Simple Linear Regression

- This algorithm is used to find the relationship between 2 continuous variables (one independent variable and one dependent variable)
- It is both a statistical algorithm and a machine learning algorithm.
- Linear regression is a linear model, e.g. a model that assumes a linear relationship between the input variables (x) and the single output variable (y). More specifically, y can be calculated from a linear combination of the input variables (x).
- The equation is Y= W1*X+b where,
  Y = Predicted value/independent Value
  W1 = Gradient/slope/Weight
  X = Input/ Dependent value
  b = Bias

Simple Linear Regression

- The equation is Y= W1*X+b The equation is the same as that of a straight line (Y=MX+c)
- The core idea is to obtain a line that best fits the data. The best fit line is the one for which total prediction error (all data points) are as small as possible. Error is the distance between the point to the regression line.
- The values W1 and b must be chosen so that they minimize the error. If the sum of squared error(explained in performance metrics series) is taken as a metric to evaluate the model, the goal to obtain a line that best reduces the error.
- We can train this algorithm using multiple methods, we can use statistics to calculate the W1 and b, or we can use ordinary least squares method to calculate W1 and b or we can use gradient descent to do that.

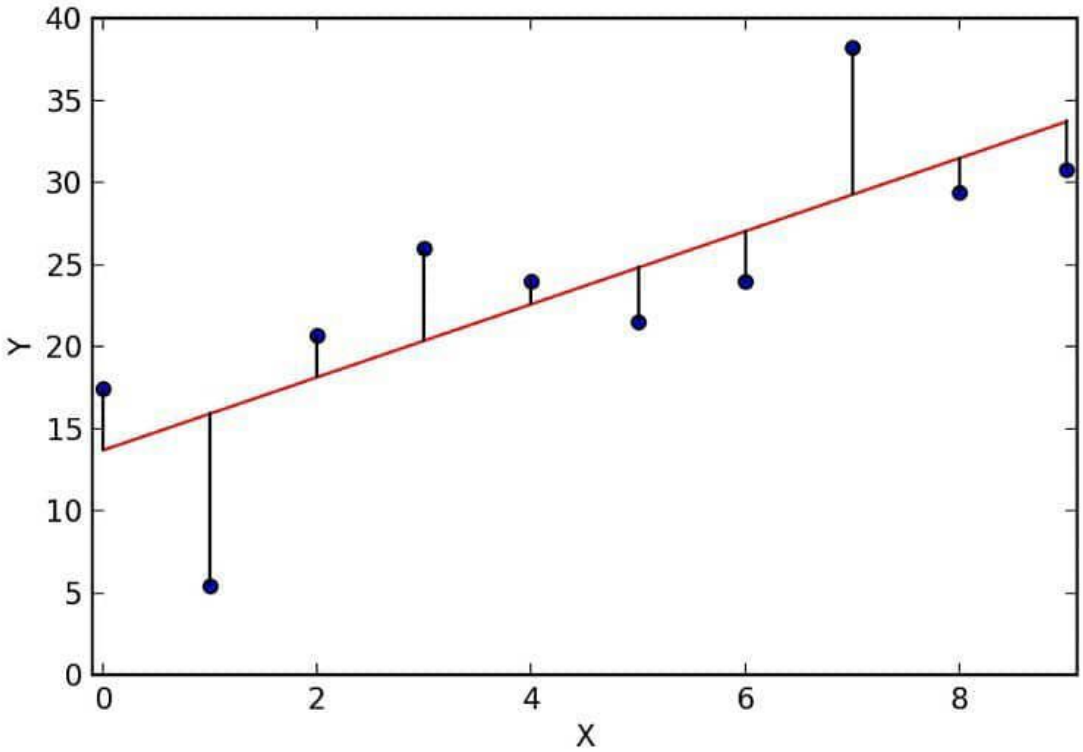
Simple Linear Regression

- Let's consider a very old example of house price prediction(Y) when given the square yards(X) as an input. We will use ordinary least squares to build the model to find W1 and b.
- The equation for ordinary least squares.

$$\text{Error} = \sum_{i=1}^{n} (actual\_output - predicted\_output) ** 2$$

- If we don't square the error, then the positive and negative point will cancel out each other.
- For calculating W1 and b we can use different methods and all will be shown in the code in future.
- And we can also use different metrics to evaluate the model like Regression sum of squares or Sum of Squared error or Total sum of squares.

