Supervised Learning
(Linear Regression)

Dr. Virendra Singh Kushwah
Assistant Professor Grade-II
School of Computing Science and Engineering
Virendra.Kushwah@vitbhopal.ac.in
7415869616

Lecture-3

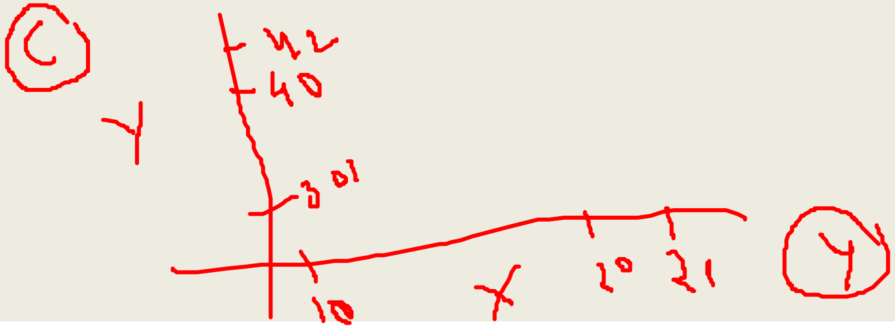
- Machine Learning Types
- Supervised Learning
- Unsupervised Learning

What is Regression?

- The main goal of regression is the construction of an efficient model to predict the dependent attributes from a bunch of attribute variables. A regression problem is when the output variable is either real or a continuous value i.e salary, weight, area, etc.

- We can also define regression as a statistical means that is used in applications like housing, investing, etc. It is used to predict the relationship between a dependent variable and a bunch of independent variables.

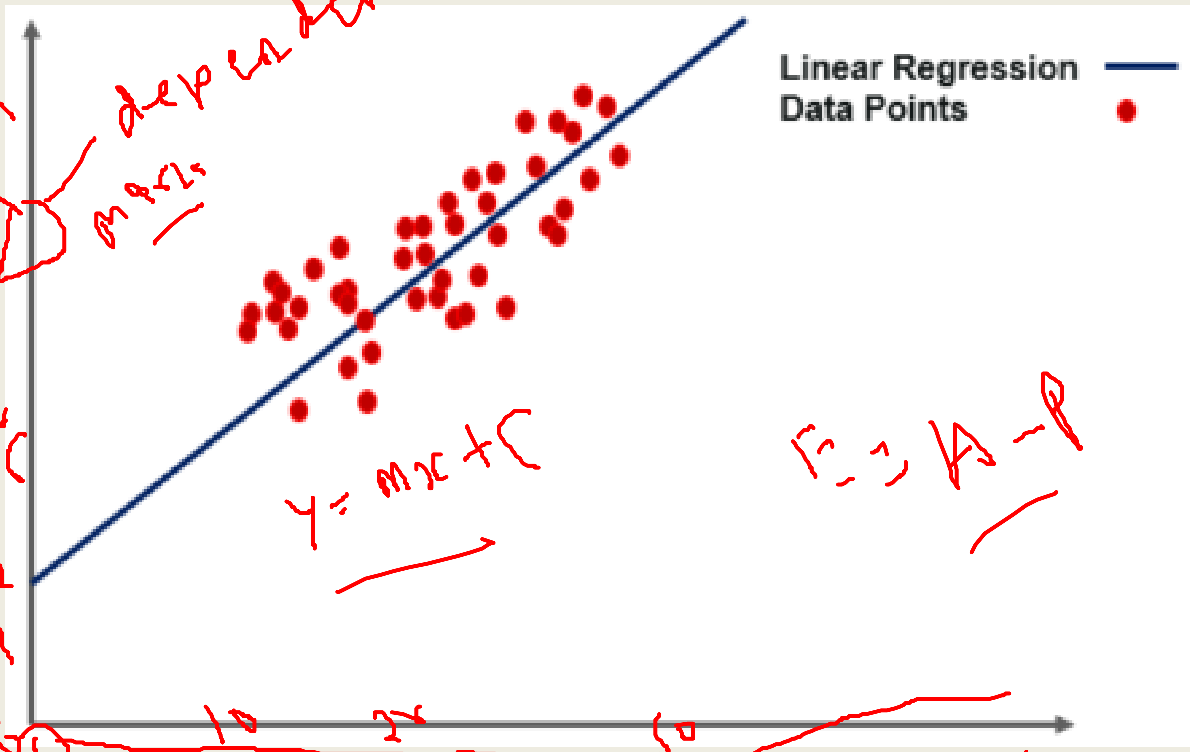
Examples



VIT BHOPAL
www.vitbhopal.ac.in

Simple Linear Regression

- One of the most interesting and common regression technique is simple linear regression. In this, we predict the outcome of a dependent variable based on the independent variables, the relationship between the variables is linear. Hence, the word linear regression.



