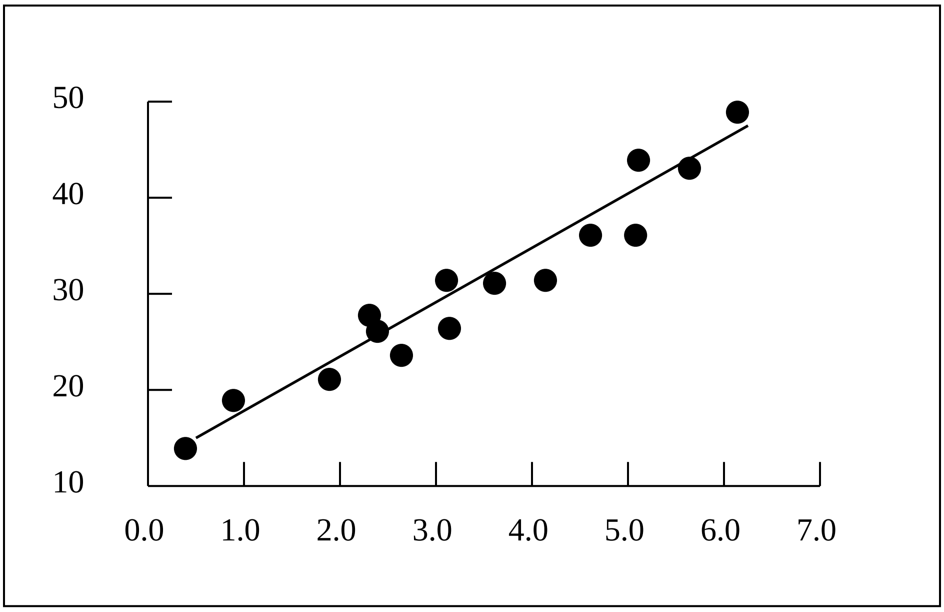
Simple Linear Regression



Prashan Rathnayaka
School of Computing,
IIT-Sri Lanka.

REFERENCE: *Applied Regression Analysis – Norman Draper, Harry Smith.*

INTRODUCTION

Modelling the relation ships

- We are concerned here with estimating the relationship between two or more variables when it is believed that some form of association exists between these variables.

Examples

1) Solid removed from a material (y) is thought to be related to the drying time $(\boldsymbol{x})$. Ten observations obtained from an experimental study are given below:

$\boldsymbol{y}$	4.3	1.5	1.8	4.9	4.2	4.8	5.8	6.2	7.0	7.9
$\boldsymbol{x}$	2.5	3.0	3.5	4.0	4.5	5.0	5.5	6.0	6.5	7.0

2) The electric power consumed each month by a chemical plant is thought to be related to the average ambient temperature (x_1), the number of days in the month (x_2) and the average product purity (x_3). The past year's historical data are available and are presented in the following table.

Y	x_1	x_2	x_3
240	25	24	91
236	31	21	90
270	45	24	88
274	60	25	87
301	65	25	91
316	72	26	94
300	80	25	87
296	84	25	86
267	75	24	88
276	60	25	91
288	50	25	90
261	38	23	89

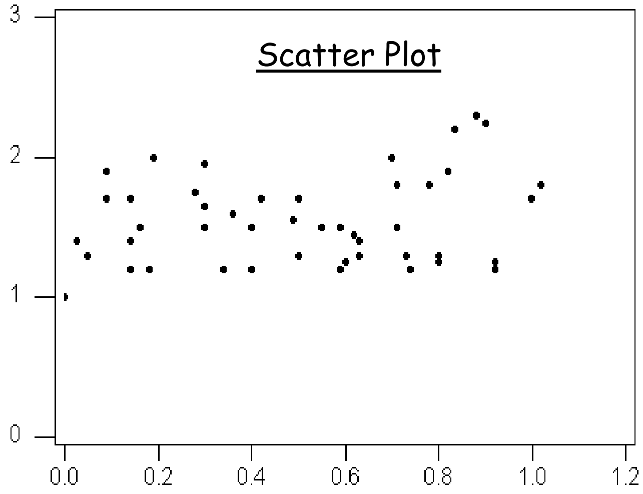
For these data the regression model is given by

