Multiple Linear Regression Analysis

EGI SAFITRI, S.MAT., M.SI

MULTIPLE REGRESSION AND CORRELATION ANALYSIS

The form of the regression equation with two independent variables is:

$$Y = a + b_1 X_1 + b_2 X_2$$

The regression equation with 3 independent variables is:

$$Y = a + b X_1 + b_2 X_2 + b_3 X_3$$

The general form of the regression equation for k independent variables can be formulated as follows:

$$Y = a + b_1 X_1 + b_2 X_2 + b_3 X_3 + \ldots + b_k X_k$$

Equation to get the regression coefficient

The principle of the ordinary least square (OLS) method is to minimize the sum of squared deviations around the regression line. The values of the regression coefficients a, b1, and b2 can be solved simultaneously.

$$\sum Y \quad = \quad na + \quad b1\sum X1 \quad + \quad b_2\sum X_2 \qquad (1)$$

$$\sum X_1 Y = a\sum X_1 + b_1\sum X_1^2 + b_2\sum X_1\, X_2 \qquad (2)$$

$$\sum X_2 Y = a\sum X_2 + b_1\sum X1\ X2 + b2\sum X_2^2 \qquad (3)$$

EXAMPLE: DEMAND IS AFFECTED BY PRICE AND INCOME

No	Deman (Y)	Price (X_1)	Income (X_2)
1	3	8	10
2	4	7	10
3	5	7	8
4	6	7	5
5	6	6	4
6	7	6	3
7	8	6	2
8	9	6	2
9	10	5	1
10	10	5	1

To get the regression coefficient, it is necessary to calculate

: $\sum Y$, $\sum X_1$, $\sum X_2$, $\sum X_1 Y$, $\sum X_1^2$, $\sum X_1 \sum X_2$, $\sum X_2 Y$, $\sum X_2^2$

Y	X_1	X_2	YX_1	YX_2	X_1^2	X_2^2	X_1X_2
3	8	10	24	30	64	100	80
4	7	10	28	40	49	100	70
5	7	8	35	40	49	64	56
6	7	5	42	30	49	25	35
6	6	4	36	24	36	16	24
7	6	3	42	21	36	9	18
8	6	2	48	16	36	4	12
9	6	2	54	18	36	4	12
10	5	1	50	10	25	1	5
10	5	1	50	10	25	1	5
$\sum Y=68$	$\sum X_1 = 63$	$\sum X_2 = 46$	$\sum X_1 Y=$ 409	$\sum X_2 Y=$ 239	$\sum X_1^2=$ 405	$\sum X_2^2=$ 324	$\sum X_1 \sum X_2 =$ 317

By elimination method

$68 = 10a + 63b1 + 46b_2 \quad \ldots(1)$
$409 = 63a + 405b_1 + 317b_2 \ldots(2)$
$239 = 46a + 317b_1 + 324b_2 \ldots(3)$

To get the value of regression coefficients a, b1, and b2 can be done by Substitution between equations

$-428,4 = -63a - 396,9\ b1 - 289,8b2 \quad \ldots$
$409 = 63a + 405b_1 + 317b_2$ .... (2)
$-19,4 = 0 + 8,1b1 + 27,2b_2$ .... (4)

$-312,8 = -46a - 289,8\ b_1 - 211,6b_2$
$239 = 46a + 317b_1 + 324b_2$ .... (3)
$-73,8 = 0 + 27,2b_1 + 112,4b_2$ .... (5)

From the equation above, the value of b2 is = -8.65/21.06 = -0.41. After finding the value of b2, the value of b1 can be found using equation 4 or 5.

$$-19,4 = 0 + 8,1b1 + 27,2(-0,41) \ldots\ldots\ldots (4)$$

$$19,4 = 8,1b1 - 11,18$$

$$8,1b1 = -19,4 + 11,18$$

$$8,1\ b1 = -\ 8,22$$

$$b1 = -8,22/8,1 = -1,015$$

$$68 = 10a + 63\ (-1,015) + 46(-0,41)\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots.\ (1)$$

$$68 = 10a - 63,96 - 18,90$$

$$10a = 63 + 92,86$$

$$a = 150,86/10 = 15,086$$

By finding the regression coefficient values a, b1, and b2, the regression equation can be expressed as follows:

$$Y = 15,086 - 1,015X_1 - 0,41\ X_2$$

COEFFICIENT OF DETERMINATION FORMULA (R^2)