dependent

$y = mx + c$

Independent

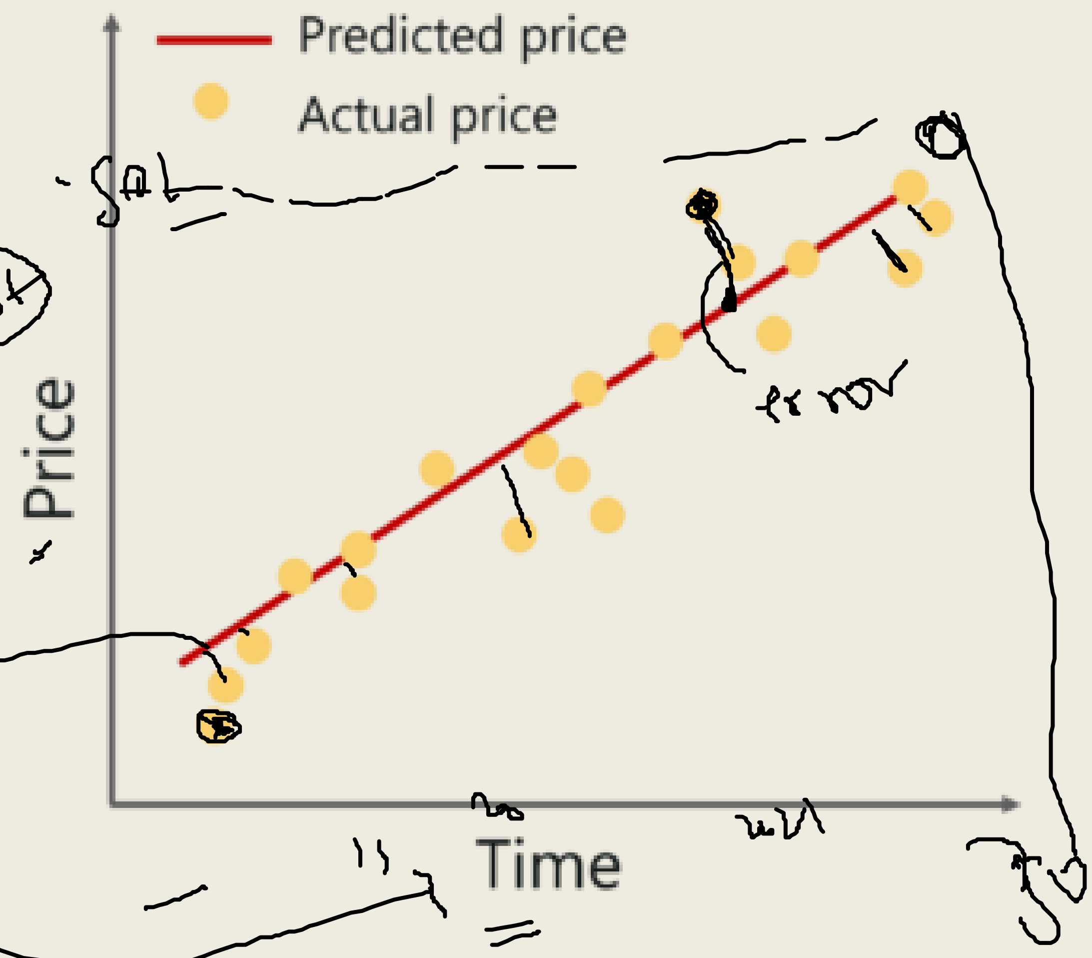
- Simple linear regression is a regression technique in which the independent variable has a linear relationship with the dependent variable. The straight line in the diagram is the best fit line.
- The main goal of the simple linear regression is to consider the given data points and plot the best fit line to fit the model in the best way possible.

What is the Line Of Best Fit?

- Line of best fit is drawn to represent the relationship between 2 or more variables. To be more specific, the best fit line is drawn across a scatter plot of data points in order to represent a relationship between those data points.
- Regression analysis makes use of mathematical methods such as least squares to obtain a definite relationship between the predictor variable (s) and the target variable. The least-squares method is one of the most effective ways used to draw the line of best fit. It is based on the idea that the square of the errors obtained must be minimized to the most possible extent and hence the name least squares method.

