

BEPEC SOLUTION PRESENTS

APPLIED MACHINE LEARNING

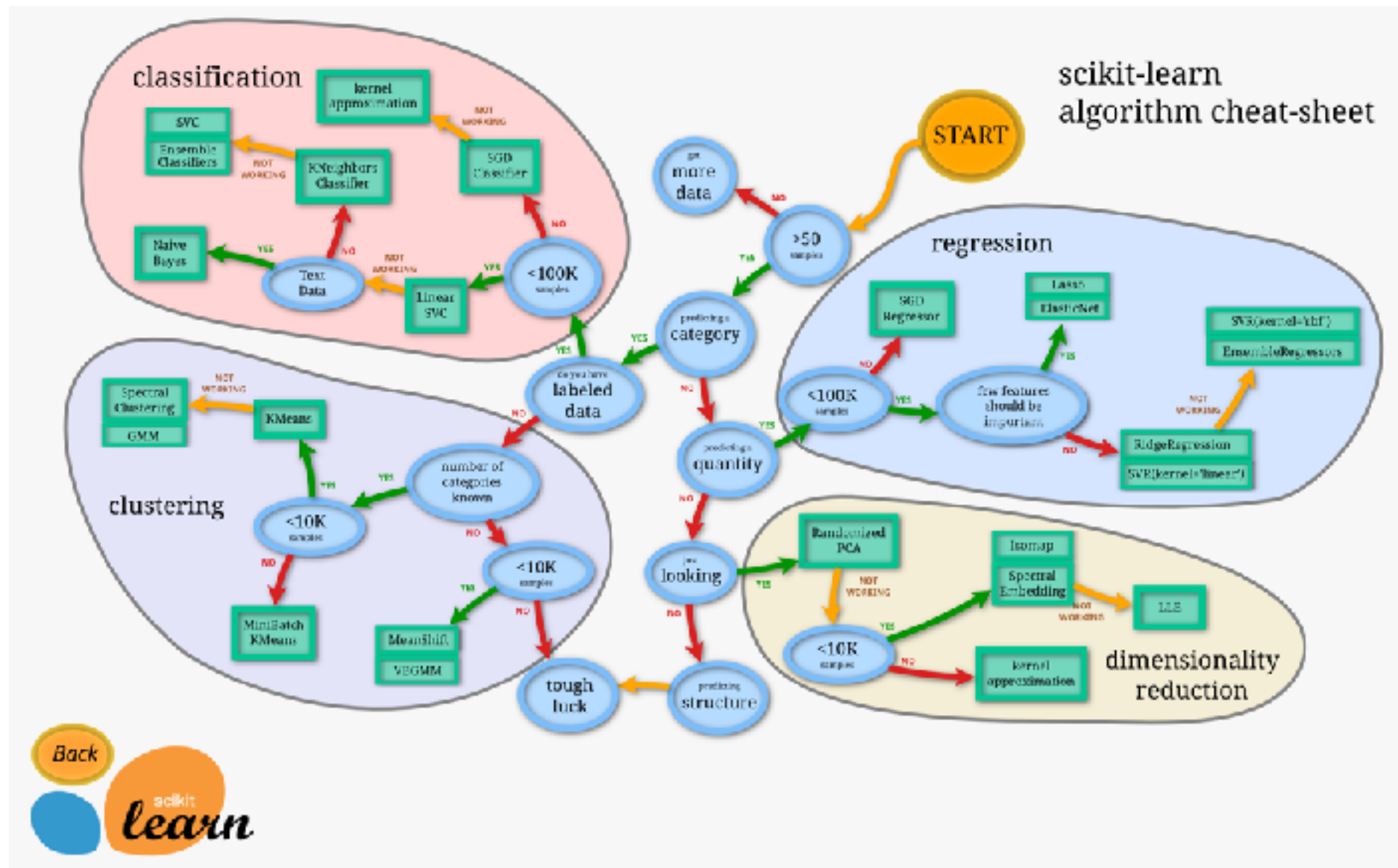
CHAPTER - 1 LINEAR REGRESSION
COMPLETE MATHS WITH POC AND THINGS
TO SUBMIT AFTER LEARNING

BY KANTH

1. Introduction to Regression vs Classification.
2. Difference between simple linear regression and multiple linear regression.
3. Mathematical understanding of Linear regression using a problem statement.
4. Assumption of linear regression.
5. Model evaluation techniques.
6. Model accuracy improving techniques.
7. Building a Linear regression model end - to - end

REGRESSION VS CLASSIFICATION VS CLUSTERING

Before starting with any of the machine learning problems or projects first we need to start with whether problem is regression based or classification based or clustering based problem. Have a look over this below chart for more detailed picture on classification or regression or clustering.



The labels(y) generally comes in categorical form and represents a finite number of classes. Consider the tasks bellow:

- Given set of input features predict whether Diabetes is Present or Absent.
- Given an image correctly unlocking a phone or not.
- From a given email predict whether it's spam email or not.

Types of classification

(1). **Binary classification**—when there is only two classes to predict, usually 1 or 0 values.

Multi-Class Classification—When there are more than two class labels to predict we call multi-classification task. E.g. predicting 3 types of iris species, image classification problems where there are more than thousands classes(cat, dog, fish, car,...).

Algorithms for classification

- Decision Trees
- Logistic Regression
- Naive Bayes
- K Nearest Neighbors
- Random Forest e.t.c.

Regression Problems

In regression problems we trying to predict continuous valued output, take this example. Given a size of the house predict the price(real value).

Regression Algorithms

- Linear Regression
- Regression Trees(e.g. Random Forest)
- Support Vector Regression (SVR)
- etc

Classification VS Regression

Classification: Discrete valued Y (e.g. A, B, C)

Regression: Continues Values Y (e.g. 12.6, 120.3, 342.4,...)

Whenever you find machine learning problem first define whether you are dealing with a classification or regression problem and you can get to know that analyzing the target variable (Y), note that here the input X can of any kind (continues or discrete) that doesn't count to define the problem. After defining the problem and getting to know the data it's much easier to chose or try out some algorithms.