The coefficient of determination shows a proportion of the variance that can be explained by the regression equation (regression of sum squares, RSS). The coefficient of determination is formulated as follows:

$$R^2 = \frac{(\hat{Y} - \overline{Y})^2}{(Y - \overline{Y})^2}$$

or

$$R^2 = \frac{n(a.\sum Y + b_1.\sum YX_1 + b_2 \Sigma YX_2) - (\sum Y)^2}{n.\sum Y^2 - (\sum Y)^2}$$

$$R^2 = \frac{n(a.\sum Y + b_1.\sum YX_1 + b_2\Sigma YX_2) - (\sum Y)^2}{n.\sum Y^2 - (\sum Y)^2}$$

$$R^2 = \frac{10\left[(15,086)(68) - 1,015(409) - (0,41)(239)\right] - (68)^2}{(10)(516) - (68)^2}$$

$$R^2 = 0,939$$

Simple correlation

$$r_{YX_1} = \frac{n\sum YX_1 - \sum Y \sum X_1}{\sqrt{\left[n\sum Y^2 - \left(\sum Y\right)^2\right].\left[n\sum X_1^2 - \left(\sum X_1\right)^2\right]}}$$

$$r_{YX_2} = \frac{n\sum YX_2 - \sum Y \sum X_2}{\sqrt{\left[n\sum Y^2 - \left(\sum Y\right)^2\right].\left[n\sum X_2^2 - \left(\sum X_2\right)^2\right]}}$$

$$r_{X_1X_2} = \frac{n\sum X_1X_2 - \sum X_1 \sum X_2}{\sqrt{\left[n\sum X_1^2 - \left(\Sigma X_1\right)^2\right].\left[n\sum X_2^2 - \left(\sum X_2\right)^2\right]}}$$

PARTIAL CORRELATION: The relationship of variables assuming other variables remain fixed

$$r_{YX_1.X_2} = \frac{r_{YX_1} - r_{YX_2} r_{X_1X_2}}{\sqrt{(1-r^2{}_{YX_2})(1-r^2{}_{X_1X_2})}}$$

$r_{yX1.X2}$= correlation Y with X_1, where X_2 constant

$$r_{YX_2.X_1} = \frac{r_{YX_2} - r_{YX_1} r_{X_1X_2}}{\sqrt{(1 - r^2_{YX_1})(1 - r^2_{X_1X_2})}}$$

$$r_{X_1X_2.Y} = \frac{r_{X_1X_2} - r_{YX_1} r_{YX_2}}{\sqrt{(1 - r^2_{YX_1})(1 - r^2_{YX_2})}}$$

