Simple Linear Correlation Analysis:

In statistics, the Pearson correlation coefficient, also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC) or the bivariate correlation, is a measure of the linear correlation between two variables X and Y. Owing to the Cauchy–Schwarz inequality it has a value between +1 and −1, where 1 is total positive linear correlation, 0 is no linear correlation, and −1 is total negative linear correlation. It is widely used in the sciences. It was developed by Karl Pearson from a related idea introduced by Francis Galton in the 1880s.

$$r = \frac{\sum(x_i - \bar{x}) * (y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 - (y_i - \bar{y})^2}} = \frac{\sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}}{\sqrt{\left[\sum x_i^2 - \frac{(\sum x_i)^2}{n}\right]\left[\sum y_i^2 - \frac{(\sum y_i)^2}{n}\right]}}$$

Simple Linear Regression Analysis

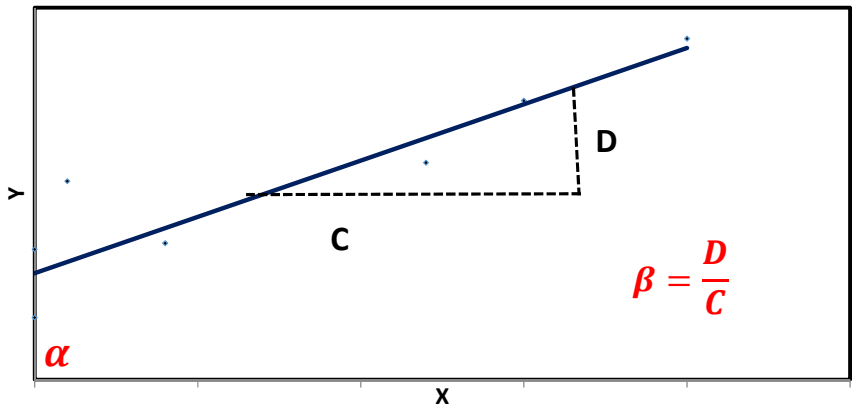
In statistics, linear regression is a linear approach to modelling the relationship between a scalar response (or dependent variable) and one or more explanatory variables (or independent variables). The case of one explanatory variable is called simple linear regression. For more than one explanatory variable, the process is called multiple linear regression. This term is distinct from multivariate linear regression, where multiple correlated dependent variables are predicted, rather than a single scalar variable.

In linear regression, the relationships are modeled using linear predictor functions whose unknown model parameters are estimated from the data. Such models are called linear models. Most commonly, the conditional mean of the response given the values of the explanatory variables (or predictors) is assumed to be an affine function of those values; less commonly, the conditional median or some other quantile is used. Like all forms of regression analysis, linear regression focuses on the conditional probability distribution of the response given the values of the predictors, rather than on the joint probability distribution of all of these variables, which is the domain of multivariate analysis.

Linear regression was the first type of regression analysis to be studied rigorously, and to be used extensively in practical applications. This is because models

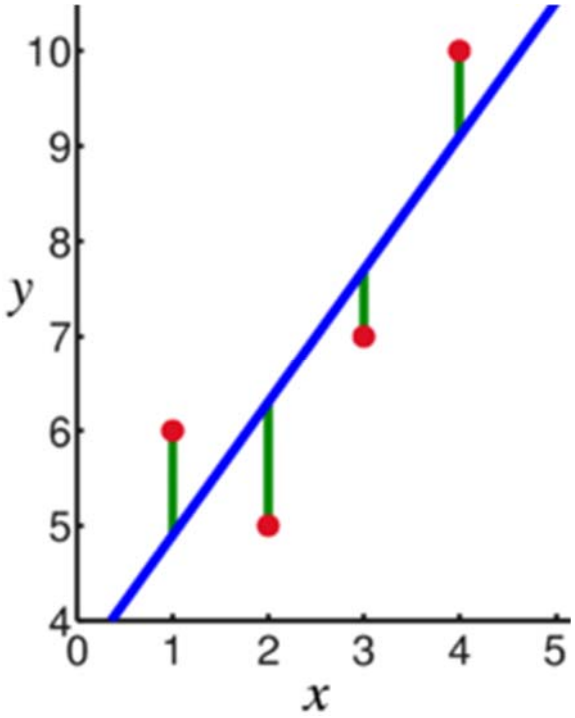
which depend linearly on their unknown parameters are easier to fit than models which are non-linearly related to their parameters and because the statistical properties of the resulting estimators are easier to determine.