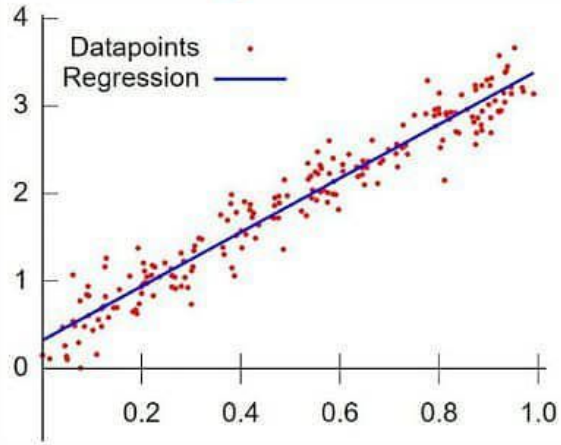
Simple Linear Regression



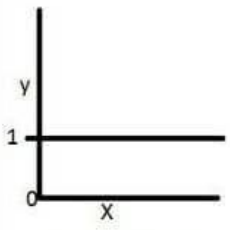
- This is how it looks like after training and finding the best fit line(Red line).
- As you can see the data points which are spread over and a line is drawn on that data points which is the best fit as it gets the lowest error metric value(Distance between the original point and the best fit line).

Simple Linear Regression

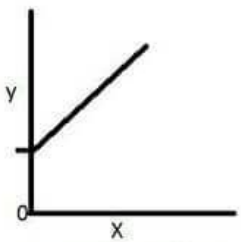
MATH behind linear regression



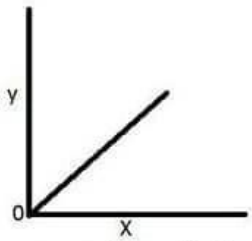
- Goal is to find that blue straight line (which is best fit) to the data. First lets talk about how to draw a linear line in the graph, In math we have an equation which is called linear equation
- y = mX+b { m->slope , b->Y-intercept }
- so we can draw the line if we take any values for m and b



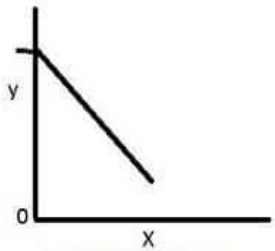
m=0 and b=1



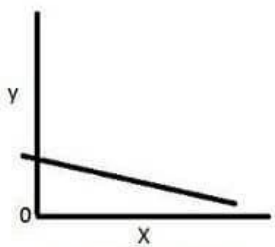
m=+value and b=1



m=+value and b=0



m=-value and b=5

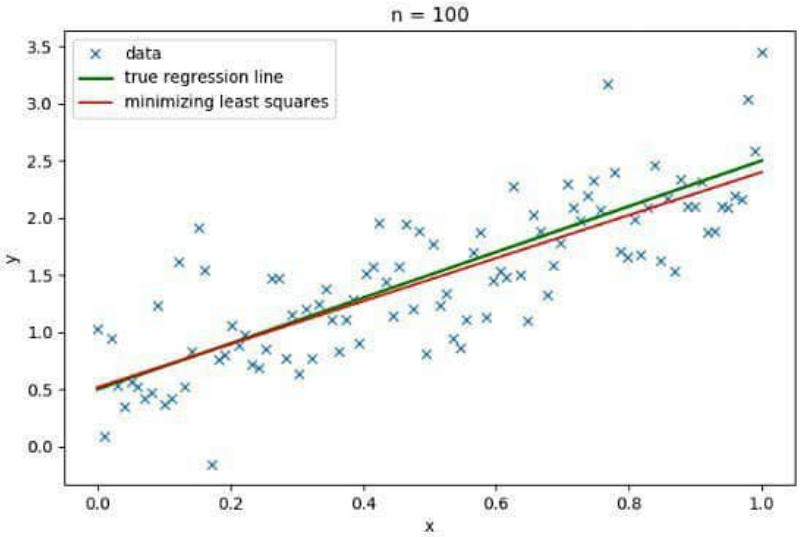


m=-value and b=1

Simple Linear Regression

MATH behind linear regression

- so how can we calculate m and b values ? and how do we know exact m and b values for the best fit line??



- so how we drew those lines, we take some random values of M and b and by taking all the X values we will find the Y values and we drew line with those Y values.
- We do this until we find the best fit line with low error value. But how many times you do this or for how many values you will check this.