Simple Linear Regression & Multiple Linear Regression

If there is only one independent variable then we call it as Simple Linear Regression. If there are more than one variable then we call it as Multiple linear regression.

$$y = f(x)$$

if $y = mx+c$ - Simple linear regression

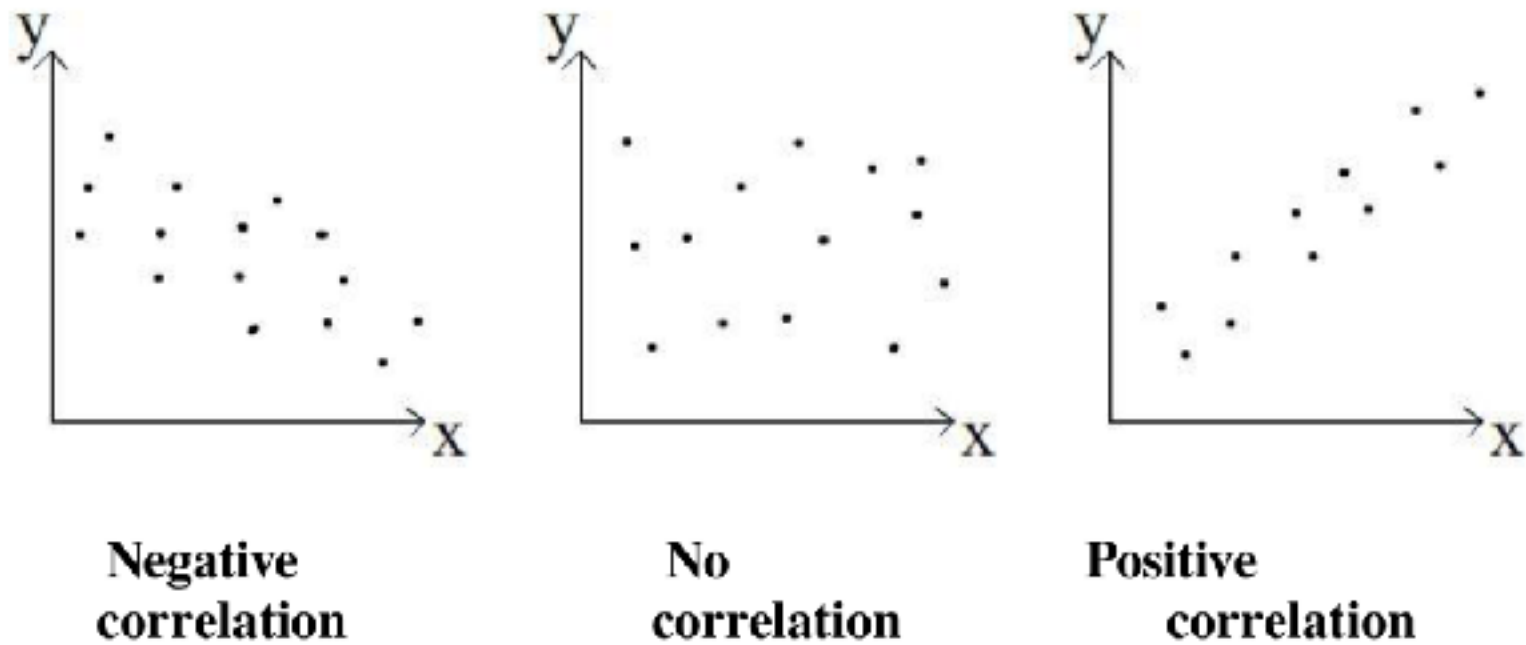
if $y = m_1x_1+m_2x_2+c$ - Multiple linear regression model.

Different types of correlation

We can categorise the type of correlation by considering as one variable increases what happens to the other variable:

- Positive correlation – the other variable has a tendency to also increase;
- Negative correlation – the other variable has a tendency to decrease;
- No correlation – the other variable does not tend to either increase or decrease.

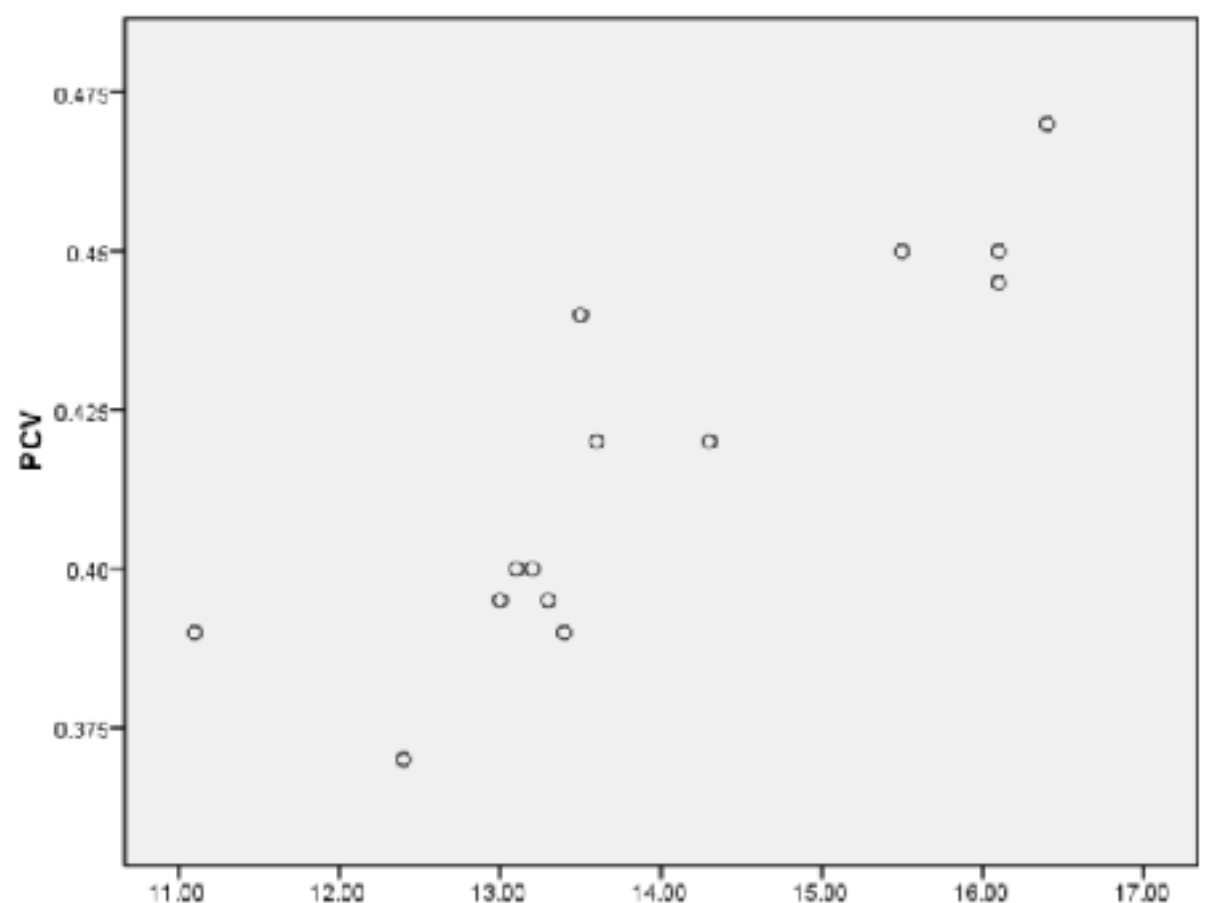
The starting point of any such analysis should thus be the construction and subsequent examination of a scatterplot. Examples of negative, no and positive correlation are as follows.



