EXPERIMENT

Regression Analysis

Objectives

- To study the linear regression method.

Introduction to regression method

Regression problem deals with the relationship between the frequency distribution of one (dependent) variable and another (independent) variable(s) which is (are) held fixed at each of several values.

The technical term “regression” has become part of the language of statistics due to the work of Sir Francis Galton. In the 1880’s, Galton laid the foundations of modern correlation techniques in a study of the relationship between the average heights of children and the heights of their parents.

In statistics, “regression” means simply average relationship, and thus between two variables, y and x,

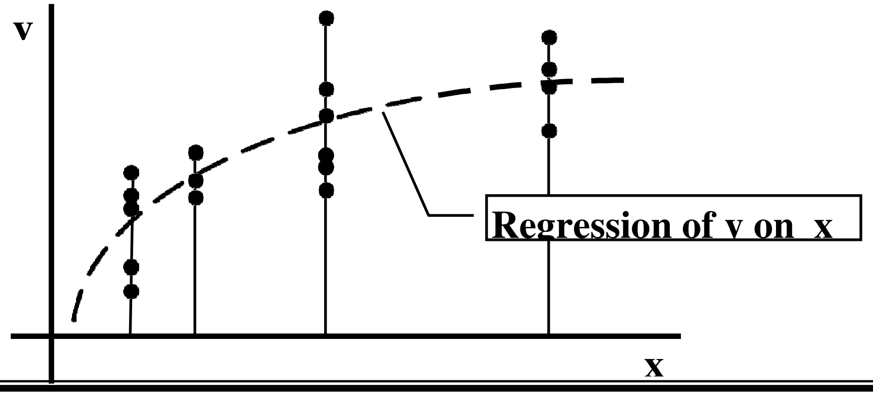
Regression of variable y on variable x

implies a relationship between

- the average of the values of the variable y for a given value of the variable x,

and

- the value of the variable x.



Note that the presumption of being dependent and independent for respective variables y and x, does not necessarily mean that a causality (a cause-effect or an input-output relationship) must exist between them, even though they may highly correlated. Because,

- both variables may be affected by the same cause, or
- two variables may be interdependent, or
- one variable is the cause, although not necessarily the sole cause, of the other.

Note also that

a regression of y on x

can only be used to estimate the mean value of y for a given x, and should not be used to estimate the mean value for x for a given y. In other words,

a regression of y on x

does not immediately imply

a regression of x on y

However, there may exists a regression of x on y but in a totally different nature.

In the practice of engineering experimentation, the regression analysis is used to estimate the "best" empirical constants, with their respective confidence limits, to fit a mathematical model to a set of measurement data.

Once a mathematical model is so established between the dependent variable y and the independent variable(s) x, it can then be used to predict y for new values of x, by treating x as a continuous variable in the implied interpolation process involved.

Linear Regression:
It is the regression in which the model used is linear; i.e.,

$$y = ax + b$$

Curvilinear Regression:
It is the regression in which the model used is a polynomial in x; i.e.,

$$y = f(x)$$

Nonlinear Regression:
It is the regression in which the model used is nonlinear (polynomial or not); e.g.,

$$y = ax + be^{-cx}$$

Multivariate (Multiple) Regression:
It is the regression in which there exist multiple independent variables used in a linear model; e.g.,

$$y = a_1x_1 + a_2x_2 + b$$

or in a nonlinear model; e.g., $y = a(1-x_1^2)x_2 + bx_3$

Method of Least Squares:

It is a general method used in a very broad class of engineering problems like

- curve fitting,
- parameter estimation,
- system identification, and
- static and dynamic optimization

Its major advantage over other techniques is that it results in a set of linear algebraic equations in terms of unknown model parameters if these parameters appear linearly in the mathematical model.

Example: Let x1, x2, ….., xn constitute n measurements. A best estimate xbest of these measurements is asked in the sense that the quantity

$$\boxed{\mathbf{E} = \sum_{i=1}^{n}(\mathbf{x}_i - \mathbf{x}_{best})^2}$$

is minimized.

