Linear Regression

The simplest way to predict values of a random variable in Statistics can be considered as Linear Regression technique.

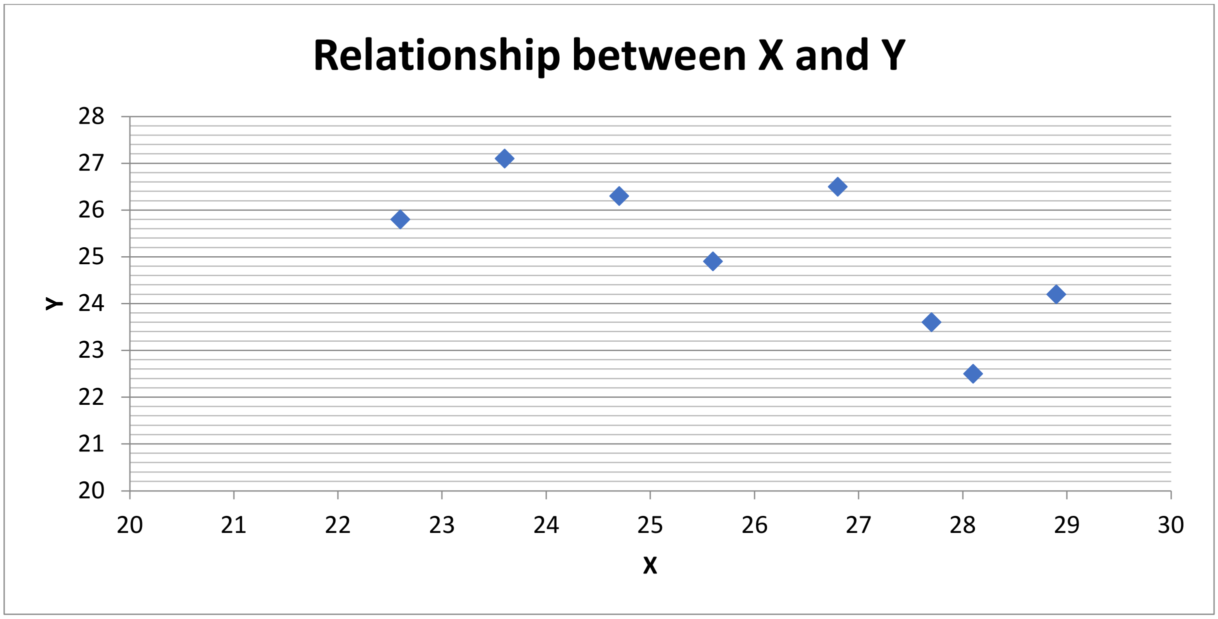
Simple Linear Regression

Mainly the relationship between two continuous variables, which can be measured simultaneously on an experimental unit in an experiment is considered, where one variable will be taken as the dependent variable (y) and the other is said to be the independent or the explanatory variable (x).

Eg: Temperature and Pressure

Indications

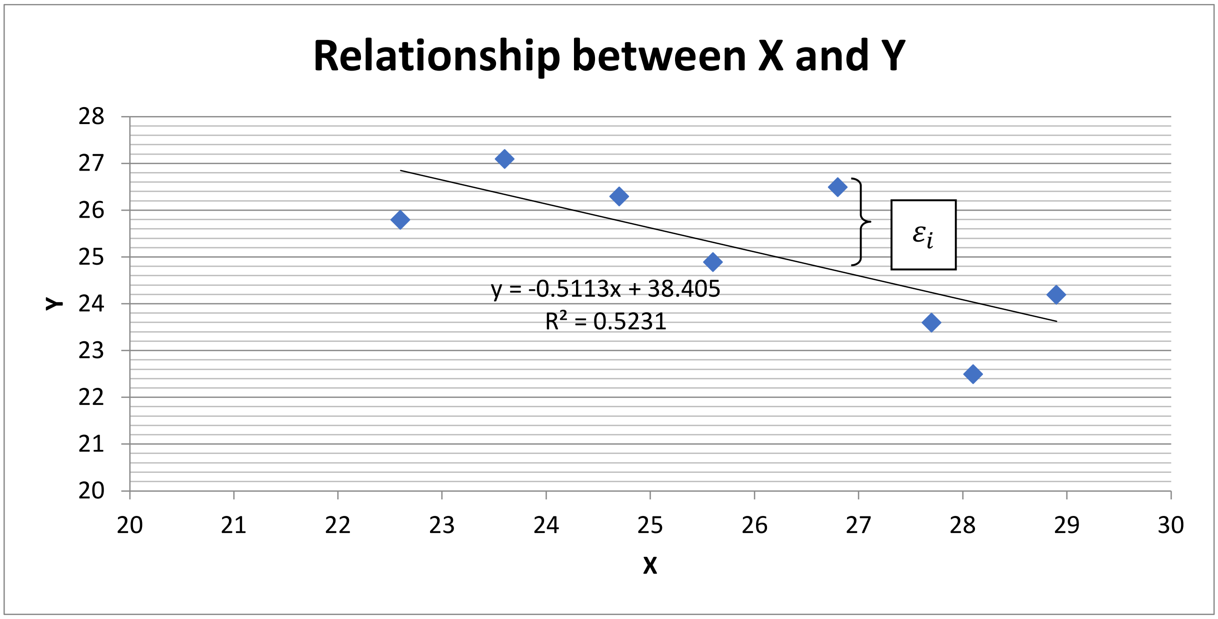
1. Scatter Diagram
2. Correlation Coefficient