Example

Let us now consider a specific example. The following data concerns the blood haemoglobin (Hb) levels and packed cell volumes (PCV) of 14 female blood bank donors. It is of interest to know if there is a relationship between the two variables Hb and PCV when considered in the female population.

Hb	PCV
15.5	0.450
13.6	0.420
13.5	0.440
13.0	0.395
13.3	0.395
12.4	0.370
11.1	0.390
13.1	0.400
16.1	0.445
16.4	0.470
13.4	0.390
13.2	0.400
14.3	0.420
16.1	0.450



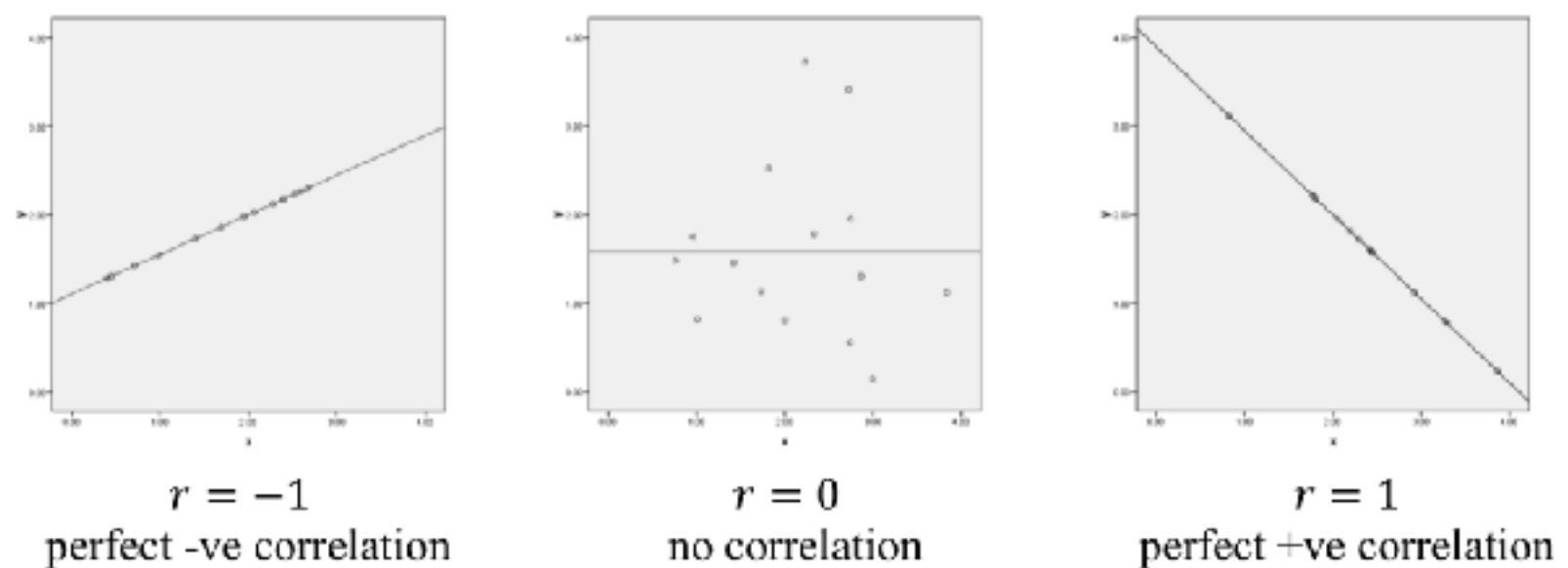
Correlation coefficient

Pearson's correlation coefficient is a statistical measure of the strength of a linear relationship between paired data. In a sample it is denoted by r and is by design constrained as follows

Furthermore:

- Positive values denote positive linear correlation;
- Negative values denote negative linear correlation;
- A value of 0 denotes no linear correlation;
- The closer the value is to 1 or -1 , the stronger the linear correlation.

In the figures various samples and their corresponding sample correlation coefficient values are presented. The first three represent the “extreme” correlation values of -1 , 0 and 1:

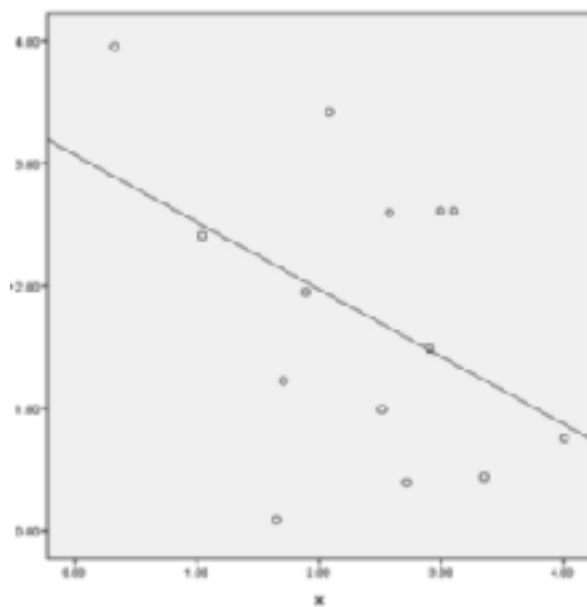


When we say we have perfect correlation with the points being in a perfect straight line.