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon.$$

- Looking for associations is not our main concern here. We are interested in determining the *exact form of the relationship* that exists between the variables of interest. Whenever possible we try to express relationships between variables by a mathematical equation. We will be concerned with the case when there is a variable 'Y' which is believed to be depended on one or more other variables say $X_1, X_2,\ldots,X_n$.

- The term regression is used to describe the relationship between the 'Y' and the 'X's. Y is called the *dependent variable or response variable*. X's are concerned with the dependence of a random variable Y on X's which are variables but *not random variables,* an equation that relates Y to X's is usually called a *regression equation*.

- We decide the mathematical equation of the relationship of the variables by direct inspection of the data. We plot the data and decide the kind of curve which will best describe the overall pattern of the data. This plot is called the '*scatter plot'*.



Simple Linear Regression

The simplest regression model is the *simple linear regression model.*

Only one independent variable

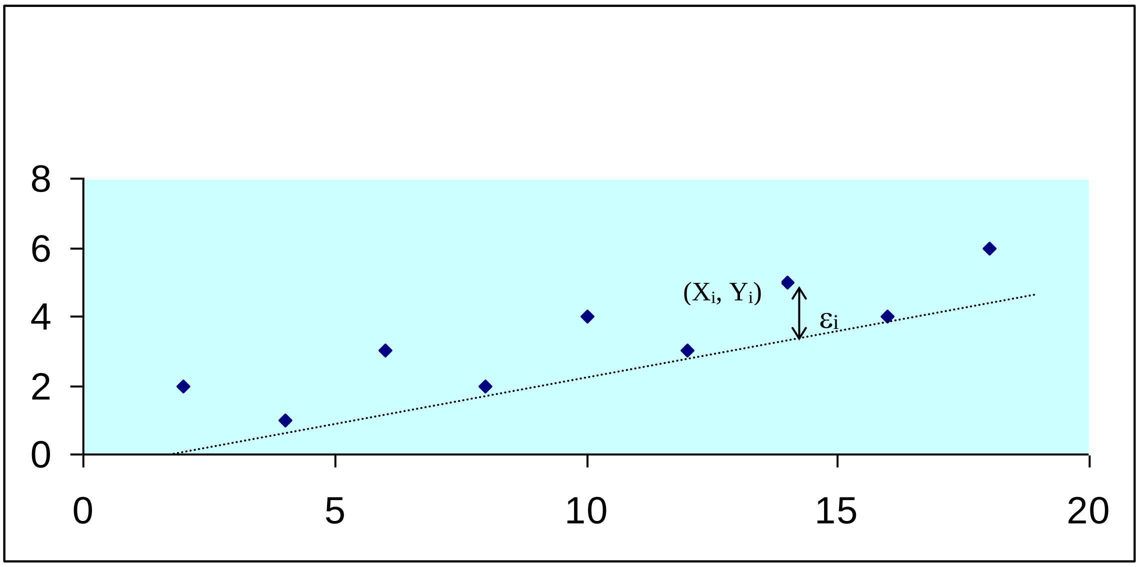
Relationship is linear

The model is

$$Y = \alpha + \beta x + \varepsilon$$

When α & β are parameters,

ε - Random error or random noise or residuals



We can draw any number of lines, but we must find the line which in some senses provides the best fit to the data, which yields the best possible predictions. The method we use to find the best fit to the data is the *least square method*. That is, we find the best fit by minimizing the sum of squares of the residuals. That is minimizing the sum of squares of the vertical distances from the points to the line.

