REGRESSION ANALYSIS

Regression analysis is used to understand which among the independent variables are related to the dependent variable, and to explore the forms of these relationships.

The Method of Least Squares

the ideas of regression analysis can be introduced in the simple setting where the distribution of a random variable $\boldsymbol{y}$ depends on the value $\boldsymbol{x}$ of one other variable. Calling the terminologies of the two variables $\boldsymbol{x}$ and $\boldsymbol{y}$ as:

$\boldsymbol{x}$ = **independent variable**, also called **input variable**.

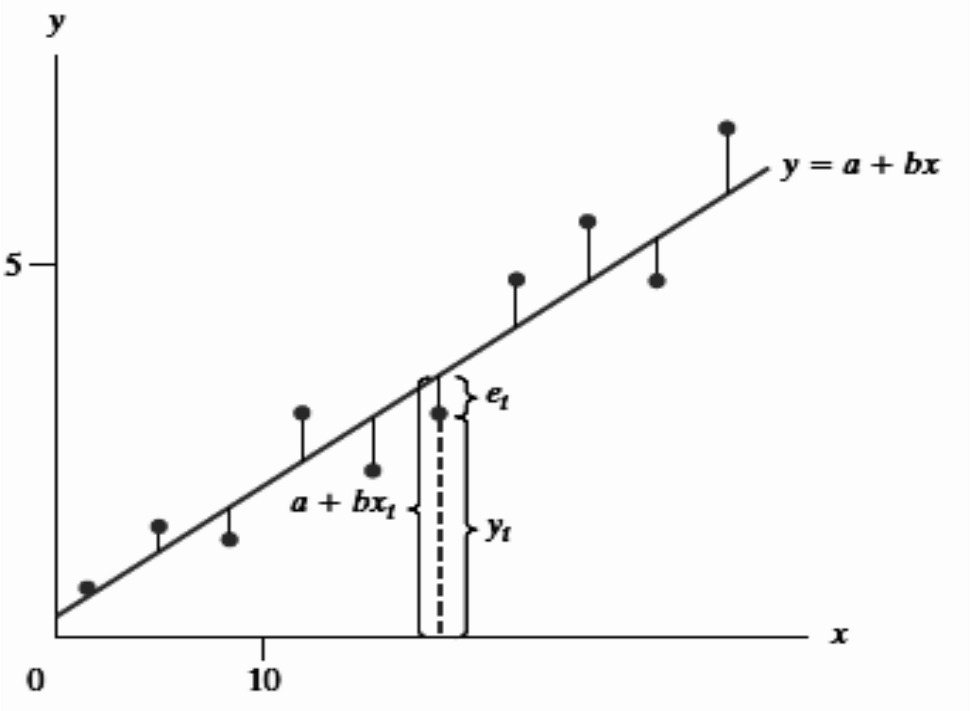
$\boldsymbol{y}$ = **dependent variable**, or **response variable**.

In most situations of this sort , it is mainly interested in the relationship between $\boldsymbol{x}$ and the mean $\boldsymbol{E}[\boldsymbol{Y} \mid \boldsymbol{x}]$ of the corresponding distribution of $\boldsymbol{Y}$. This relationship is mentioned as the **regression of $\boldsymbol{Y}$ on $\boldsymbol{x}$**.

To state the problem formally, assume n paired observations (x_i, y_i) for which the regression of Y on x is linear. It is required to determine the line (that is, the equation of the line) which in some sense provides the best fit. There are several ways in to interpret the word "best," and the meaning that shall be explained as follows. If it is predicted y by means of the linear equation: $\hat{y} = a + b\,x$; where a and b are constants of the equation, then e_i, the error in predicting the value of y correspondingto the given x_i, is $e_i = y_i - \hat{y}$ and it is required to determine a and b so that these errors are in some sense as small as possible. Since this sum can be made equal to zero as result of positive and negative errors cancel, the sum of the squares of the e_i will be minimize (for the same reason we worked with the squares of the deviations from the mean in the definition of the standard deviation). In other words, the choose a and b so that:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} [y_i - (a + b\,x_i)]^2$$

is a minimum. This is equivalent to minimizing the sum of the squares of the vertical distances from the points to the line in any scatter plot as shown in the figure. The procedure of finding the equation of the line which best fits a given set of paired data, called the **method of least squares**, yields values for a and b.



It is convenient to introduce some notation for the sums of squares and sums of cross-products.

$$S_{xx} = \sum_{i=1}^{n} (x_i - \overline{x})^2$$

$$S_{yy} = \sum_{i=1}^{n} (y_i - \overline{y})^2$$

$$S_{xy} = \sum_{i=1}^{n} (x_i - \overline{x})\,(y_i - \overline{y})$$

where $\overline{x}$ and $\overline{y}$ are, respectively, the means of the values of x and y.

Least squares estimates for constant of linear equation $\hat{y} = a + b\,x$ are:

$$b = \frac{S_{xy}}{S_{xx}}$$

$$a = \overline{y} - b\,\overline{x}$$

SSE = residual sum of squares= $\sum_{i=1}^{n} (y_i - a - b\,x_i)^2$

The minimum value of the sum of squares is called the **residual sum of squares** or **error sum of squares**.

$$\mathbf{SSE} = S_{yy} - \frac{S_{xy}^2}{S_{xx}}$$

Examole 1 Least squares calculations for the given paired observations (x and y)

Calculate the least squares estimates and sum of squares error for the given paired (x_i, y_i) observation.

x	0	1	2	2	4	4	5	6
y	25	20	30	40	45	50	60	50

Solution The structure of the table guides the calculations

x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})^2$	$(x-\bar{x})(y-\bar{y})$	$(y-\bar{y})^2$	residual
0	25	−3	−15	9	45	225	3
1	20	−2	−20	4	40	400	−8
2	30	−1	−10	1	10	100	−4
2	40	−1	0	1	0	0	6
4	45	1	5	1	5	25	−1
4	50	1	10	1	10	100	4
5	60	2	20	4	40	400	8
6	50	3	10	9	30	100	−8
$\bar{x}=3$	$\bar{y}=40$	0	0	$S_{xx}=30$	$S_{xy}=180$	$S_{yy}=1350$	

$$b = \frac{S_{xy}}{S_{xx}} = \frac{180}{30} = 6 \qquad \text{and} \qquad a = \bar{y} - b\,\bar{x} = 40 - 6 \text{ x } 3 = 22$$

Accordingly, the best fit linear equation between x and y is : $\boxed{y = 22 + 6x}$

$$\text{SSE} = \sum_{i=1}^{n} (y_i - a - b\,x_i)^2 = 3^2 + (-8)^2 + (-4)^2 + 6^2 + (-1)^2 + 4^2 + 8^2 + (-8)^2 = 270$$

$$\text{SSE} = S_{yy} - \frac{S_{xy}^2}{S_{xx}} = 1350 - \frac{180^2}{30} = 270$$