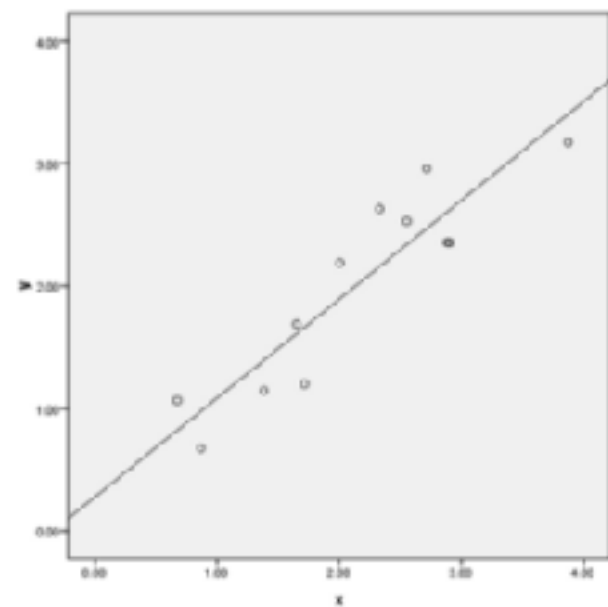
Invariably what we observe in a sample are values as follows:

Correlation is an effect size and so we can verbally describe the strength of the correlation using

- .00-.19. “very weak”
- .20-.39. “weak”



$r = -.45$
moderate -ve correlation



$r = .92$
very strong +ve correlation

•

40-.59. “moderate”

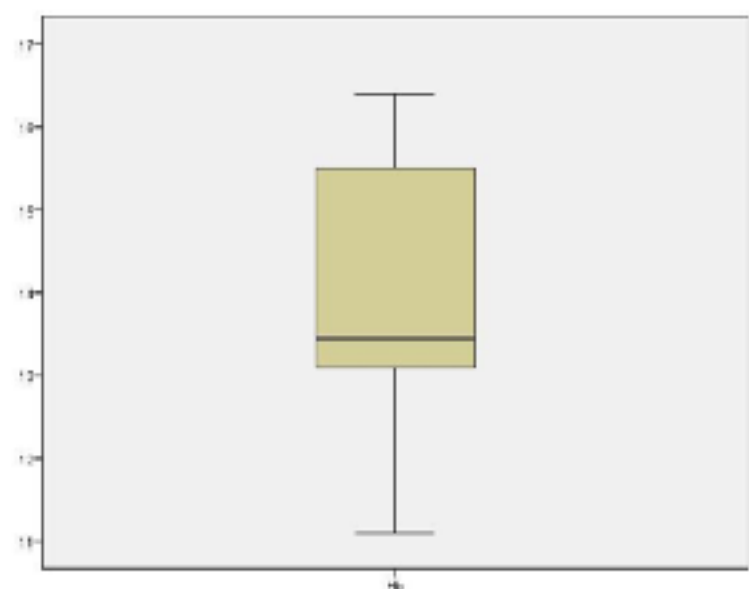
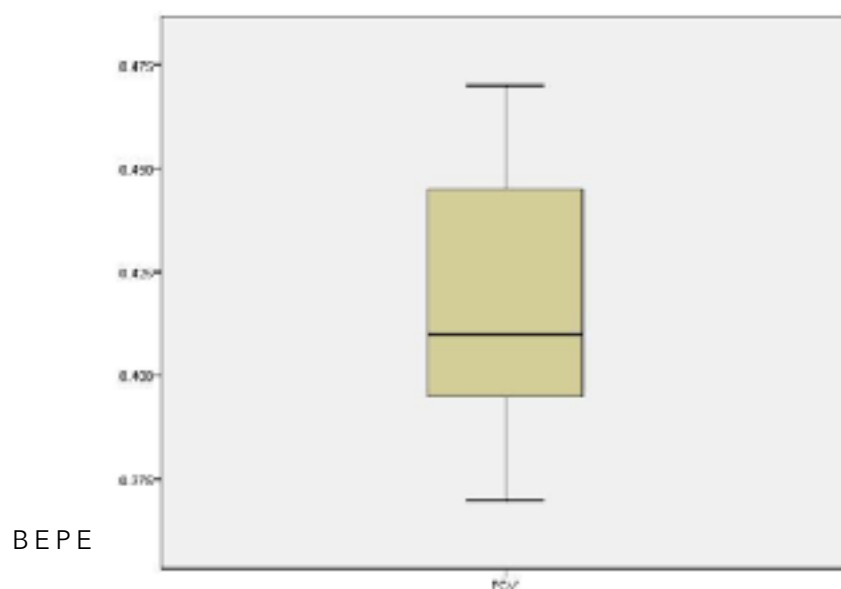
- .60-.79 “strong”
- .80-1.0. “very strong”

Assumptions:

The calculation of Pearson’s correlation coefficient and subsequent significance testing of it requires the following data assumptions to hold:

- interval or ratio level;
- linearly related;
- bivariate normally distributed.

We have no concerns over the first two data assumptions, but we need to check the normality of our variables. One simple way of doing is to examine boxplots of the data. These are given below.



The boxplot for PCV is fairly consistent with one from a normal distribution; the median is fairly close to the centre of the box and the whiskers are of approximate equal length.

The boxplot for Hb is slightly disturbing in that the median is close to the lower quartile which would be suggesting positive skewness. Although countering this is the argument that with positively skewed data the lower whisker should be shorter than the upper whisker; this is not the case here.

Since we have some doubts over normality, we shall examine the skewness coefficients to see if they suggest whether either of the variables is skewed.