Find the Linear Association Between Two Variables

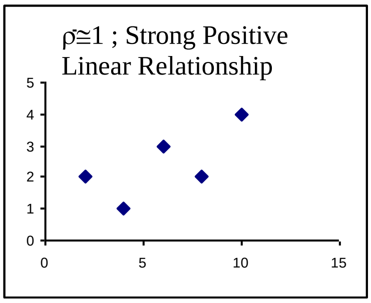
Let X and Y be the two variables. We use correlation coefficients denoted by ρ or γ and

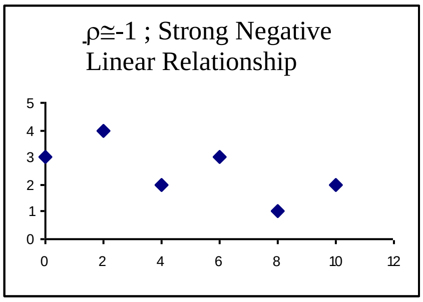
$$\rho = \frac{COV(X,Y)}{\sqrt{[V(X)]}\sqrt{[V(Y)]}}$$ where

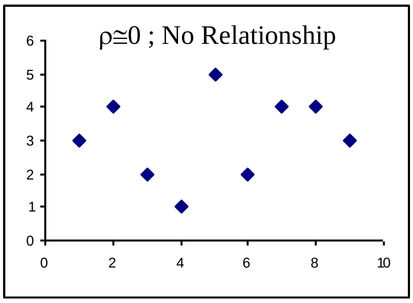
$$COV(X,Y) = E[(X-\overline{X})(Y-\overline{Y})]$$

