LINEAR REGRESSION

CONTENTS

- Introduction
- Regression –Definition
- Linear Regression
- Scatter Graph
- Slope and Intercept
- Least square method
- Example

Introduction

- Analyze the specific relationships between the two or more variables .

- This is done to gain the information about one through knowing values of the others

Regression

- A statistical measure that attempts to determine the strength of the relationship between one dependent variable (usually denoted by Y) and a series of other changing variables (known as independent variables).

- Forecast value of a dependent variable (Y) from the value of independent variables (X_1, X_2,....).

Regression Analysis

- In statistics, regression analysis includes many techniques for modeling and analyzing several variables, when the focus is on the relationship between a dependent variable and one or more independent variables
- Regression analysis is widely used for prediction and forecasting,

Dependent & independent variable

- Independent variables are regarded as inputs to a system and may take on different values freely.
- Dependent variables are those values that change as a consequence of changes in other values in the system.
- Independent variable is also called as predictor or explanatory variable and it is denoted by X.
- Dependent variable is also called as response variable and it is denoted by Y.

Linear regression

- The simplest mathematical relationship between two variables x and y is a linear relationship.
- In a cause and effect relationship, the independent variable is the cause, and the dependent variable is the effect.
- Least squares linear regression is a method for predicting the value of a dependent variable *Y*, based on the value of an independent variable *X*.

The first order linear model

$$Y = b_0 + b_1 X + \epsilon$$

Y = dependent variable

X = independent variable

b_0 = Y-intercept

b_1 = slope of the line

e = error variable

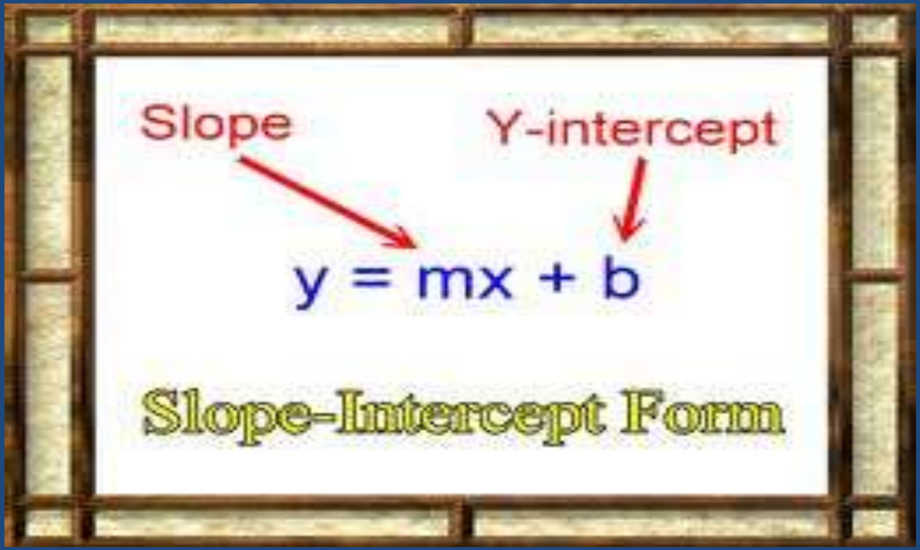
Slope & Intercept

SLOPE:

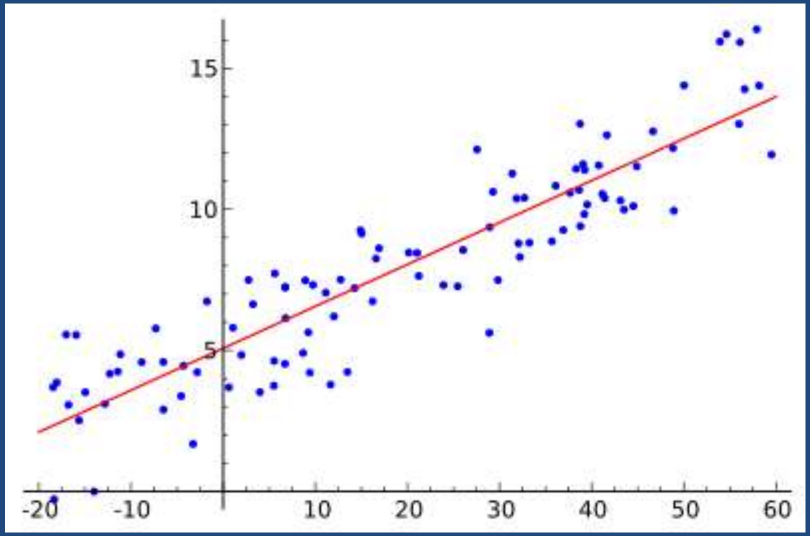
The slope of a line is the change in y for a one unit increase in x.

