Regression Line Analysis

Nirdosh Bhatnagar

1. **Simple linear regression model**

In performance modeling, and other disciplines a relationship often exists between two or more set of variables.

A relationship can be developed among these variables using statistical techniques.

We are given a set of points $\{(x_i, y_i) : 1 \leq i \leq n\}$.

For example x_i's can be the number tasks which attempt service at a CPU, and y_i''s can be the CPU utilization.

The first step is to plot these points on a graph.

The resulting plot is generally called a *scatter diagram.*

We will assume that the points fall approximately on a straight line.

Our goal is to fit these points approximately to a straight line.

Before a linear regression model is developed, an analyst should do a visual test of the scatter diagram.

It should be approximately linear.

2. **Analysis**

Let the equation of the desired line be

$$y = a + bx$$

This equation is called a *regression equation of* y *on* x.

Method of least-square technique is used to find the values of a and b.

Here, the aim is to have

$$y_i = a + bx_i + e_i \qquad\qquad 1 \le i \le n$$

The e_i's are said to be error terms. Define

$$\widehat{y}_i = a + bx_i \qquad\qquad 1 \le i \le n$$

where $\widehat{y}_i$ is the estimated value of y_i.

The goal of least-square technique is to minimize

$$E = \sum_{i=1}^{n} (\widehat{y}_i - y_i)^2$$

Then

$$E = \sum_{i=1}^{n} e_i^2$$

Define

$$\begin{aligned}\overline{x} &= \frac{1}{n}\sum_{i=1}^{n} x_i \\ \overline{y} &= \frac{1}{n}\sum_{i=1}^{n} y_i\end{aligned}$$

The values a and b can be obtained as follows.

$$\begin{aligned}b &= \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \\ a &= \overline{y} - b\overline{x}\end{aligned}$$

3. **Least-Squares Line in Terms of Sample Variances and Covariance**

The sample variances and covariances of the x-sequence and y-sequence are

$$\begin{aligned} S_x^2 &= \frac{\sum_{i=1}^n (x_i - \overline{x})^2}{n} \\ S_y^2 &= \frac{\sum_{i=1}^n (y_i - \overline{y})^2}{n} \\ S_{xy} &= \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{n} \end{aligned}$$

Then

$$b = \frac{S_{xy}}{S_x^2}$$

Define the sample correlation coefficient r by

$$r = \frac{S_{xy}}{S_x S_y}$$

The regression line equation can be written as

$$\frac{(y - \overline{y})}{S_y} = r\frac{(x - \overline{x})}{S_x}$$

The value r^2 is sometimes referred to as *coefficient of determination.*

The net error in the regression line is

$$E = nS_y^2\left(1 - r^2\right)$$

4. **Observations**

We make the following observation.

- The regression line passes through the point $(\overline{x}, \overline{y})$.

- Since $e_i = -(\widehat{y}_i - y_i)$, for $1 \leq i \leq n$, we have $\sum_{i=1}^{n} e_i = 0$.

- The sample correlation coefficient r has the following properties:

    1. The sample coefficient r is dimensionless.

    2. $0 \leq r^2 \leq 1$, that is $-1 \leq r \leq 1$.

    3. If all points in the scatter diagram lie on the straight line then, $r = 1$ (positive slope) or $r = -1$ (negative slope).

4. If all points in the scatter diagram do not lie on the regression line, then $-1 < r < 1$.

5. if $|r|$ is close to 0, then the points in the scatter diagram show no straight-line trend, that is no linear correlation.

6. If $0 < r$, then the regression line has a positive slope. However, if $r < 0$, then the regression line has a negative slope.

7. The magnitude of r is not an indicator of the steepness or slope of the regression line, rather r is a measure of how closely the data points cluster about the line.

8. The following expression gives a quantitative interpretation of r^2.

$$\begin{aligned}\left(1-r^2\right) &= \frac{\sum_{i=1}^{n}\left(\widehat{y}_i - y_i\right)^2}{nS_y^2} \\ r^2 &= \frac{\sum_{i=1}^{n}\left(\widehat{y}_i - \overline{y}\right)^2}{\sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2} = \frac{\text{explained variation}}{\text{total variation}}\end{aligned}$$

Therefore, r^2 can be interpreted as the fraction of the total variation that is explained by the least-squares regression line.

Alternately, r measures how well the least-squares regression line fits the sample data.

5. **Example**

The use of the above formula is illustrated in this example.

Let number of data points be $n = 6$.

The data points are

$$(2,2)\,,(4,6)\,,(5,4)\,,(7,8)\,,(8,10)\,,\text{ and }\;(10,12)$$

Find the equation of the regression line. The relevant quantities are:

$$
\begin{aligned}
\overline{x} &= 6,\\
\overline{y} &= 7,\\
S_x^2 &= 7,\\
S_y^2 &= 11.6667,\\
S_{xy} &= 8.6667,\\
r^2 &= 0.9197,\\
r &= 0.9590,\\
E &= 5.6190,\\
a &= -0.42857,\\
b &= 1.2381.
\end{aligned}
$$

The equation of the regression line is

$$y = -0.42857 + 1.2381x$$

We have

$$
\begin{aligned}
\widehat{y}_1 &= 2.047619, \\
\widehat{y}_2 &= 4.5238095, \\
\widehat{y}_3 &= 5.7619048, \\
\widehat{y}_4 &= 8.2380952, \\
\widehat{y}_5 &= 9.4761905, \\
\widehat{y}_6 &= 11.952381.
\end{aligned}
$$

Also since $e_i = -\left(\widehat{y}_i - y_i\right)$, for $1 \leq i \leq 6$, we have

$$
\begin{aligned}
e_1 &= -0.047619, \\
e_2 &= 1.476190, \\
e_3 &= -1.761905, \\
e_4 &= -0.238095, \\
e_5 &= 0.523810, \\
e_6 &= 0.047619
\end{aligned}
$$

It can be checked that $\sum_{i=1}^{6} e_i = 0$.