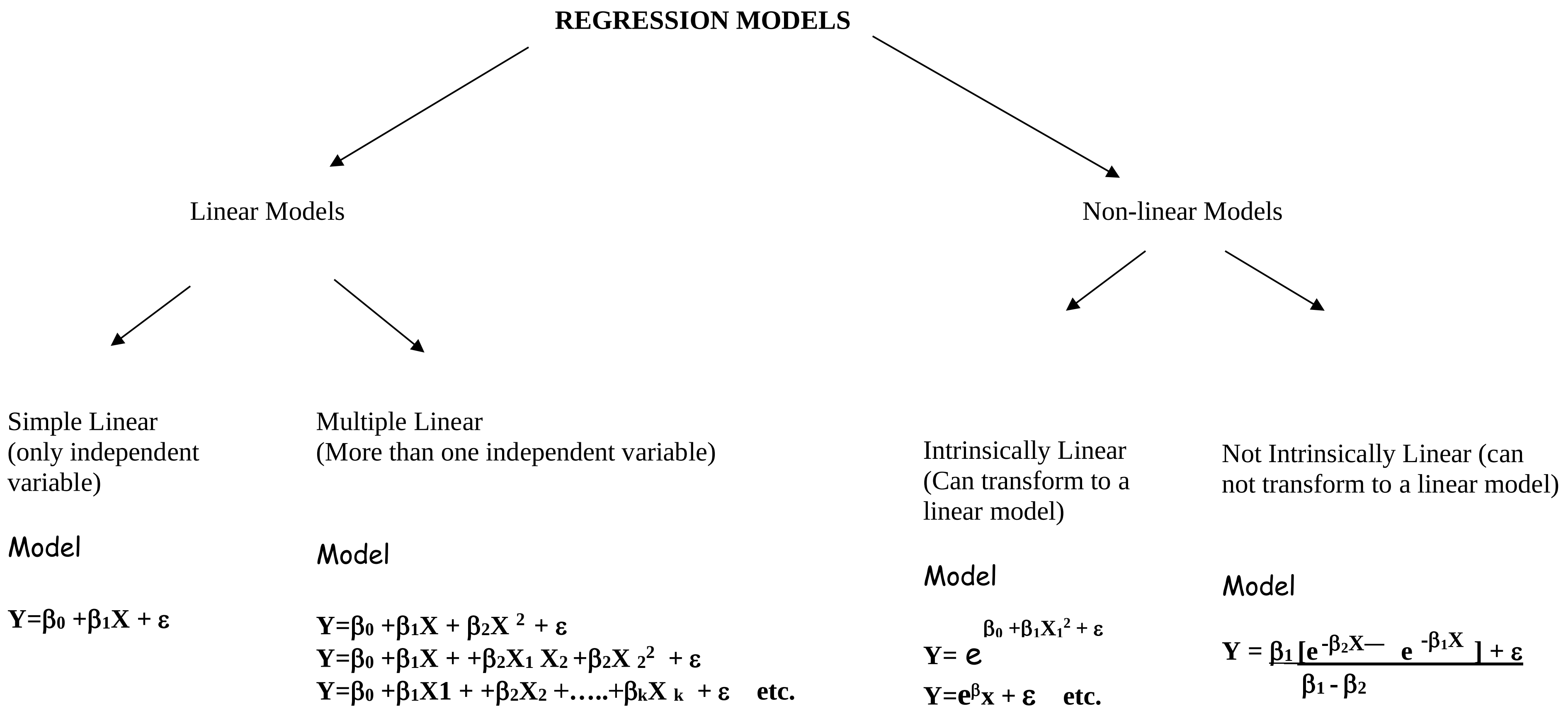
$$V(X) = E(X-\overline{X})^2$$

$$V(Y) = E(Y-\overline{Y})^2 \quad ; \quad (-1 \le \rho \le 1)$$


ρ≅1 ; Strong Positive Linear Relationship


ρ≅-1 ; Strong Negative Linear Relationship


ρ≅0 ; No Relationship



REGRESSION MODELS

Linear Models

Non-linear Models

Simple Linear
(only independent
variable)

Model

$Y=\beta_0+\beta_1X+\varepsilon$

Multiple Linear
(More than one independent variable)

Model

$Y=\beta_0+\beta_1X+\beta_2X^2+\varepsilon$

$Y=\beta_0+\beta_1X++\beta_2X_1X_2+\beta_2X_2^2+\varepsilon$

$Y=\beta_0+\beta_1X1++\beta_2X_2+\ldots..+\beta_kX_k+\varepsilon$ etc.

Intrinsically Linear
(Can transform to a
linear model)

Model

$Y=e^{\beta_0+\beta_1X_1^2+\varepsilon}$

$Y=e^{\beta}x+\varepsilon$ etc.

Not Intrinsically Linear (can
not transform to a linear model)

Model

$Y=\frac{\beta_1[e^{-\beta_2X}-e^{-\beta_1X}]+\varepsilon}{\beta_1-\beta_2}$

SIMPLE LINEAR REGRESSION

Regression model;

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Suppose we have available n sets of observations $(X_1, Y_1), \quad (X_2, Y_2), \quad \ldots, \quad (X_n, Y_n)$

Then,

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad i = 1, 2, \ldots, n$$

Least Squares Estimation

- The parameters β_0 and β_1 are estimated by the method of least squares.
- From the many straight lines that can be drawn through a scatter gram we wish to pick the one that 'best fits' the data.
- The fit is 'best' in the sense that the values of β_0 and β_1 chosen are those that minimize the sum of squares of the residuals.