The minimization of E with respect to $\mathbf{x}_{best}$ requires that

$$\frac{dE}{dx_{best}} = \frac{d\sum_{i=1}^{n}(x_i - x_{best})^2}{dx_{best}} = -2\sum_{i=1}^{n}(x_i - x_{best}) = 0$$

$$\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_{best} = \sum_{i=1}^{n} x_i - nx_{best} = 0$$

$$\boxed{x_{best} = \frac{1}{n}\sum_{i=1}^{n} x_i}$$

which is nothing but the arithmetic mean of the data set given.

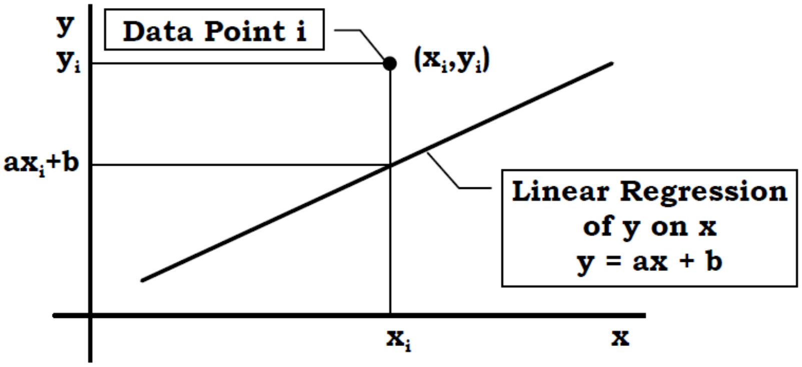
Example (Linear Regression): Let (x1,y1), (x2,y2), ….., (xn,yn) constitute n paired data points. A best linear regression of y on x as

$$\mathbf{y = ax + b}$$

is asked (i.e., the best estimates of a and b are asked) in the sense that the quantity

$$\boxed{E = \sum_{i=1}^{n}\left[y_i - (ax_i + b)\right]^2}$$

is minimized.



Note that the quantity

$$y_i - (ax_i + b)$$

is the vertical distance between the data point i and the regressed value of ***y* (=a*xi*+*b*)** at xi in the y versus x plane.
The minimization of E with respect to a and b requires that

$$\boxed{\frac{dE}{da} = 0} \quad \text{and} \quad \boxed{\frac{dE}{db} = 0}$$

The solution of last two equations in terms of a and b gives:

$$\boxed{a = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}}$$

$$\boxed{\mathbf{b} = \frac{\left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} x_i^2\right) - \left(\sum_{i=1}^{n} x_i y_i\right)\left(\sum_{i=1}^{n} x_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}}$$

or by defining

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

$$\overline{x^2} = \frac{1}{n}\sum_{i=1}^{n} x_i^2 \qquad \overline{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i$$

$$\boxed{a = \frac{\overline{xy} - \bar{x}\,.\,\bar{y}}{\overline{x^2} - \bar{x}^2}} \quad \textbf{and} \quad \boxed{b = \frac{\overline{x^2}\,.\,\bar{y} - \bar{x}\,.\,\overline{xy}}{\overline{x^2} - \bar{x}^2}}$$

or

$$\boxed{\begin{bmatrix} a \\ b \end{bmatrix} = \frac{1}{\overline{x^2} - \bar{x}^2}\begin{bmatrix} -\bar{x} & 1 \\ \overline{x^2} & -\bar{x} \end{bmatrix}\begin{bmatrix} \bar{y} \\ \overline{xy} \end{bmatrix}}$$

Note also that

$$\boxed{b = \bar{y} - a\bar{x}}$$

Example: **Let the following data show the result of an experiment as 6 pairs of values.**

i	1	2	3	4	5	6
x_i	0.5	1.0	1.5	2.0	2.5	3.0
y_i	0.71	1.33	1.68	1.88	2.31	2.48

Example:

The data *y* has been observed for various values of *x*, as follows:

y	240	181	193	155	172	110	113	75	94
x	1.6	9.4	15.5	20.0	22.0	35.5	43.0	40.5	33.0

Fit the simple linear regression model using least squares.