Linear regression

Sunday, 10 March 2024 1:05 PM

Independent Variables

The feature of the dataset are known as independent variables

dependent Variables

The target variables are known as dependent variables

What is linear Regression?

We have to determine two things

1) Do the independent variable predict the dependable variable with right accuracy

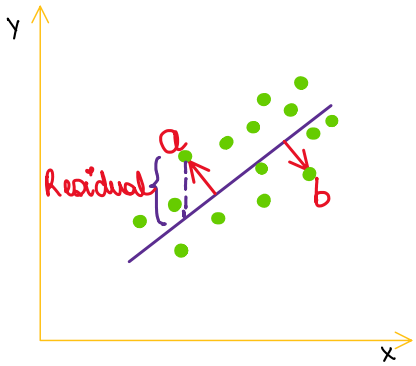
2) which independent variable are best fitted to predict the dependable variable

Linear Regression is a statistical model

$$Y = mX + C$$

Multiple Independent variable

$$Y = (m_1X1 + m_2X2 + m_3X3 + \cdots m_nX_n) + C$$



The distance from a line to the data point is called Residual

Residual Sum of Square (RSS)

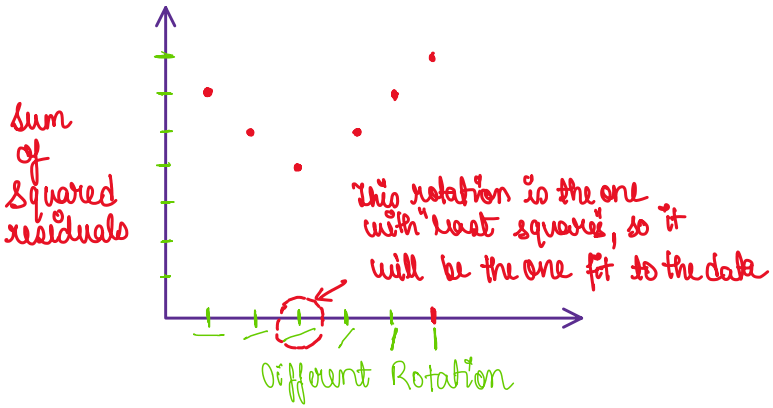
$$RSS = \sum_{i=1}^{n} (y_i - f(x_i))^2$$

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

y_i = i^{th} value of variable to be predicted

$f(x_i)$ is the predicted value

n is the number of terms or variable



Total Sum of Squares

$$TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

y_i = value of sample
$\bar{y}$ = The mean value of sample

Regression sum of squares

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

$\hat{y}_i$ = The value estimated by regression line
$\bar{y}$ = mean value of sample

$$TSS = SSR + RSS$$

R^2 score

$$R^2 = 1 - \frac{RSS}{TSS}$$

$$\text{Variance (around mean)} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y})^2$$

Average sum of squares

$$\text{Variance (fit)} = \frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y)^2$$

Average sum of squares

$$R^2 = \frac{\text{Var(mean)} - \text{Var(fit)}}{\text{var(mean)}}$$

An R-Squared value shows how well the model predicts the outcome of the dependent variable. R^2 value range from 0 to 1

An R-Squared value of 0 means that the model explains or predicts 0% of the relationship between the dependent and independent variables

A value of 1 indicates that the model predicts 100% of the relationship between the dependent and independent variables

A value of 0.5 indicates that the model predicts 50% of the relationship between the dependent and independent variables

Root mean square Error

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}$$

y_i = actual value
$\hat{y}_i$ = predicted value
n = number of observations

X	Y
1	42
3	50
10	75
16	100
26	150
36	200

$n = 6$
$\Sigma x = 92$
$\Sigma y = 617$
$\Sigma xy = 13642$
$\Sigma x^2 = 2338$
$\Sigma y^2 = 82389$

