Lecture 10

Linear Regression

Regression is a technique used to predict the value of a response (dependent) variables, from one or more predictor (independent) variables, where the variable are numeric.

For example, you can observe several employees of some company and try to understand how their salaries depend on the **features**, such as experience, level of education, role, city they work in, and so on.

This is a regression problem where data related to each employee represent one **observation**. The presumption is that the experience, education, role, and city are the independent features, while the salary depends on them.

Similarly, you can try to establish a mathematical dependence of the prices of houses on their areas, numbers of bedrooms, distances to the city center, and so on.

Generally, in regression analysis, you usually consider some phenomenon of interest and have a number of observations. Each observation has two or more features. Following the assumption that (at least) one of the features depends on the others, you try to establish a relation among them.

In other words, **you need to find a function that maps some features or variables to others sufficiently well**.

The dependent features are called the **dependent variables**, **outputs**, or **responses**.

The independent features are called the **independent variables**, **inputs**, or **predictors**.

Regression problems usually have one continuous and unbounded dependent variable. The inputs, however, can be continuous, discrete, or even categorical data such as gender, nationality, brand, and so on.

There are various forms of regression such as linear, multiple, logistic, polynomial, non-parametric, etc.

Linear and Logistic regression are the most basic form of regression which are commonly used. The essential difference between these two is that Logistic regression is used when the dependent variable is binary in nature. In contrast, Linear regression is used when the dependent variable is continuous and nature of the regression line is linear.

Definition of Linear Regression

The **linear regression** technique involves the continuous dependent variable and the independent variables can be continuous or discrete. By using best fit straight line linear regression sets up a relationship between dependent variable (Y) and one or more independent variables (X). In other words, there exist a linear relationship between independent and dependent variables.

Linear Regression uses a linear function to map input variables to continuous response/dependent variables. Once fitted, a Linear Regression model can be used to predict the values of response/dependent variables for new values of the input variables. An example application, in a financial trading analytics context, might be to predict the total value of trades that will occur by end of day based on the number and size of orders that have been submitted so far. The output of Linear Regression is a continuous value, and due to the use of a straight line to map the input variables to the dependent variables, the output can be any one of an infinite number of possibilities. This means that outputs can be positive or negative, with no maximum or minimum bounds.

Objective for linear regression is to make a line that is able to be as close as possible to all the points.

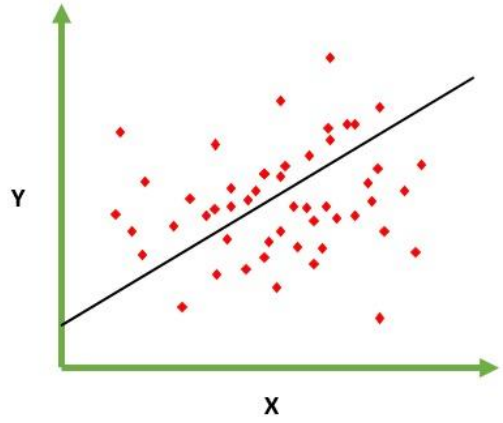
The difference between linear and multiple linear regression is that the linear regression contains only one independent variable while multiple regression contains more than one independent variables. The best fit line in linear regression is obtained through least square method.

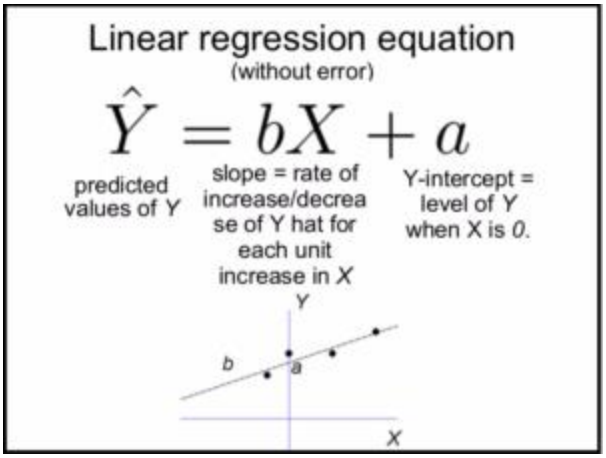
The following equation is used to represent a linear regression model:

$$Y = b_0 + b_1 \times X + e$$

Where $\mathbf{b_0}$ is the intercept, $\mathbf{b_1}$ is the slope of the line and **e** is the error. Here **Y** is dependent variable and **X** is an independent variable.

The following graph can be used to show the linear regression model.

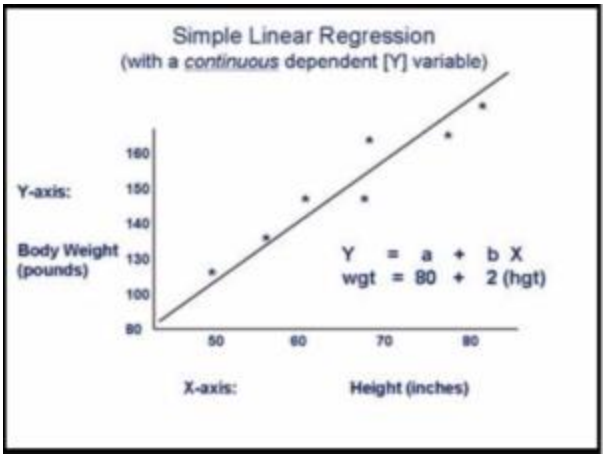


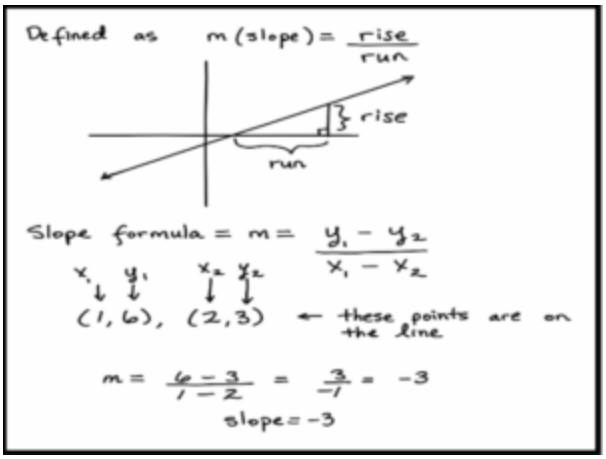


$$Slope \quad a = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

$$\text{int}ercept \quad b = \bar{y} - a\bar{x}$$

$$\_\_where \quad overbar \quad denotes \quad average$$





Applications:

Below are some real world applications of Simple Linear Regression:

1. Linear Regression can be used to predict the sale of products in the future based on past buying behaviour.

2. Economists use Linear Regression to predict the economic growth of a country or state.

3. Sports analyst use linear regression to predict the number of runs or goals a player would score in the coming matches based on previous performances.

4. An organisation can use linear regression to figure out how much they would pay to a new joinee based on the years of experience.

5. Linear regression analysis can help a builder to predict how much houses it would sell in the coming months and at what price.

Advantages of linear regression:

When we know the relationship between the independent and dependent variable have a linear relationship, this algorithm is the best to use because it's the least complex to compared to other algorithms that also try finding the relationship between independent and dependent variable.

Disadvantage:

1. By its definition, linear regression only models relationships between dependent and independent variables that are linear. It assumes there is a straight-line

relationship between them which is incorrect sometimes. Linear regression is very sensitive to the anomalies in the data (or outliers).

2. Take for example most of your data lies in the range 0-10. If due to any reason only one of the data item comes out of the range, say for example 15, this significantly influences the regression coefficients.