Y-Intercept:

It is the height at which the line crosses the vertical axis and is obtaining by setting x=0 in the equation.



EXAMPLE



- Example of simple linear regression which has one independent variable.

Error variable

- Random error term:

  1.The quantity Є in the model equation is a random varible assumed to be normally distributed with

  $E(\epsilon)=0$ and $V(\epsilon)=\sigma^2$

  2.Є-random deviation or random error term.

  3. Without Є ,any observed pair (x,y) would correspond to a point falling exactly on the line

  $Y=b_0 + b_1 X$, called true regression line.

  The inclusion of the random error term allows (x,y) to fall either above the true regression line (when $\epsilon>0$) or below the line (when $\epsilon<0$).

Scatter plot

Definition of Scatter Plot

1. Scatter plot or Scattergraph is a type of mathematical diagram to display values for two variables for a set of data.

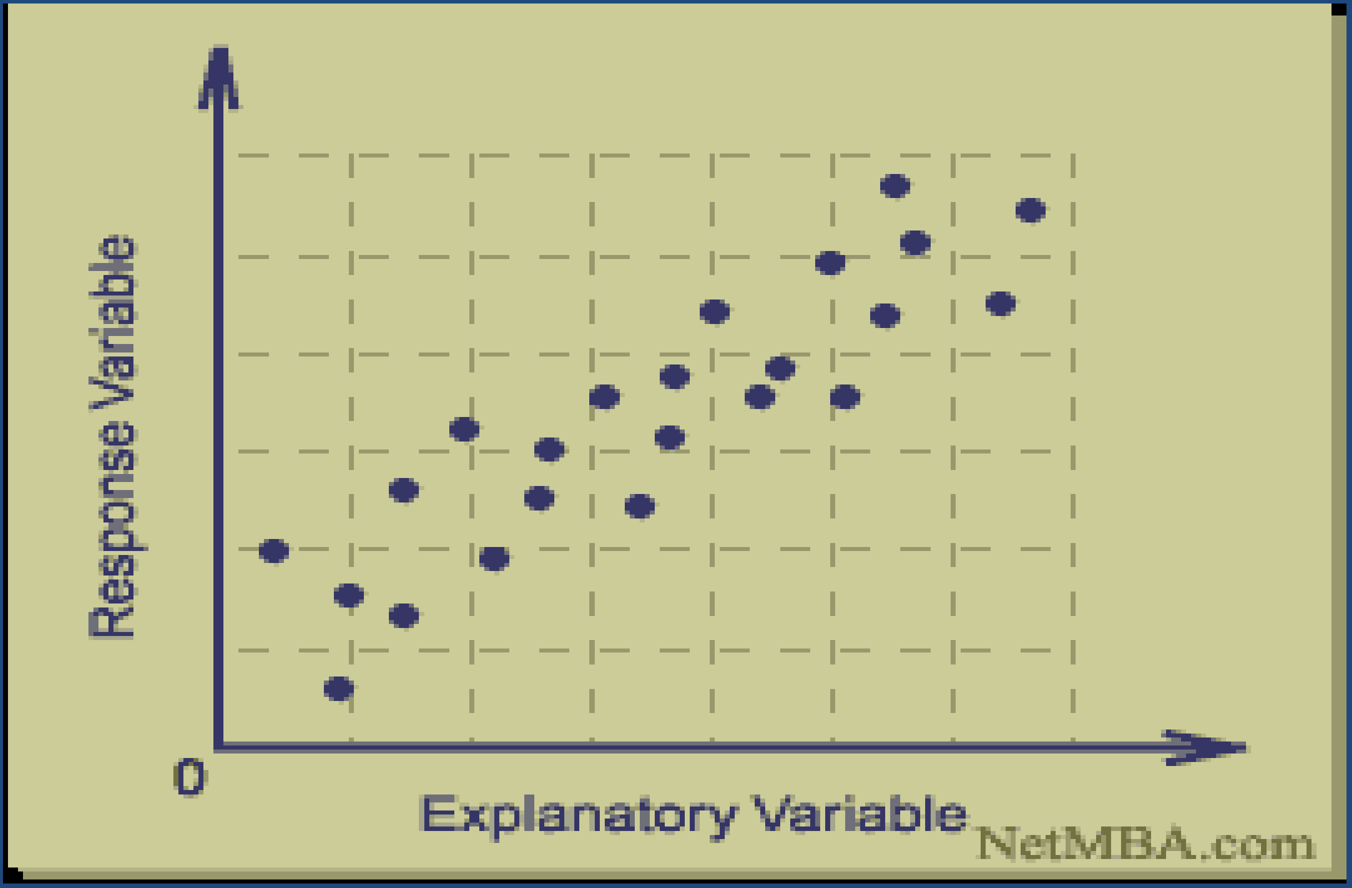
2. A scatter plot is a graph made by plotting ordered pairs in a coordinate plane to show the correlation between two sets of data.