$\bar{x} = \frac{1}{n}(\Sigma x)$

$\bar{y} = \frac{1}{n}(\Sigma y)$

$y = a + bx$

$$b = \frac{n\Sigma xy - \Sigma x \Sigma y}{n\Sigma x^2 - (\Sigma x)^2}$$

$\bar{y} = a + b\bar{x}$

$a = 33.83$

$b = 4.51$

$y = 33.83 + 4.51x$

R^2 Calculation

$\bar{y} = 3.511$

x	y	$\hat{y} = 0.1 + 0.78x$	$y - \hat{y}$	$(y - \hat{y})^2$	$y - \bar{y}$	$(y - \bar{y})^2$
0.4	1.4	0.802	0.598	0.3576	2.111	4.4521
1.8	2.6	1.504	1.096	1.2012	0.911	0.8299
2.4	1.0	1.972	0.972	0.9506	2.511	6.3001
3.5	3.7	2.83	0.87	0.7569	0.189	0.03572
3.9	5.5	3.142	2.358	5.5601	1.989	3.9561
4.4	3.2	3.532	0.332	0.1102	0.311	0.0967
5.1	3.0	4.078	1.078	1.1620	0.511	0.2611
5.6	4.9	4.468	0.432	0.1866	1.389	1.9293
6.3	6.3	5.014	1.286	1.6537	2.789	7.7785
			RSS	11.918	TSS	25.56

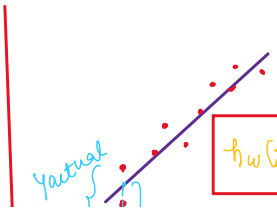
$$R^2 = 1 - \frac{RSS}{TSS}$$

$= 1 - 0.466$

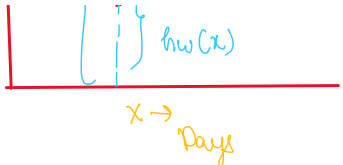
$= 0.53$

$y = h_w(x)$

no. of cases ↑



$h_w(x) = w_0 + w_1 x$



Define an objective function (also called Error/Cost function)

$$J(w_0, w_1)$$

objective is to find values of w_0 & w_1, so that $J(w_0, w_1)$ becomes optimal

$$J = h_w(x) - Y_{actual}$$

∴ Sometimes J will be positive & some times J will be negative

Therefore

$$J = \sum (h_w(x) - Y_{actual})^2$$

$$J(w_0, w_1) = (w_0 + w_1 x_1 - y_1)^2 + (w_0 + w_1 x_2 - y_2)^2 + (w_0 + x_1 x_3 - y_3)^2 \cdots + (w_0 + w_1 x_m - y_m)^2 \quad -- (1)$$

To minimize $J(w_0, w_1)$ find

$$\frac{\partial J}{\partial w_0} = 2\{(w_0 + w_1 x_1 - y_1) + (w_0 + w_1 x_2 - y_2) + \cdots + (w_0 + w_1 x_m - y_m)\}$$

$$m w_0 + w \sum_{i=1}^{m} x_i - \sum_{i=1}^{m} y_i = 0$$

$$\frac{\partial J}{\partial w_1} = 2\{(w_0 + w_1 x - y_1) x_1 + (w_0 + w_1 x_2 - y_2) x_2 + \cdots + (w_0 + w_1 x_m - y_m) x_m\} = 0$$

$$w_0 \sum_{i=1}^{m} x_i^0 + w_i \sum_{i=1}^{m} x_i^2 - \sum_{i=1}^{m} x_i y_i = 0$$

Rewriting the Equation ② & ③ after substituting

$$A = \sum_{i=1}^{m} x_i^0 \qquad m \leftarrow \text{number of terms.}$$

$$B = \sum_{i=1}^{m} y_i$$

$$C = \sum_{i=1}^{m} x_i^2$$

$$D = \sum_{i=1}^{m} x_i y_i$$

$$m w_0 + A w_1 = B \quad \text{—— (4)}$$

$$w_0 A + C w_1 = D \quad \text{—— (5)}$$

solving 4 & 5

$$w_1 = \frac{AB - DM}{A^2 - Cm}$$

$$w_0 = \frac{BC - AC}{Cm - A^2}$$

$$J = \sum (h_w(x) - y_{actual})^2$$

$$J(w_0, w_1) = \frac{1}{2m} \sum_{i=1}^{m} (h_i(x) - y_i)^2$$

Because of differenciation

no. of samples