$Y = mx + c + e$

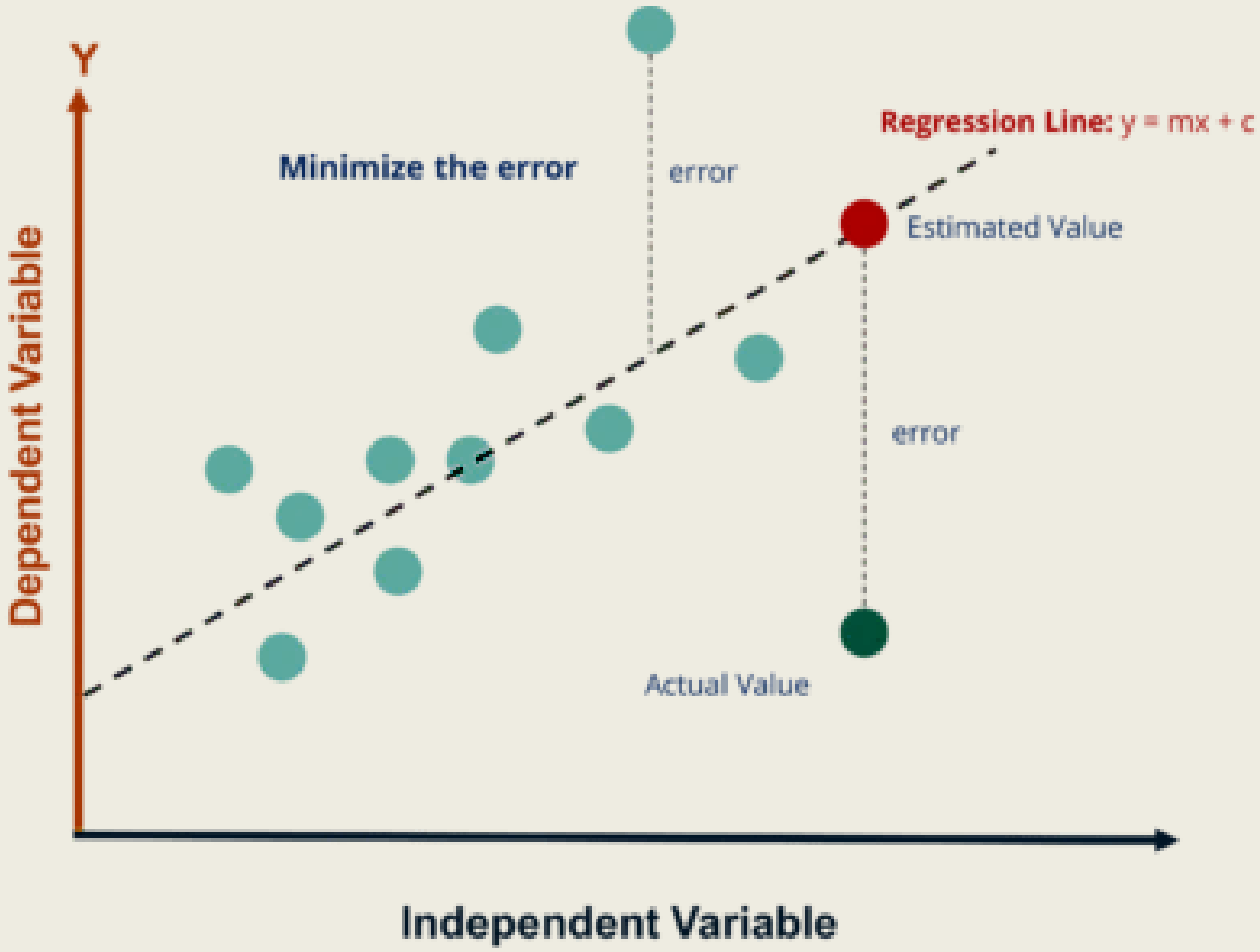
- If we were to plot the best fit line that shows the depicts the sales of a company over a period of time, it would look something like this:
- Notice that the line is as close as possible to all the scattered data points. This is what an ideal best fit line looks like.
- To better understand the whole process let's see how to calculate the line using the Least Squares Regression.



Cost

$Y = mx + c$

$$\frac{(e_1 + e_2 + \cdots e_n)^2}{n}$$


Y
Regression Line: y = mx + c
Minimize the error
error
Estimated Value
error
Dependent Variable
Actual Value
Independent Variable

Steps to calculate the Line of Best Fit

- To start constructing the line that best depicts the relationship between variables in the data, we first need to get our basics right. Take a look at the equation below:

$$y = mx + c$$

- Surely, you've come across this equation before. It is a simple equation that represents a straight line along 2 Dimensional data, i.e. x-axis and y-axis. To better understand this, let's break down the equation:

- y: dependent variable
- m: the slope of the line
- x: independent variable
- c: y-intercept

- So the aim is to calculate the values of slope, y-intercept and substitute the corresponding ‘x’ values in the equation in order to derive the value of the dependent variable.

- Let’s see how this can be done.
- As an assumption, let’s consider that there are ‘n’ data points.

