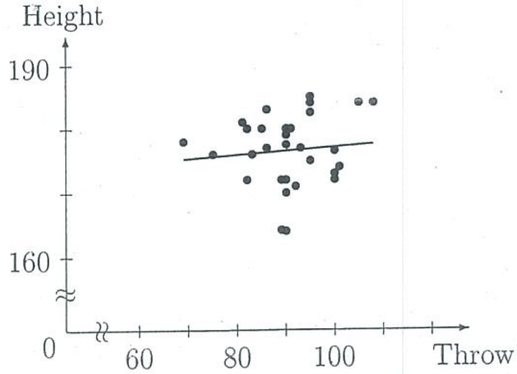

Fig.4.3 Relation between throwing a ball and height

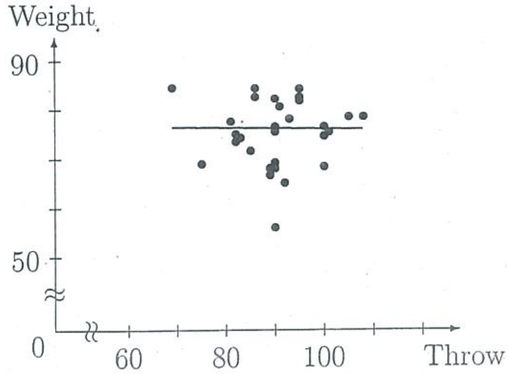

Fig.4.4 Relation between throwing a ball and weight

Example 2

When the value of the explanatory variables x_1, x_2, x_3, x_4, are given by the Table 4.2 for the student number 31, 32, 33, calculate the criterion variable (y) and the prediction value (Y) by the multiple regression equation in example 1.

Table 4.2 Criterion variable and explanatory variable

Data:No.	Criterion variable	Explanatory variable			
	y	x_1	x_2	x_3	x_4
31	?	58.0	170.0	176.0	80.5
32	?	56.0	190.0	185.0	72.5
33	?	50.0	160.0	173.0	78.5

(Answer)

The prediction value of y_{31}, y_{32}, y_{33}, are calculated by using the multiple regression equation (4.2) as

$$\text{Exact value} \quad y_{31} = 93.0, \quad y_{32} = 99.0, \quad y_{33} = 94.0.$$

$$\text{Prediction value} \quad Y_{31} = 90.605, \quad Y_{32} = 93.637, \quad Y_{33} = 91.27.$$

I wonder the prediction method like this comes together with the problem how much the prediction value calculated by this example has the confidence. Here we calculate the multiple regression equation that is the prediction equation by the only data of n=30 in a descriptive statistics point of view. However, we usually consider that the number of n is quite many in prediction and the data of these as the random sample drawn randomly from the population in an inferential statistics point of view. Then we can evaluate randomly the confidence of the prediction value and construct that confidence interval. Also we can discuss the confidence of regression coefficient similarly.

References

[1] Johnson, R. A. and Wichern, D. W. (1988). Applied Multivariate Statistical Analysis, Part III, Fourth Edition, Prentice Hall.

[2] Tanaka, Y. Tarumi, T. and Wakimoto, K. (1994). Handbook of Personal-Computer Statistical Analysis Part II Multivariate Analysis Series, Kyoritsu.Co.

[3] Tanaka, Y. and Wakimoto, K. (1983). Method of Multivariate Statistical Analysis, Gendai Suugakusya.