$$\hat{y} = \alpha + \beta x$$



$$\alpha = y_{intercept} = \bar{y} - \beta * \bar{x}$$

$$\text{Regression Coefficient} = slope = \beta = \frac{\sum(x_i - \bar{x}) * (y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{\sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}}{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}$$



Example: Use simple linear regression with Correlation to analysis data in the underneath, where r tabulation = 0.497, then test its significantly at 0.05, where t tabulation at this level = 2.228 and F tabulation at 0.05 level = 4.96, 0.01 level = 10.04 from this data table:

X_i	65	50	55	65	55	70	65	70	55	70	50	55
Y_i	85	74	76	90	85	87	94	98	81	91	76	74

Solve

Rank	x_i	y_i	x_i^2	y_i^2	$x_i y_i$	$\hat{y}$
1	65	85	4225	7225	5525	88.36
2	50	74	2500	5476	3700	74.91
3	55	76	3025	5776	4180	79.39
4	65	90	4225	8100	5850	88.36
5	55	85	3025	7225	4675	79.39
6	70	87	4900	7569	6090	92.85
7	65	94	4225	8836	6110	88.36
8	70	98	4900	9604	6860	92.85
9	55	81	3025	6561	4455	79.39
10	70	91	4900	8281	6370	92.85
11	50	76	2500	5776	3800	74.91
12	55	74	3025	5476	4070	79.39
$\sum$	725	1011	44475	85905	61685	1011

$$\beta = \frac{\sum(x_i - \bar{x}) * (y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{\sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}}{\sum x_i^2 - \frac{(\sum x_i)^2}{n}} = \frac{61685 - \frac{(725) * (1011)}{12}}{44475 - \frac{(725)^2}{12}}$$

$$= \frac{\frac{740220 - 732975}{12}}{\frac{533700 - 525625}{12}} = \frac{603.75}{672.92} = 0.897$$

$$\alpha = \bar{y} - \beta * \bar{x} = 84.25 - (0.897) * (60.417) = 84.25 - 54.194049 = 30.056$$

$$\therefore Linear\ equation\ is\ \hat{y} = \alpha + \beta x = 30.056 + 0.897(x)$$

$$Sum\ of\ Squeres\ due\ to\ regression = SSR = \beta * \left[\sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}\right]$$

$$= \beta^2 * S_x^2 = (0.897)^2 * (672.92) = 541.44$$

$$SST = S_y^2 = \sum y_i^2 - \frac{(\sum y_i)^2}{n} = 728.25$$

$$Sum\ of\ Sequares\ due\ to\ deviations\ from\ regression = SS_e = SST - SSR$$
$$\rightarrow SS_e = 728.25 - 541.44 = 186.81$$

S.O.V.	d.f.	SS	MS	Cal.F	F_tab	
					0.05	0.01
Regression	1	541.44	541.44	28.98^{**}	4.96	10.04
Residual(Error)	10	186.81	18.681			
Total	11	728.25				

$$r = \frac{\sum(x_i - \bar{x}) * (y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 - (y_i - \bar{y})^2}} = \frac{\sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}}{\sqrt{\left[\sum x_i^2 - \frac{(\sum x_i)^2}{n}\right]\left[\sum y_i^2 - \frac{(\sum y_i)^2}{n}\right]}} = 0.862^{*}$$