Simple Linear Regression

MATH behind linear regression

- How do we change m and b values for the best fit line??
- Either we can use an algorithm called Gradient Descent Or we can use direct formulas from statistics. lets first use the statistics formulas then we can go to GD.

$$m = \frac{\sum_{i=1}^{n}(x_i-\overline{X})(y_i-\overline{Y})}{\sum_{i=1}^{n}(x_i-\overline{X})^2} \qquad b = \overline{Y} - m\overline{X}$$

- Here $\bar{x}$ is the mean of all the values in the input X and $\bar{y}$ is the mean of all the values in the desired output Y. This is the Least Squares method. Now we will implement this in python and make predictions.
- So how to do prediction for new values, we just use the formula Y = mX + b and use the m and b values which we got from the above formula. And we take new values as input as X and we calculate the Y value(Predicted value).

Simple Linear Regression

MATH behind linear regression using Gradient descent

- Ordinary Least Square method(previous method) looks simple and computation is easy. But, this OLS method will only work for a univariate dataset which is single independent variables and single dependent variables. Multi-variate dataset contains a single independent variables set and multiple dependent variables sets(This is called multiple linear regression), require us to use a machine learning algorithm called “Gradient Descent”. We can also use GD for simple linear regression.
- Gradient descent algorithm’s main objective is to minimise the cost function. It is one of the best optimisation algorithms to minimise errors (difference of actual value and predicted value). (we discussed more abput GD in our previous posts

Simple Linear Regression

MATH behind linear regression using Gradient descent

- Here we want to minimize the error function i.e sumof(Y - Y*^)*2 where Y is actual value and Y*^ is predicted value.
- We will use gradient descent to find the m and b values as old_m = new_m - partial_derivative_of_the_loss_function_with_respect_to_m(loss function).
- loss function = sumof(Y - (mX+b))**2
- We repeat the above formula multiple times until we dont see any change in the error value. We calculate b also using the same method.
- After stopping we get the m and b values and we do same thing as we done in previous method for drawing the best fit line and for predicting.

Simple Linear Regression

Preparing Data For Linear Regression

Try different preparations of your data using these heuristics and see what works best for your problem.

- **Linear Assumption:** Linear regression assumes that the relationship between your input and output is linear. It does not support anything else. This may be obvious, but it is good to remember when you have a lot of attributes. You may need to transform data to make the relationship linear (e.g. log transform for an exponential relationship).
- **Remove Noise:** Linear regression assumes that your input and output variables are not noisy. Consider using data cleaning operations that let you better expose and clarify the signal in your data. This is most important for the output variable and you want to remove outliers in the output variable (y) if possible.

Simple Linear Regression

Preparing Data For Linear Regression

Try different preparations of your data using these heuristics and see what works best for your problem.

- **Remove Collinearity:** Linear regression will over-fit your data when you have highly correlated input variables. Consider calculating pairwise correlations for your input data and removing the most correlated. (we use when we have multiple input features)
- **Gaussian Distributions:** Linear regression will make more reliable predictions if your input and output variables have a Gaussian distribution. You may get some benefit using transforms (e.g. log or BoxCox) on your variables to make their distribution more Gaussian looking.
- **Rescale Inputs:** Linear regression will often make more reliable predictions if you rescale input variables using standardization or normalization.

Simple Linear Regression

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.

Disadvantages of linear regression

- In real life, there aren't many problems in the world that exhibit a clear relationship between the independent and dependent variables.
- linear regression most of the time can be only used when we deal with relationships that graphically look like a line because "linear" means according to the mathematical graphical definition is a straight line. Outliers are other that make linear regression more limited in terms of its' use because linear regression always considers the case that tends to be the most frequent. least squares create more error because of outliers

Simple Linear Regression

Real world Cases

- Most important failure case is when we have multiple input features with collinearity. The model may over fit because of collinearity, to get rid of this we can use multiple techniques like removing collinearity or using PCA to reduce the dimensions and making collinearity features to independent features or using regularization techniques like L1 or L2(we will discuss more in next series).
- we can also use L1 or L2 regularizers to get important features when we have multiple features as input.
- Another case where it fails to fit best model is when it haves outliers. because of them we get the high error and squared error gives more penalty to them. We can get rid of this problem by removing outliers using any outlier detection technique.