Simple Linear Regression

In simple linear regression, we allow only one independent variable to predict the dependent variable. Under multiple linear regression there can be many independent variables predicting a sing dependant variable.

In here, a set of measurements(x_i, y_i); i=1,2,3…$n$on n individuals are taken and if evidence is available on a scatter diagram for a linear relationship between x and y, a regression function will be established to model the relationship.



The **coefficient of determination** (denoted by **R^2**) is a key output of regression analysis. It is interpreted as the proportion of the variance in the dependent variable that is predictable from the independent variable.

General Model

$$\boldsymbol{Y = \alpha + \beta X}$$

For an existing individual

$$y_i = \alpha + \beta x_i + \varepsilon_i;$$

$$\varepsilon_i = the\ observation's\ deviation\ from\ the\ model\ Y = \alpha + \beta X$$

OR

$\varepsilon_i = error\ in\ prediction$

Fitting a Linear Regression Function

First Step : Plot the scatters and look for any evidence for a linear relationship.

Assumption:

Error terms are independently and identically distributed as Normal with mean zero and varianceσ^2.

i.e.$\varepsilon \sim N(0, \sigma^2)$

Parameter Estimation (estimating $\boldsymbol{\alpha}$ &$\boldsymbol{\beta}$):

Estimation will be carried out based on the principle of **"least squares".**

In least square estimation, the "sum of squares of errors" will be minimized.

i.e we will find α, β such that $\sum \varepsilon_i^2$ is at minimum.

Let *ESS* = Error Sum of Squares

$$\boldsymbol{ESS} = \sum_{i=1}^{n} \varepsilon_i^2 \quad = \sum_{1}^{n} (y_i - \alpha - \beta x_i)^2$$

ESS is at minimum when $\frac{\partial(ESS)}{\partial\alpha} = 0$ and $\frac{\partial(ESS)}{\partial\beta} = 0$

Computing $\hat{\beta}$; the least square estimate of β (regression coefficient)

$$\frac{\partial(ESS)}{\partial\beta} = \frac{\partial(\sum_1^n (y_i - \alpha - \beta x_i)^2)}{\partial\beta} = 2\sum_{1}^{n} x_i (y_i - \alpha - \beta x_i)^1 = 0$$

$$\sum_{1}^{n} x_i y_i - \alpha \sum_{1}^{n} x_i - \beta \sum_{1}^{n} x_i^2 = 0$$

$$\sum_{1}^{n} x_i y_i = \alpha \sum_{1}^{n} x_i + \beta \sum_{1}^{n} x_i^2 ---------(1)$$

Let $\hat{\alpha}$ be the least square estimate of α and

$$\frac{\partial(ESS)}{\partial\alpha} = \frac{\partial(\sum_1^n(y_i - \alpha - \beta x_i)^2)}{\partial\alpha} = 2\sum_1^n(y_i - \alpha - \beta x_i) \quad = 0$$

$$\frac{\partial(ESS)}{\partial\alpha} = \sum_1^n y_i - n\alpha - \beta\sum_1^n x_i = 0$$

$$\sum_1^n y_i = n\alpha + \beta\sum_1^n x_i - - - - - - - - - -(2)$$

Equation (1) and (2) are called the Normal Equations.