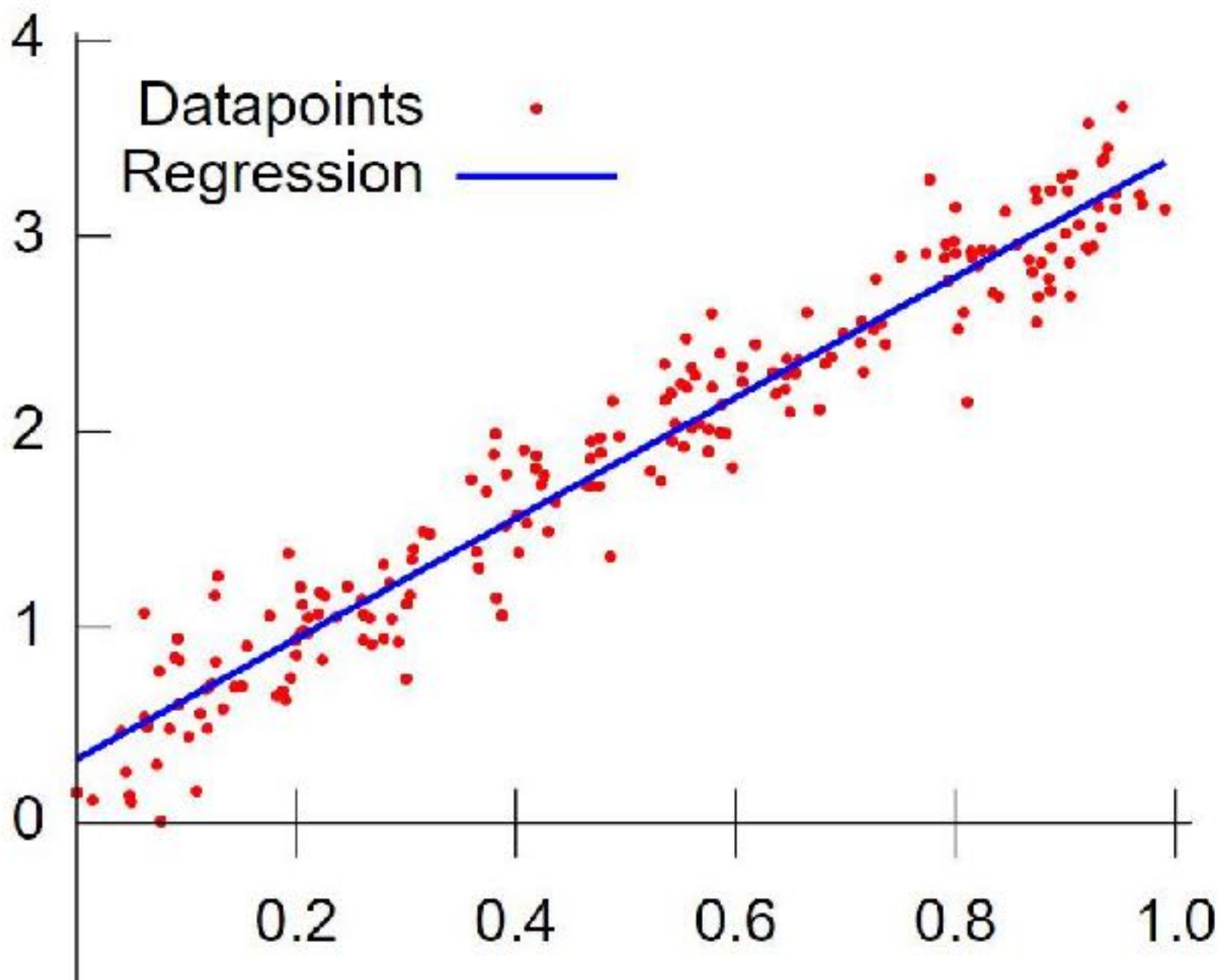
$$r = \frac{N\sum xy - (\sum x)(\sum y)}{\sqrt{[N\sum x^2 - (\sum x)^2][N\sum y^2 - (\sum y)^2]}}$$

Where:

N	=	number of pairs of scores
$\sum xy$	=	sum of the products of paired scores
$\sum x$	=	sum of x scores
$\sum y$	=	sum of y scores
$\sum x^2$	=	sum of squared x scores
$\sum y^2$	=	sum of squared y scores

Maths behind Linear Regression:

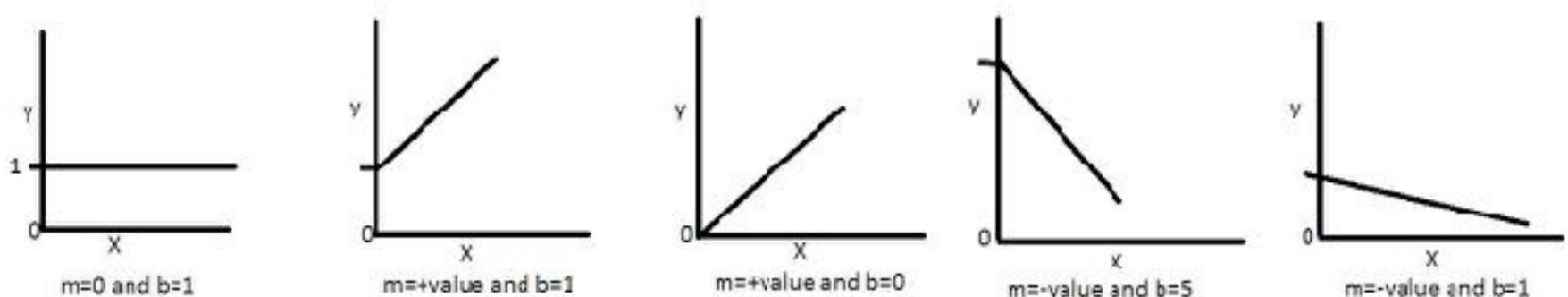
Linear Regression: it is a linear model that establishes the relationship between a dependent variable y (*Target*) and one or more independent variables denoted X (*Inputs*).



Goal is to find that blue straight line (which is best fit) to the data.

Our Training Data consists of X and y values so we can plot them on the graph, that's damn easy.
now *what's next?* how to find that blue line????