- Step 1: Calculate the slope ‘m’ by using the following formula:

$$m = \frac{n\sum xy - (\Sigma x)(\Sigma y)}{n\Sigma x^2 - (\Sigma x)^2}$$

- Step 2: Compute the y-intercept (the value of y at the point where the line crosses the y-axis):

$$c = y - mx$$

- Step 3: Substitute the values in the final equation:

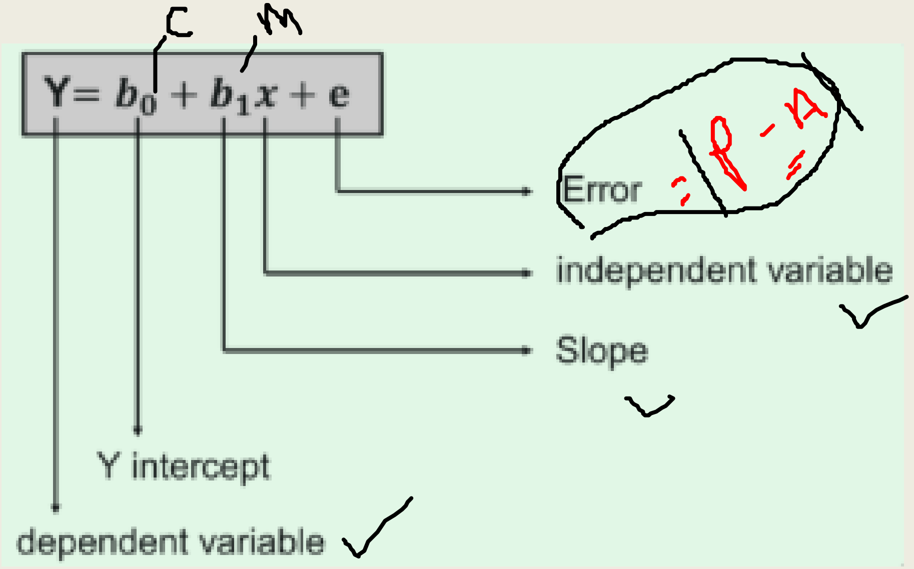
$$y = mx + c$$

Linear Regression Terminologies

- Cost Function

$$mSE = \frac{\sum_{i=1}^{N} (P-A)^2}{N}$$

$$mSE \quad \frac{\sum_{i=0}^{n} (P_i - A_i)^2}{n}$$



- The dependent variable that is to be predicted is denoted by Y.
- A line that touches the y-axis is denoted by the intercept b_0.
- b_1 is the slope of the line, x represents the independent variables that determine the prediction of Y.
- The error in the resultant prediction is denoted by e.

- The cost function provides the best possible values for b_0 and b_1 to make the best fit line for the data points. We do it by converting this problem into a minimization problem to get the best values for b_0 and b_1. The error is minimized in this problem between the actual value and the predicted value.
- We choose the function above to minimize the error. We square the error difference and sum the error over all data points, the division between the total number of data points. Then, the produced value provides the averaged square error over all data points.
- It is also known as MSE(Mean Squared Error), and we change the values of b_0 and b_1 so that the MSE value is settled at the minimum.

$$minimize \frac{1}{n}\sum_{i=1}^{n}(pred_i - y_i)^2$$

$$J = \frac{1}{n}\sum_{i=1}^{n}(pred_i - y_i)^2$$

Actual value

Pred

Advantages And Disadvantages

Advantages	**Disadvantages**
Linear regression performs exceptionally well for linearly separable data	The assumption of linearity between dependent and independent variables
Easier to implement, interpret and efficient to train	It is often quite prone to noise and overfitting
It handles overfitting pretty well using dimensionally reduction techniques, regularization, and cross-validation	Linear regression is quite sensitive to outliers
One more advantage is the extrapolation beyond a specific data set	It is prone to multicollinearity

Where is Linear Regression Used?

- 1. Evaluating Trends and Sales Estimates

- Linear regressions can be used in business to evaluate trends and make estimates or forecasts.
- For example, if a company’s sales have increased steadily every month for the past few years, conducting a linear analysis on the sales data with monthly sales on the y-axis and time on the x-axis would produce a line that that depicts the upward trend in sales. After creating the trend line, the company could use the slope of the line to forecast sales in future months.

- 2. Analyzing the Impact of Price Changes

- Linear regression can also be used to analyze the effect of pricing on consumer behavior.

- For example, if a company changes the price on a certain product several times, it can record the quantity it sells for each price level and then performs a linear regression with quantity sold as the dependent variable and price as the explanatory variable. The result would be a line that depicts the extent to which consumers reduce their consumption of the product as prices increase, which could help guide future pricing decisions.

- 3. Assessing Risk