3. The data is displayed as a collection of points,

More about Scatter Plot

- A scatter plot describes a positive trend if, as one set of values increases, the other set tends to increase.
- A scatter plot describes a negative trend if, as one set of values increases, the other set tends to decrease.
- The position on the vertical axis. This kind of plot is also called a *scatter chart*, *scattergram*, *scatter diagram* or *scatter graph*.

Scatter graph



Least Squares Estimation of b_0, b_1

- $\beta_0 \equiv$ Mean response when x=0 (y-intercept)
- $\beta_1 \equiv$ Change in mean response when x increases by 1 unit (slope)
- β_0, β_1 are unknown parameters (like μ)
- $\beta_0+\beta_1 x \equiv$ Mean response when explanatory variable takes on the value x
- Goal: Choose values (estimates) that minimize the sum of squared errors (*SSE*) of observed values to the straight-line:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \qquad SSE = \sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2 = \sum_{i=1}^{n}\left(y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right)\right)^2$$

The least squares estimate of the slope coefficient β_1 of true regression line is

$$\beta_1 = \frac{\Sigma(X_i - X')(Y_i - Y')}{\Sigma\,(X_i - X')^2}$$

The least squares estimate of the intercept β_0 of true regression line is

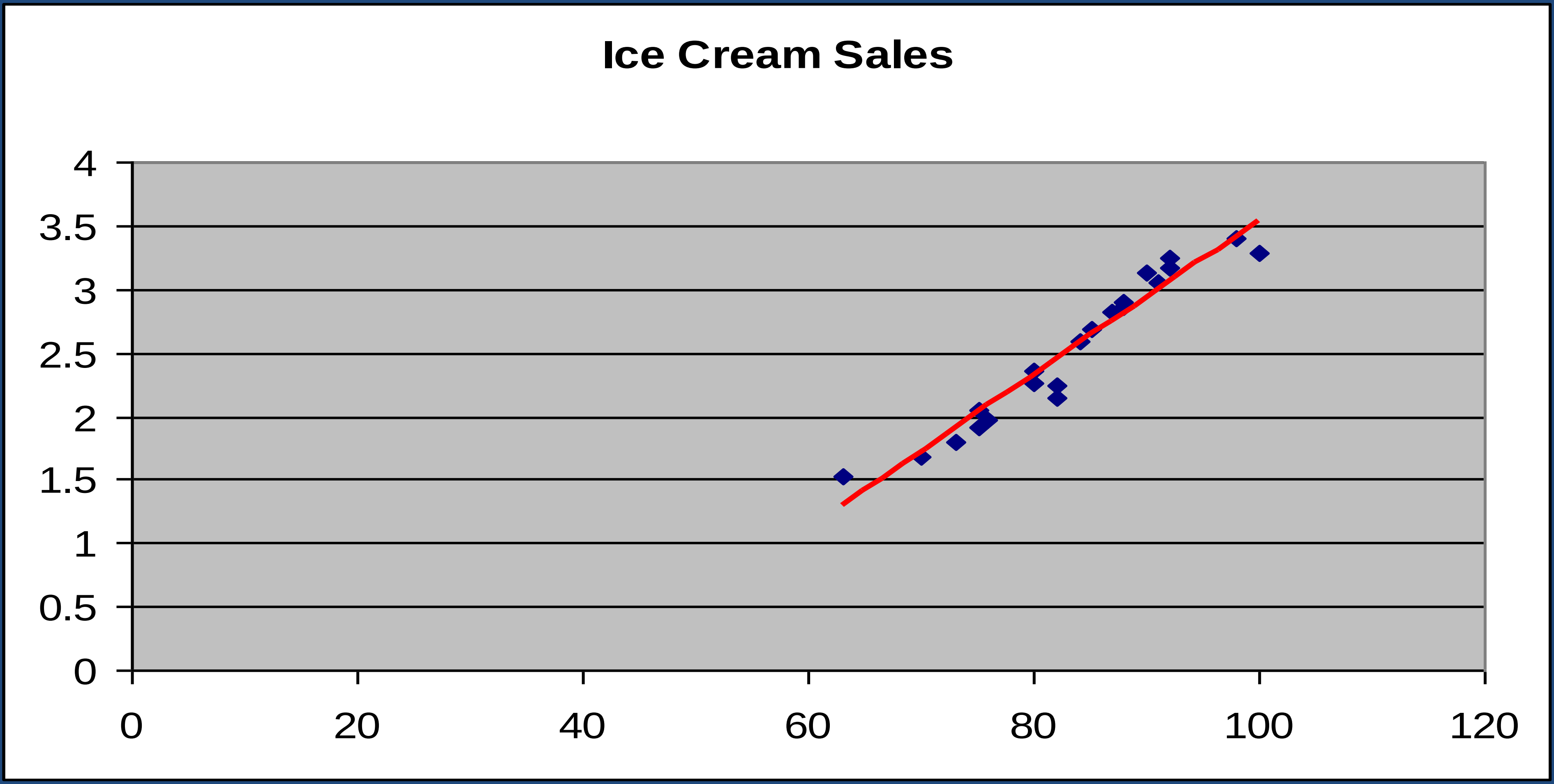
$$\beta_0 = Y' - \beta_1 x'$$

- Regression generates what is called the "least-squares" regression line.
- The regression line takes the form: = a + b*X, where a and b are both constants, (pronounced y-hat) is the predicted value of Y and X is a specific value of the independent variable.
- Such a formula could be used to generate values of for a given value of X. For example, suppose a = 10 and b = 7. If X is 10, then the formula produces a predicted value for Y of 45 (from 10 + 5*7).