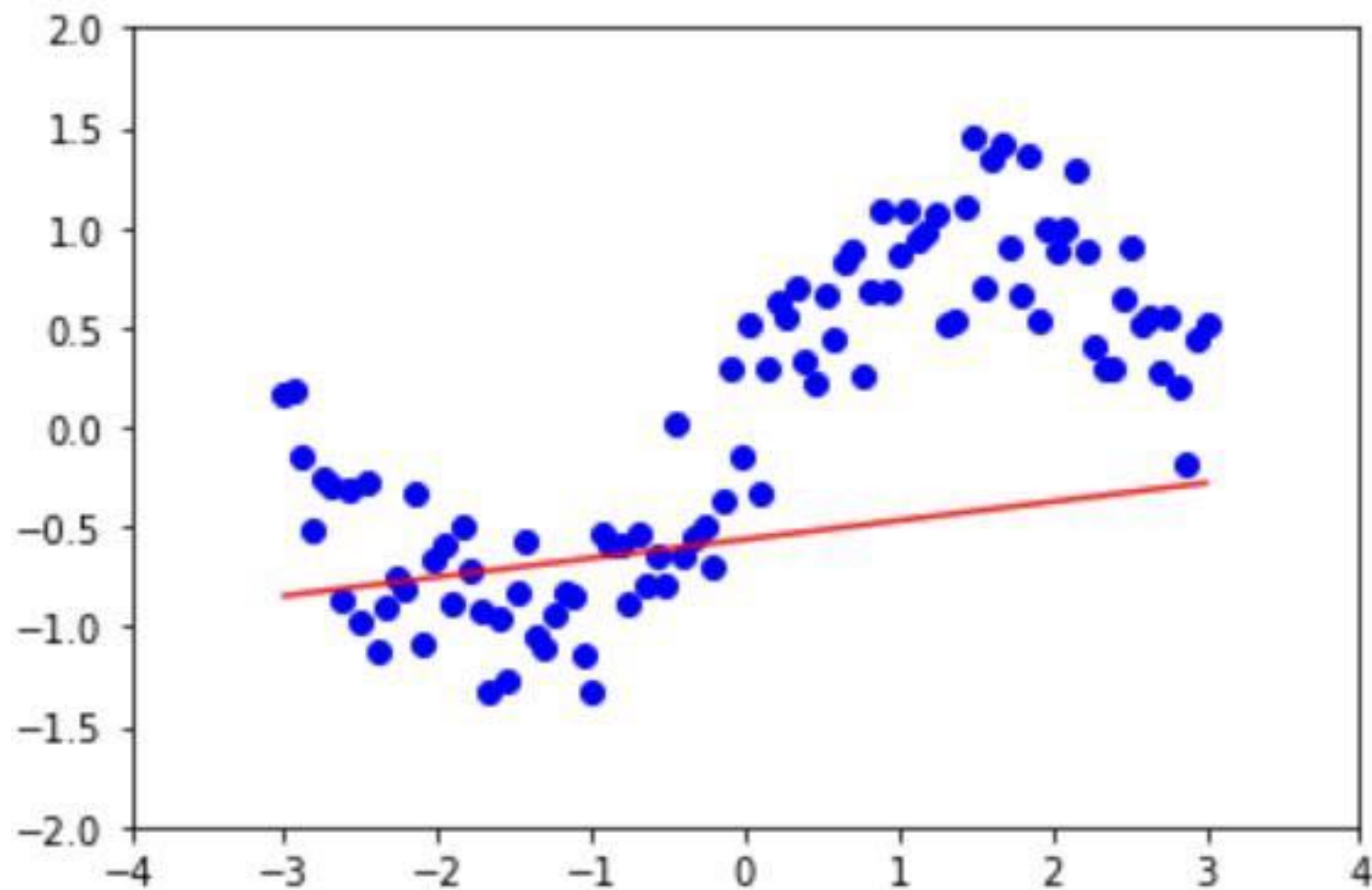
First lets talk about how to draw a linear line in the graph,
In math we have an equation which is called linear equation
 $y = mX + b$ { m->slope , b->Y-intercept }



so we can draw the line if we take any values for m and b

How do we get the m and b values ??? and how do we know exact m and b values for the best fit line??

Lets take a simple data set (sine wave form -3 to 3) and First time we take random values of m and b values and we draw a line something like this.



Random line for m and b

How we drew the above line?

we take the first X value(x_1) from our data set and calculate y value(y_1)

$y_1 = m \cdot x_1 + b$ {m,b->random values lets say 0.5,1

x_1 ->lets say -3 (first value from our data-set)

$$y_1 = (0.5 \cdot -3) + 1$$

$$y_1 = -0.5$$

by applying all x values for m and b values we get our first line.

Above picture has its own random variables (I hope you understand the concept)

That line is *not* fitting well to the data so we need to change m and b values to get the best fit line.

How to change this m and b values to get best fit line:

1. Gradient Descent
2. Ordinary least square

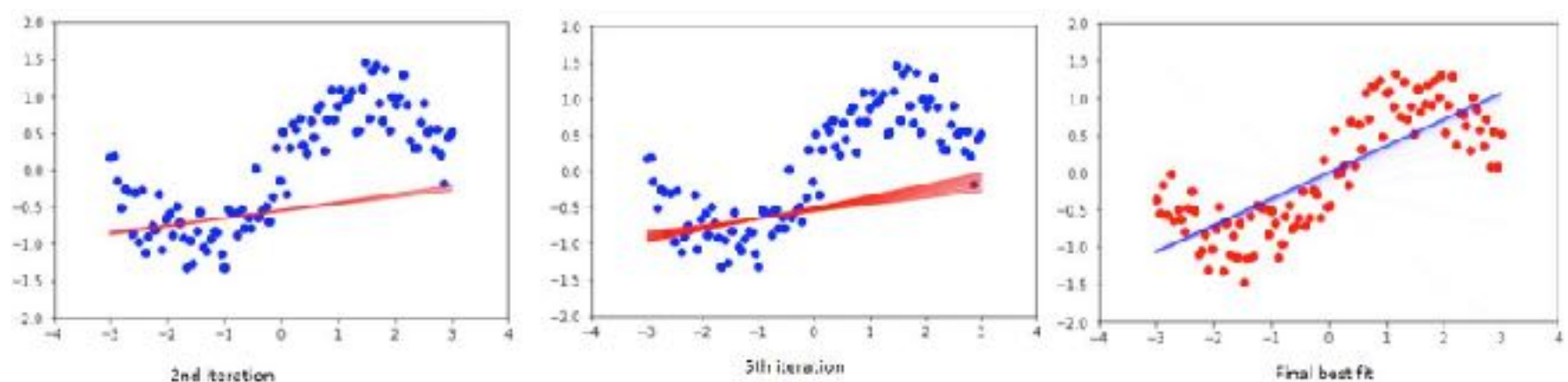
OLS:

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

$$b = \bar{Y} - m\bar{X}$$

\bar{X} is mean of X values , \bar{Y} mean of y values

Right now lets black box, we assume that we are getting the m and b values, Every time when the m and b values change we may get a different line and finally we get the best fit line



What's next??? Predicting new data, remember?? so we give new X values we get the predicted y values how does it work ??

same as above $y = mX + b$, we now know the final m and b values.

This is called **simple linear regression** as we have only one independent X value. Lets say we wanna predict housing price based the size of house

X= Size (in sqft's) y= Price (in dollar's)

X	y
1000	40
2000	70
500	25
.....	

What if we have more independent values of X????

Lets say we wanna predict housing price not only by the size of house but also by no of bedrooms

x1= Size (in sqft's), x2=N_rooms and y= Price (in dollar's)

x1	x2	y
----	----	---

1000	2	50
2000	4	90
500	1	35
.....		

The process same as above but the equation changes a bit

Note: Lets alias b and m as θ_0 and θ_1 (theta 0 and theta 1) respectively.