Standard Error of Estimation

$$S_{Y.X_1X_2} = \sqrt{\frac{\Sigma(\hat{Y} - Y)^2}{n-(k+1)}}$$

S: Standard error of estimating variable Y based on variables X1 and X2

$\hat{Y}$: The expected value of Y where X1 and X2 are known

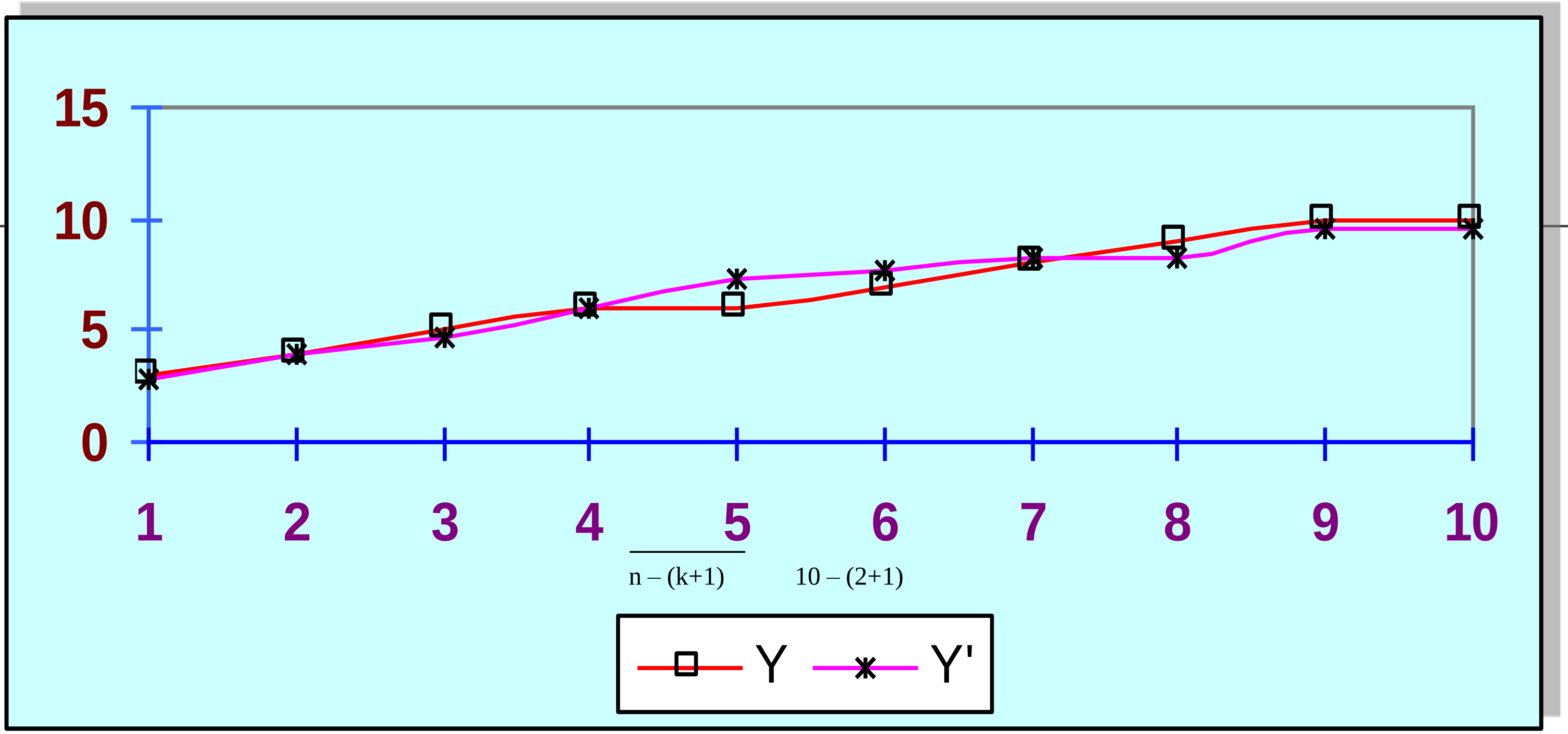
Y : Observed value of Y

N : Number of samples or data

K : Number of independent variables

equation $\hat{Y} = 15{,}086 – 1{,}015\ X_1 – 0{,}41\ X_2$

Y	X_1	X_2	$\hat{Y} = 15{,}086 – 1{,}015\ X_1 – 0{,}41\ X_2$	$(\hat{Y}–Y)$	$(\hat{Y}–Y)^2$
3	8	10	2,86=15,086 – 1,015 (8) – 0,41(10)	0,14	0,02
4	7	10	3,87=15,086 – 1,015 (7) – 0,41(10)	0,13	0,02
5	7	8	4,69=15,086 – 1,015 (7) – 0,41 (8)	0,31	0,09
6	7	5	5,92=15,086 – 1,015 (7) – 0,41 (5)	0,08	0,01
6	6	4	7,35=15,086 – 1,015 (6) – 0,41 (4)	-1,35	1,83
7	6	3	7,76=15,086 – 1,015 (6) – 0,41 (3)	-0,76	0,58
8	6	2	8,17=15,086 – 1,015 (6) – 0,41 (2)	-0,17	0,03
9	6	2	8,17=15,086 – 1,015 (6) – 0,41 (2)	0,83	0,68
10	5	1	9,60=15,086 – 1,015 (5) – 0,41 (1)	0,40	0,16
10	5	1	9,60=15,086 – 1,015 (5) – 0,41 (1)	0,40	0,16
				$\sum (\hat{Y} – Y)^2$	3,58



$$S_{Y.X_1X_2} = \sqrt{\frac{\sum(\hat{Y}-Y)^2}{n-(1+k)}} = \sqrt{\frac{3.58}{10-(2+1)}} = 0,72$$