From (2)

$$\hat{\alpha} = \frac{\sum_1^n y_i}{n} - \widehat{\beta}\frac{\sum_1^n x_i}{n}$$

$$\boldsymbol{\widehat{\alpha} = \overline{y} - \widehat{\beta}\overline{x}}$$

Substituting for $\hat{\alpha}$ in equation (1)

$$\sum_1^n x_i y_i = \left(\frac{\sum_1^n y_i}{n} - \widehat{\beta}\frac{\sum_1^n x_i}{n}\right)\sum_1^n x_i + \hat{\beta}\sum_1^n x_i^2$$

$$\sum_1^n x_i y_i = \frac{\sum_1^n y_i \sum_1^n x_i}{n} - \hat{\beta}\frac{(\sum_1^n x_i)^2}{n} + \hat{\beta}\sum_1^n x_i^2$$

$$\hat{\beta} = \frac{n\sum_1^n x_i y_i - \sum_1^n x_i \sum_1^n y_i}{n\sum_1^n x_i^2 - (\sum_1^n x_i)^2}$$

$$\boldsymbol{\widehat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sum(x_i - \overline{x})^2}}$$

$$\hat{\alpha} = \bar{y} - \widehat{\beta}\bar{x} \qquad and \qquad \hat{\beta} = \frac{\sum_1^n x_i y_i - n\bar{x}\bar{y}}{\sum_1^n x_i^2 - n\bar{x}^2}$$

Example:

X	y
26.8	26.5
28.9	24.2
23.6	27.1
28.1	22.5
22.6	25.8
27.7	23.6
24.7	26.3
25.6	24.9

Confidence Interval for $\boldsymbol{\beta}$

It can be proved that under the assumption of $\varepsilon \sim N(0, \sigma^2)$; $\hat{\beta} \sim N(\beta, \frac{\sigma^2}{S_{xx}})$.

When $\sigma^2 is\ unknown$ it can be estimated by $\widehat{\sigma^2} = \frac{S_{yy} - \widehat{\beta} S_{xy}}{n-2}$.

$$P\left(t_{\frac{\alpha}{2}, n-2} \quad \leq \quad \frac{\widehat{\beta} - \beta}{\sqrt{\frac{\widehat{\sigma}^2}{S_{xx}}}} \quad \leq \quad t_{1-\frac{\alpha}{2}, n-2} \right) = (1 - \alpha)$$

Hypothesis Testing on Regression Coefficient

$H_0 : \beta = 0$

$H_1 : \ \beta \neq 0; regression\ coefficient\ is\ significantly\ different\ from\ 0$

Test Statistic and its distribution

$$\frac{\hat{\beta}-\beta}{\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}} \sim t_{n-2}$$

Analysis of Variance (ANOVA)

H_0 : $Regression\ Line\ does\ not\ fit\ the\ data\ well$

H_1 : $Regression\ Line\ fits\ the\ data\ well$

Source of variation	Sum of Squares (SS)	Degrees of Freedom (df)	Mean Sum of Squares(MS)	F-ratio	p-value (prob.>F)
REGRESSION **(estimation via reg. line)**	$\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$	**1**	**RSS/1**	$F_{cal} = \dfrac{\text{RSS/1}}{\text{ESS}/(n-2)} \sim F_{1,n-2}$	$P(F_{1,n-2} \geq F_{cal})$
ERROR(Residual) **(error in estimation)**	$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	**n-2**	**ESS/(n-2)**		
TOTAL (estimation + error)	$\sum_{i=1}^{n}(y_i - \bar{y})^2$	**n-1**			

Example :

Description: These data are on the production of power from wind mills. Direct Current (DC) output was measured against wind speed (in miles per hour).

Number of observations: 25

Variable	**Description**
output	Current output produced by the wind mill
speed	Windspeed (in miles per hour)

Source: Joglekar, G., Schuenemeyer, J.H. and LaRiccia, V. (1989) Lack-of-fit testing when replicates are not available, *American Statistician*, 43, pp. 135-143.

(speed,output)≡ (x, y)
0.123,2.45
0.500,2.70
0.653,2.90
0.558,3.05
1.057,3.40
1.137,3.60
1.144,3.95
1.194,4.10
1.562,4.60
1.582,5.00
1.501,5.45
1.737,5.80
1.822,6.00
1.866,6.20
1.930,6.35
1.800,7.00
2.088,7.40
2.179,7.85
2.166,8.15
2.112,8.80
2.303,9.10
2.294,9.55
2.386,9.70
2.236,10.00
2.310,10.20

ANOVA^a

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	134.282	1	134.282	160.257	.000^b
	Residual	19.272	23	.838		
	Total	153.554	24			

a. Dependent Variable: Electricity Production at the Wind Mill

b. Predictors: (Constant), Wind Speed (in miles per hour)