# 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, $\hat{Y}_i = a + bX_i$, 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

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

By substitution,

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

Applying calculus,

$$\frac{\partial \Sigma e_i^2}{\partial a} = -2\Sigma (Y_i - a - bX_i) = 0, \tag{5-39}$$

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

From (5-39)

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

From (5-40)

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

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

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

## *Determining coefficients – contd..*

which substituted into (5-42) yields

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

Thus, the values of $a$ and $b$ in (5-43) and (5-44) correspond to the point 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 indeed where $\Sigma e_i^2$ is at a minimum, as can be verified by computing the second derivatives, and showing that

$$\frac{\partial^2 \Sigma e_i^2}{\partial a^2} > 0 \quad \text{and} \quad \frac{\partial^2 \Sigma e_i^2}{\partial b^2} > 0.$$

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

$$x_i = X_i - \bar{X} \quad \text{or} \quad X_i = x_i + \bar{X}$$

and

$$y_i = Y_i - \bar{Y} \quad \text{or} \quad Y_i = y_i + \bar{Y}.$$

The regression equation, $Y = a + bX$, then becomes

$$y_i + \bar{Y} = a + bx_i + b\bar{X},$$

which simplifies to

$$y_i = a + bx_i + b\bar{X} - \bar{Y}.$$

But, since

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

$$y_i = \bar{Y} - b\bar{X} - bx_i + b\bar{X} - \bar{Y}$$

and

$$y_i = bx_i.$$

Similarly, by substitution

$$\Sigma e_i^2 = \Sigma (y_i - bx_i)^2$$

$$= \Sigma (y_i - bx_i)^2,$$

$$\frac{d\Sigma e_i^2}{db} = 2\Sigma x_i (y_i - bx_i)$$

$$= -2\Sigma x_i y_i + 2b\Sigma x_i^2 = 0,$$

$$b = \frac{\Sigma x_i y_i}{\Sigma x_i^2}. \tag{5-46}$$

Equation (5-46) is similar to (5-44) except that it is expressed in terms of deviations. Equation (5-43), on the other hand, is not needed because the