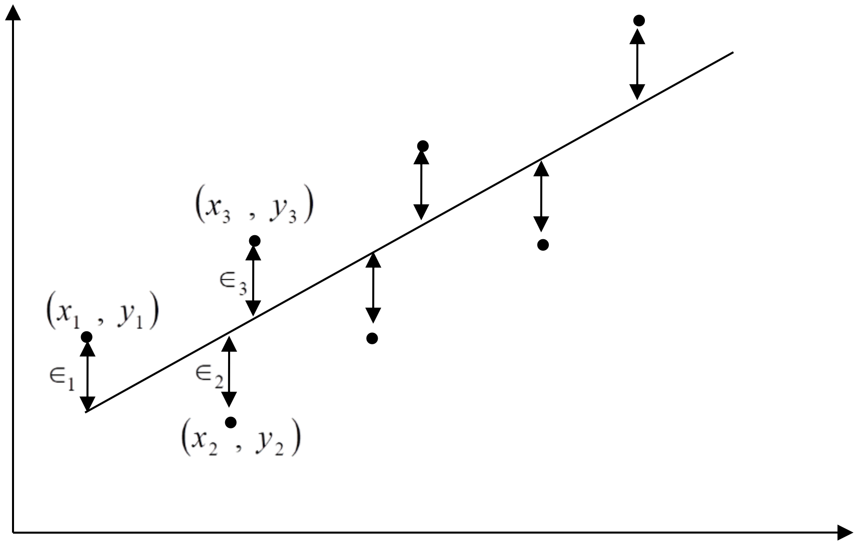


Fig 1.1: The Least Squares procedure minimizes the sum of the residuals ε_i

In this way we are essentially picking the line that comes as close as it can to all data points simultaneously. For example, if we consider the sample of five data points shown in Fig 1.1, then the least-squares procedure selects that line which causes $\varepsilon_1^2 + \varepsilon_2^2 + \varepsilon_3^2 + \varepsilon_4^2 + \cdots$ be as small as possible.

The residuals are squared before summing for a very practical reason. Notice that the residuals for a data point that lies above the estimated regression line is positive; for a point that lies below the line is negative. If the residuals themselves are summed, the negative and positive values will counter act one another and the sum will always be 0.

Least Square Method

Residual sum of squares S,

$$S = \sum_{i=1}^{n} \varepsilon_i^2$$

$$\varepsilon_i = Y_i - \beta_0 - \beta_1 X_i$$

$$S = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

To find the best line which fits the data, we have to find β_0 and β_1 so that $\sum \varepsilon_i^2$ is a minimum. We will denote the estimated values of β_0 and β_1 by $\overset{\wedge}{\beta}_0$ and $\overset{\wedge}{\beta}_1$.

$$\overset{\wedge}{\beta}_0 = \overline{Y} - \overset{\wedge}{\beta}_1 \overline{X}$$

$$\overset{\wedge}{\beta}_1 = \frac{\sum_{i=1}^{n} X_i Y_i - n\overline{X}\overline{Y}}{\sum_{i=1}^{n} X_i^2 - n\overline{X}^2}$$

We can show that $\overset{\wedge}{\beta}_1$ can be expressed as;

(I) $$\overset{\wedge}{\beta}_1 = \frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2}$$

(II) $$\overset{\wedge}{\beta}_1 = \frac{\sum_{i=1}^{n} Y_i \left(X_i - \overline{X}\right)}{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2}$$

So the fitted regression line is

$$\hat{Y} \quad = \quad \hat{\beta}_0 + \hat{\beta}_1 X$$

- To estimate the parameters in the model we do not have to make any assumption.
- But after we estimate, we need to know the distance between the estimator and the parameter.
- We want this distance to be small with high probability.
- So, we need to find the confidence intervals for the estimators.
- For this we want to know the distribution of the estimators.

So for the rest of the work, we need some basic assumptions in the model.

Model

$$Y = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad ; i = 1, 2, 3, \ldots, n$$

Random variables

β_0 and β_1 are estimators. So, they are random variables.

Assumptions:

(i) $E(\varepsilon_i) = 0$ for all i

(ii) $Cov(\varepsilon_i, \varepsilon_j) = 0$; $i \neq j$ uncorrelated

$Var(\varepsilon_i) = \sigma^2$ or $E(\varepsilon_i^2) = \sigma^2$ for all i

(iii) $\varepsilon_i \sim N(0, \sigma^2)$