Or it can also be found with the following formula:

$$S_{Y.X_1X_2} = \sqrt{\frac{\sum Y^2 - a\sum Y - b_1\sum X_1Y - b_2\sum X_2Y}{n-3}}$$

$$S_{Y.X_1X_2} = \sqrt{\frac{516 - (15{,}086 \text{ x } 68) - (-1{,}01524 - 409) - (-0{,}41 \text{ x } 239)}{10-3}}$$

$$S_{Y.X_1X_2} = 0{,}72$$

Standard error of estimation (Sb)

That is how much the estimated values, b1 and b2, are from the true values (population parameters).

$$Sb_1 = \frac{S_{y.x_1x_2}}{\sqrt{(\Sigma X_1^2 - n\bar{X}_1^2)(1 - r_{X_1X_2}^2)}} = 0{,}580$$

$$Sb_2 = \frac{S_{y.x_1x_2}}{\sqrt{(\Sigma X_2^2 - n\bar{X}_2^2)(1 - r_{X_1X_2}^2)}}$$

$$Sb_2 = 0{,}156$$

Partial or individual significance test

$$t_{hit}.b_1 = \frac{b_1 - B_1}{Sb_1} = \frac{-1{,}015 - 0}{0{,}58} = -1{,}75$$

$$t_{hit}.b_2 = \frac{b_2 - B_2}{Sb_2} = \frac{-0{,}410 - 0}{0{,}1558} = -2{,}673$$

Partial Test

In the pasial test, each coefficient t_{hit} for each coefficient is tested, whether it is significant or not

The form of testing is the same as before two-way or one-way

Hypothesis tested:

Ho: b1=0 and

Ha: b2 $\neq$ 0

GLOBAL TEST OR SIMULTANEOUS SIGNIFICANCE TEST (F TEST)

The hypothesis to be tested is the ability of the independent variable to explain the behavior of the independent variable, if the independent variable cannot affect the independent variable, it can be considered that the regression coefficient value is equal to zero, so that any value of the independent variable will not affect the independent variable. For the equation in example one, Y = 15.086 - 1.015X1 - 0.41 X2, the independent variables X1, and X2 are said to be able to affect Y if the coefficient values b1 and b2 are not equal to zero, if they are equal to zero, then it is said that they are not able to affect the independent variable Y.

F Table

The F-table value needs to know the degree of freedom of the numerator in the column, the degree of freedom of the denominator in the row and the real level. It is known that there are three variables, namely Y, X1, and X2, so k = 3, while the number n = 10.

So the degree of numerator k - 1 = 3 - 1 = 2, while the degree of denominator n - k = 10 - 3 = 7 with a real level of 5%. The F-table value with a numerator degree of 2 and a denominator of 7 at a real level of 5% is 4.74.