- It turns out that with any two variables X and Y, there is one equation that produces the "best fit" linking X to Y.
- We use the criterion is called the least squares criterion to measure best .

- You can imagine a formula that produces predictions for Y from each value of X in the data. Those predictions will usually differ from the actual value of Y that is being predicted (unless the Y values lie exactly on a straight line).
- If you square the difference and add up these squared differences across all the predictions,
- you get a number called the residual or error sum or squares (or SS_{error}). The formula above is simply the mathematical representation of SS_{error}. Regression generates a formula such that SS_{error} is as small as it can possibly be
- Minimising this number (by using calculus) minimises the average error in prediction.

- **Example:**

X Temperature	Y Sales
63	1.52
70	1.68
73	1.8
75	2.05
80	2.36
82	2.25
85	2.68
88	2.9
90	3.14
91	3.06
92	3.24
75	1.92
98	3.4
100	3.28
92	3.17
87	2.83
84	2.58
88	2.86
80	2.26
82	2.14
76	1.98


Ice Cream Sales
4
3.5
3
2.5
2
1.5
1
0.5
0
0
20
40
60
80
100
120

Most applications of linear regression :

- If the goal is prediction, or forecasting, linear regression can be used to fit a predictive model to an observed data set of *y* and *X* values.
- After developing such a model, if an additional value of *X* is then given without its accompanying value of *y*, the fitted model can be used to make a prediction of the value of *y*.
- Given a variable *y* and a number of variables $X_1, \ldots, X_p$ that may be related to *y*, linear regression analysis can be applied to quantify the strength of the relationship between *y* and the X_j, to assess which X_j may have no relationship with *y* at all, and to identify which subsets of the X_j contain redundant information about *y*.

THANK YOU