$y = \theta_0 + \theta_1 * X \rightarrow b + mX \rightarrow \text{Simple LR} \rightarrow \text{Single variable LR}$

$y = \theta_0 + \theta_1 * x_1 + \theta_2 * x_2 + \dots \theta_n * x_n \rightarrow \text{Multiple LR} \rightarrow \text{Multi variable LR}$

Gradient Descent:

First time we take random values for θ_0 and θ_1 , and we calculate y

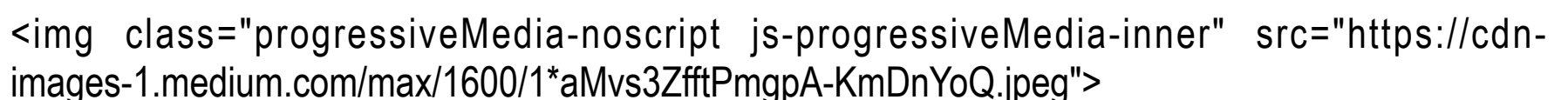
$y = \theta_0 + \theta_1 * X$

In machine learning we say hypothesis so **$h(X) = \theta_0 + \theta_1 * X$**

$h(X) = y$ but this y is not actual value in our data-set, this is predicted y from our hypothesis.

For example lets say our data-set is something like below and we take random values which are **1** and **0.5** for **θ_0** and **θ_1** respectively.

x	y		$h(x) = \theta_0 + \theta_1 * x$	predicted y
10	5		$1 + 0.5 * 10$	6
12	6.6			
3	1			
...	...	(Actual y value is 5 and predicted y value is 6)		



From this we calculate the error which is

$\text{error} = (h(x) - y)^2 \rightarrow (\text{Predicted} - \text{Actual})^2$

$\text{error} = (6 - 5)^2 = 1$

² is to get rid of negative values (what if Actual $y = 6$ and $P_y = 5$)

we just calculated the error for one data point in our data-set , we need to repeat this for all data points in our data set and sum up the all errors to one error which is called **Cost Function 'J(θ)'** in machine learning.

$$J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)})^2$$

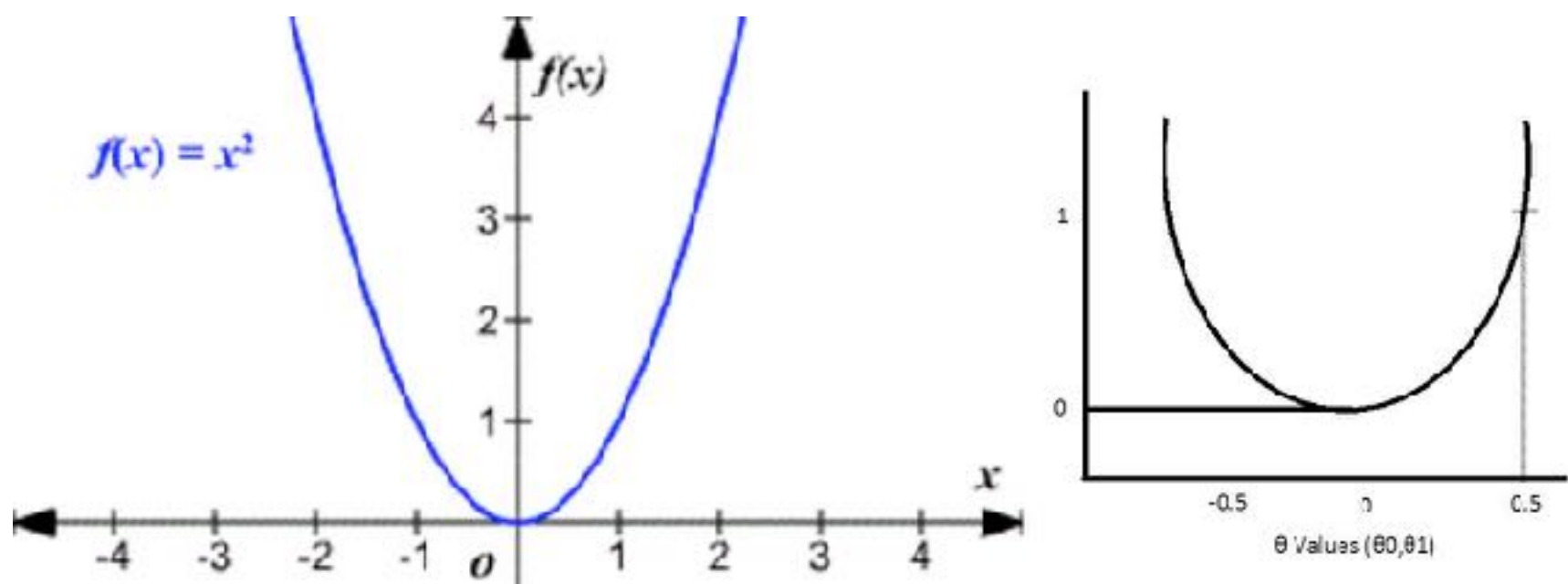
Cost Function.

- m The number of training examples
- $x^{(i)}$ The input vector for the i^{th} training example
- $y^{(i)}$ The class label for the i^{th} training example
- θ The chosen parameter values or “weights” ($\theta_0, \theta_1, \theta_2$)
- $h_{\theta}(x^{(i)})$ The algorithm’s prediction for the i^{th} training example using the parameters θ .

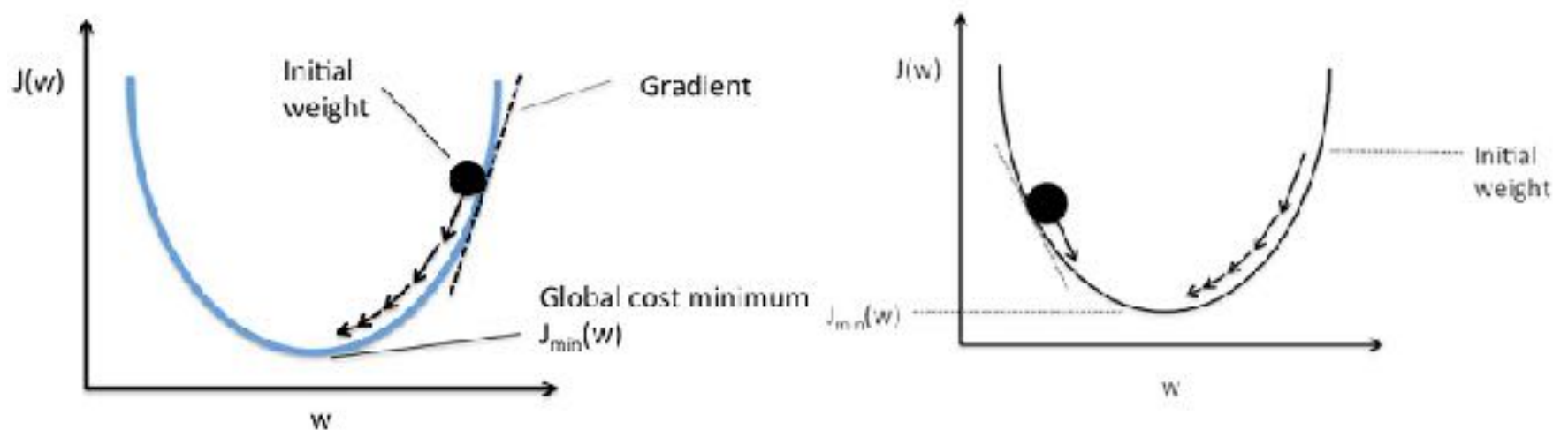
Our goal is to minimize the cost function (error) **we want our error close to zero Period.**