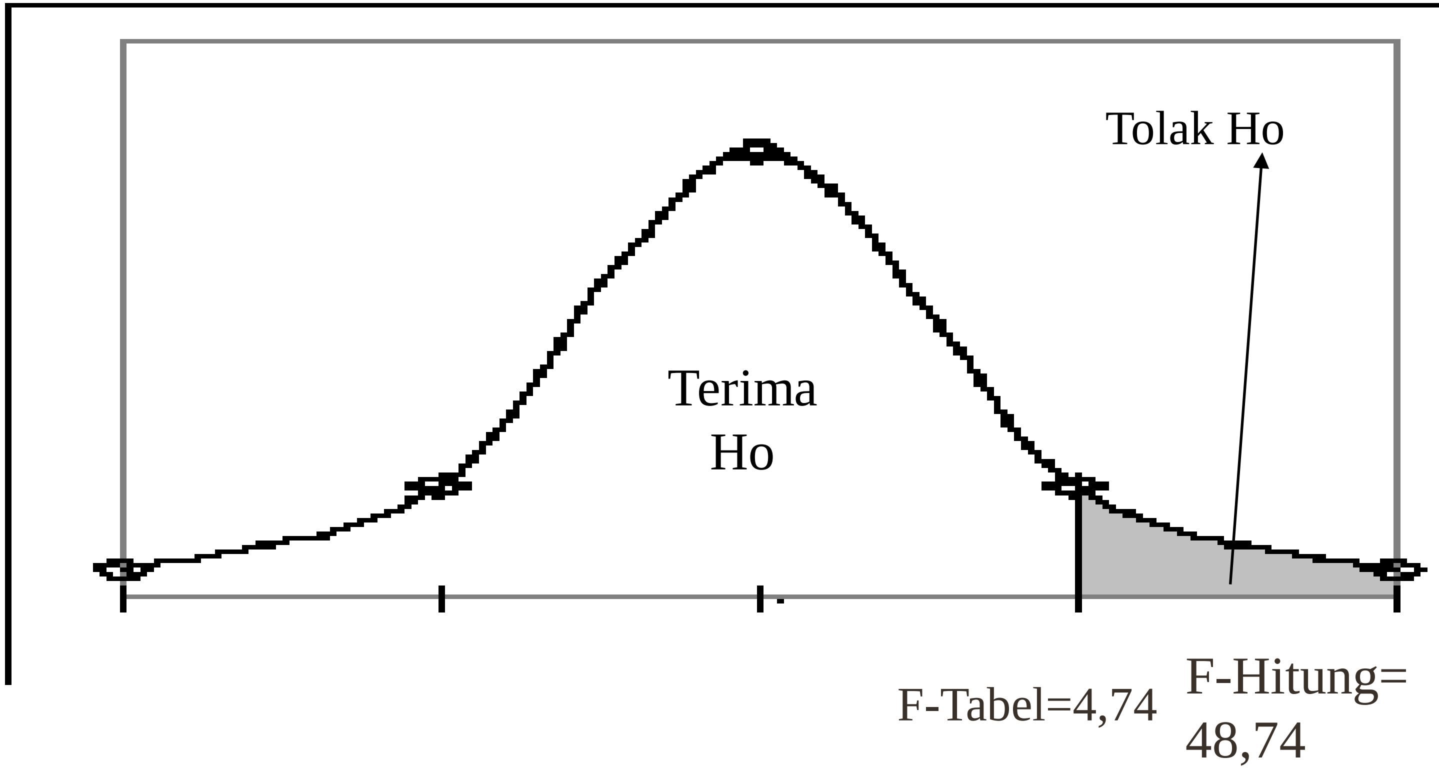
The F-count value is determined by the following formula:

:

$$F = \frac{R^2/(k-1)}{(1-R^2)/(n-3)}$$

$$F = \frac{0,933/(3-1)}{(1-0,933)/(10-3)} = 0,4665/0,0096 = 48,73881$$

GLOBAL TEST OR SIMULTANEOUS SIGNIFICANCE TEST (F TEST)



Multiple Regression printout display

Model	R	R Square	Adjusted R Square	Sts. Eror of the Estimate
1	.647[a]	.419	.379	537549.548

Based on the data The table above shows the results of multiple regression tests:
(1) R square value = 0.419.
(2) Ry12 value = 0.647.
(3) SEE value = 537,549,547.

Model	Unstandardized Coefficients		Standardized Coefficeient	t	Sig.
	B	Std.Error	Beta		
1 (Constant) Harga Biaya promosi	100004 792852.6 3.378	1059097 186021.7 2.785	.752 .214	.944 4.262 1.213	.353 .000 .235

From the data above,

a = 1,000,044.

b1 = 792,852.6.

b2 = 3.378.

The regression equation obtained is as follows:

$$Y = 1{,}000{,}044 + 792{,}852.6\, X1 + 3.378\, X2$$

Where, Y = Sales, X1 = Price and X2 = Promotion Costs