



Regression Analysis

Prof. D. Nagesh Kumar
Dept. of Civil Engg.

IISc, Bangalore – 560 012, India
URL: <http://www.civil.iisc.ernet.in/~nagesh>

Simple Regression Determining coefficients

Assuming n data points denoted by Y_i , the regression equation, $Y_i = a + bX_i$, can be estimated so as to minimize the sum of the squared deviations. Defining

$$e_i = Y_i - \hat{Y}_i$$

then

$$e_i^2 = (Y_i - \hat{Y}_i)^2$$

and

$$\sum e_i^2 = \sum (Y_i - \hat{Y}_i)^2$$

By substitution,

$$\sum e_i^2 = \sum (Y_i - a - bX_i)^2$$

Applying calculus,

$$\frac{\partial \sum e_i^2}{\partial a} = -2\sum Y_i - a - bX_i = 0 \tag{5-39}$$

$$\frac{\partial \sum e_i^2}{\partial b} = -2\sum X_i Y_i - a - bX_i = 0 \tag{5-40}$$

From (5-39)

$$-\sum Y_i + na + b\sum X_i = 0 \tag{5-41}$$

From (5-40)

$$-\sum X_i Y_i + a\sum X_i + b\sum X_i^2 = 0 \tag{5-42}$$

Equations (5-41) and (5-42) can be solved simultaneously to obtain the values of a and b . Solving (5-41) for a gives

$$a = \frac{\sum Y_i}{n} - b \frac{\sum X_i}{n} \tag{5-43}$$

Determining coefficients – contd..

$$b = \frac{n\sum X_i Y_i - \sum X_i \sum Y_i}{n\sum X_i^2 - (\sum X_i)^2} \tag{5-44}$$

Thus, the values of a and b in (5-43) and (5-44) correspond to the means where the first derivatives of (5-41) and (5-42) are zero, that is, where the sum of the squared errors is at a minimum.

Since the value of b is known through (5-44), it can be substituted into (5-43) to get a . The solution point for a and b is defined where $\sum e_i^2$ is at a minimum, as can be verified by computing the second derivatives, and showing that

$$\frac{\partial^2 \sum e_i^2}{\partial a^2} > 0 \text{ and } \frac{\partial^2 \sum e_i^2}{\partial b^2} > 0$$

A convenient way of expressing the regression equation is in terms of deviations from the mean values of X and Y . The data are transformed by substituting

$$x_i = X_i - \bar{X} \text{ and } y_i = Y_i - \bar{Y}$$

and

$$y_i = \hat{Y}_i - \bar{Y} \text{ or } \hat{Y}_i = y_i + \bar{Y}$$

The regression equation, $Y_i = a + bX_i$, then becomes

$$\hat{Y}_i = \bar{Y} + a + bx_i + \bar{Y}$$

which simplifies to

$$\hat{y}_i = a + bx_i + \bar{Y} - \bar{Y}$$

But since

$$a = \bar{Y} - b\bar{X} \tag{5-45}$$

$$\hat{y}_i = \bar{Y} - b\bar{X} + bx_i + \bar{Y} - \bar{Y}$$

and

$$\hat{y}_i = bx_i$$

Similarly, by substitution

$$\sum e_i^2 = \sum (y_i - \hat{y}_i)^2$$

$$= \sum (y_i - bx_i)^2$$

$$= \sum (y_i^2 - 2bx_i y_i + b^2 x_i^2)$$

$$= \sum y_i^2 - 2b \sum x_i y_i + b^2 \sum x_i^2 = 0 \tag{5-46}$$

Equation (5-46) is similar to (5-41) except that it is expressed in terms of deviations. Equation (5-43), on the other hand, is not needed because the