we have the error 1 for first data-point so lets treat that as whole error and reduce to zero for sake of understanding.

for $(h(x)-y)^2$ function we get always positive values and graph will look like this(Left) and lets plot the error graph.



Here is the gradient descent work comes into the picture.



+ θ values (Left), - θ values(Right)

By taking the little steps down to reach the minimum value (bottom of the curve) and changing the θ values in the process.

Assumptions of Linear Regression:

Assumption 1 : The Regression model is linear in its parameters (which are Coefficients and the error term).

Linearity: The change in the response variable due to one unit change in the predictor variable(X_k) is always constant irrespective of the current value of X_k

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k + \epsilon$$

The defining characteristic of linear regression is its functional form and to satisfy this assumption, the model should be correctly defined.

Assumption 2 : Independent variables should not be perfectly correlated with each other (No Multicollinearity)

Two variables are said to be correlated when one variable changes, the other variable also changes in fixed proportions. They are said to be perfectly correlated when they have *pearson correlation coefficient* between them as +1 or -1.

Perfect correlation between two variables suggest that they contain same information in them, in other words both the variables are different forms of same variable. Perfect correlation is a show stopper and regression cannot be applied in this case.

However, in many of the cases, variables are not perfectly correlated but have a strong correlation between them. This condition is known as multicollinearity.

Why is multicollinearity a problem?

The interpretation of a regression coefficient is that it represents the mean change in the response variable for 1 unit change in a predictor variable when you hold all of the other predictor variables

constant. But in case of multicollinearity, as the variables are strongly correlated, it is difficult to hold the other variables constant. So it is hard to estimate the parameters of the variables.

Checking for Multicollinearity:

Multicollinearity can be checked using *Variance Inflation factor(VIF)*. VIF is calculated using the below formula.

$$VIF(\hat{\beta}_j) = \frac{1}{1 - R^2_{X_j|X_{-j}}}$$

VIF > 5 is generally considered problematic and VIF > 10 suggests a definite presence of collinearity

Assumption 3 : Mean of the residuals should be Zero

Residuals refer to the difference between actual value and predicted value

Error term actually refers to the variance present in the response variable which the independent variables failed to explain. Our model is said to be unbiased if the mean of the error variable is zero. This assumption is by default taken care of in all the packages and libraries and we need not worry much about it.

Assumption 4 : Residuals should not be correlated with the independent variables

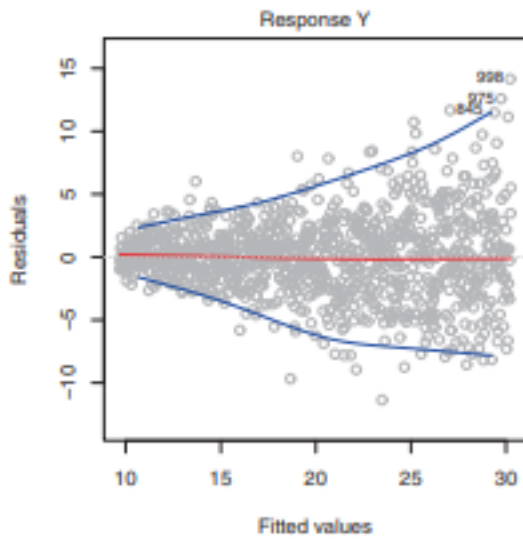
As mentioned above, error term represents the unexplained variance in the response variable. Now if residuals are correlated with the independent variable, we can use the independent variables to predict the error which is fundamentally wrong. This correlation between error terms and independent variables is known as *endogeneity*.

When this kind of correlation occurs our model might attribute the variance present in error to the independent variable, which in turn produces incorrect estimates.

Assumption 5 : Standard Deviation of the residuals should be constant (Homoscedasticity)

The variance of the errors should be consistent for all the observations. This condition is known as homoscedasticity. If the variance changes, we refer to that as heteroscedasticity.

The validity of this assumption can be checked using fitted vs residual plot. In this plot, if there is a cone shape, i.e. the spread of the residuals increases in one direction then heteroscedasticity is present.



Assumption 6 : Residuals should not be correlated with each other.

Residual of one observation should not predict the next observation. This problem is also known as auto correlation. In this scenario, the estimated standard errors underestimate the true standard errors

If residual can be predicted, that information should actually go into the model and not the error term. This problem can be solved by adding extra (Relevant) variables.

There are other assumptions too which are not very important, but are good to have.

Assumption 7 : Residuals should be normally distributed.

Assumption 8 : Independent variables should have positive variance.

Assumption 9: Number of observations should be more than the number of features.

Problem Statement: POC for OLX

The OLX marketplace is a platform for buying and selling services and goods such as electronics, fashion items, furniture, household goods, cars and bikes. In 2014, the platform had 11 billion page views, 200 million **monthly active users**, 25 million listings, and 8.5 million transactions per month. Even though they have good active users but OLX ended up with 8.5 million transactions. Now they want to grow the 8.5 million transactions to 16 Million transactions. To grow this transactions they need to identify best parameters which are impacting transactions of OLX. By tuning those parameters we need plan for 16 Million transactions in 2019.

1. Build a model end to end
2. Improve the model accuracy
3. Share the document on list of model improving techniques
4. Different Evaluation techniques in Linear Regression
5. Share the document on evaluation techniques

6. Submit the entire POC Script + Dataset + Documentation

