Linear Regression Assignment

- Aniket Verma

1. **Explain the linear regression algorithm in detail.**

Linear Regression is a Machine Learning Algorithm which is based on supervised learning. It works upon a given input and output data set, which act as a learning basis for future predictions. It is a regression algorithm, which tries to predict the change in dependent variables on any change in the one or more independent variables.

There are two kinds of Linear Regression models:

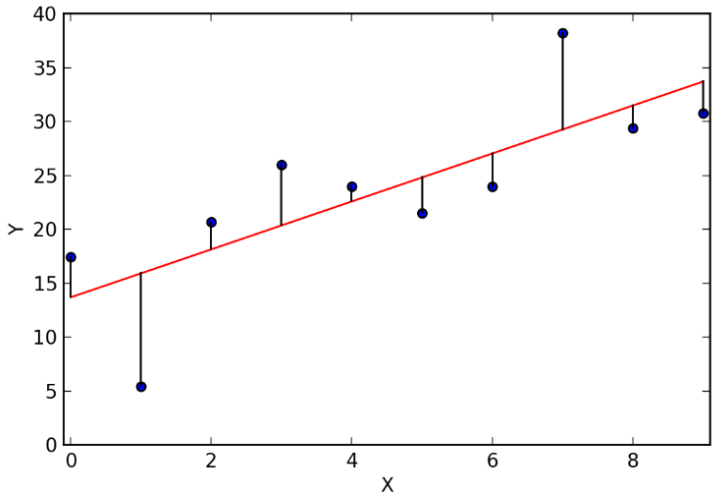
- Simple Linear Regression (1 independent variable):

$$y = b_0 + b_1 x$$

- Multiple Linear Regression (> 1 independent variable):

$$y = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3$$

Let us say that we have a simple data set where we have one independent variable x and the dependent variable y. The linear regression algorithm attempts to find a linear relationship between x and y and give us a regression line (best fit line). Any point on this best fit line is the predicted value for every independent datapoint x and using this line future predictions can be made. Since more than one best fit lines are possible, we usually follow the approach of reduced Mean Square Error.



MSE = mean ((predicted value <point on the line> - actual value <value of dependent value y>) ^ 2)

In the above figure, the red line is the “best possible” (with minimum MSE) and the blue points are the actual values.

2. **What are the assumptions of linear regression regarding residuals?**

The assumptions of linear regression residuals are as follows:

- Normality: The residuals are assumed to follow a normal distribution.
- Homoscedasticity: The residuals are assumed to have a constant variance.
- Independence: The residuals are assumed to be independent of each other and the magnitude of one residual in no way impacts that of the other.
- Zero Mean: The mean of residuals is assumed to be zero.
- No autocorrelation: The current value of residuals is assumed to not depend on any previous values.
- No correlation between independent variables and residuals.

3. **What is the coefficient of correlation and the coefficient of determination?**

The correlation coefficient tells us how closely data in a scatterplot fall along a straight line. If the absolute value of the correlation is equal to or nearly equal to one, it is said that the two variables in question have a high correlation. A correlation value of zero or near zero shows little or no correlation and a linear relation cannot be established. The sign of the correlation coefficient tells us whether the correlation is positive or negative.

The Coefficient of determination is the square of the coefficient of correlation. It explains the level of variance in the dependent variable

caused or explained by the independent variables in question. A higher value signifies that most of the variance in dependent variable is explained by the chosen independent variables.

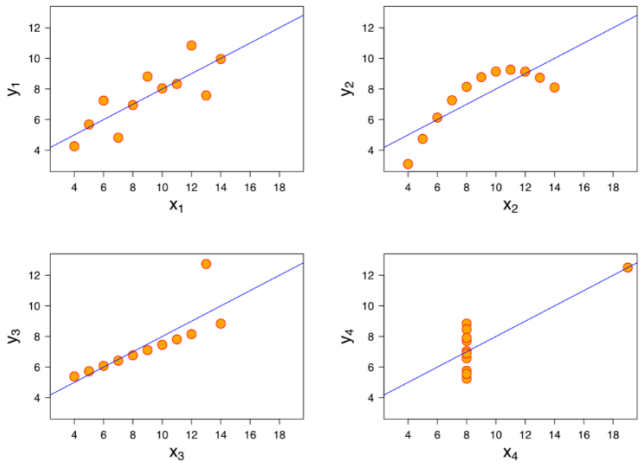
4. **Explain the Anscombe's quartet in detail.**

Anscombe's quartet comprises four data sets that identical descriptive statistics but exhibit entirely different distributions when plotted on graph. They were developed by statistician Francis Anscombe to demonstrate the importance of graphing data before analyzing it and the effect of outliers and other influential observations on statistical properties.

Following are the common descriptive statistics of the quartet.

Property	Value
Mean of X	9
Mean of Y	7.50
Sample variance of X	11
Sample variance of Y	4.125
Correlation between X and Y	0.816
Linear Regression Line	y = 3 + 0.5x
Coefficient of determination of linear regression	0.67

Following are the graph distributions of the quartet.



As you see from the above plots that though the descriptive statistics are same, the distributions are completely different. Anscombe came up with this concept to break the myth among statisticians that numerical calculations give more accurate results than graphical distributions. It laid emphasis on the importance of graphing data before analyzing it.

5. **What is Pearson's R?**

Pearson's R is also known as Pearson's correlation coefficient (or bivariate correlation). It denotes the linear correlation between two variables X and Y and the value lies between -1 and +1. Mathematically, it is defined as covariance of the two variables divided by the product of their standard deviations.

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_x \sigma_y}$$

Where,

cov = covariance,

σ_x = standard deviation of X

σ_y = standard deviation of Y

A higher absolute value signifies that the two variables are highly correlated. This means that the change in one variable "strongly impacts" a change in another variable. This change may be direct or inverse as per the coefficient's sign.

6. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?**

Scaling (or Feature Scaling) means adjusting the data with the intention to overcome any bias from outliers. It is usually performed in the data

preprocessing stage and is used to standardize the independent variables to a fixed range.

Feature scaling is a very essential component of data preprocessing for several reasons.

- It helps to eliminate any bias due to outliers in the data. The outliers in the data column may tend to interfere with the results and produce error prone predictions.
- Some variables are in dataset have higher range of values (say 10K-1000K) whereas other may have a relatively lower range (say 10-100). The relative impact on the outcome of the two variables on the dependent variable will be considerably different and may be a false inference that one may be more important than the other even if it is not the case.

The different types of scaling are as follows:

- Min-max normalization: It rescales the features in the range [0,1] or [-1,1]

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

- Mean normalization: It finds the normalized value of x.

$$x' = \frac{x - mean(x)}{\max(x) - \min(x)}$$

- Z-score normalization: It also finds the normalized value of x however we use standard deviation here. Hence, the process is also called standardization.

$$x' = \frac{x - \bar{x}}{\sigma}$$

- Unit length scaling: Each value is divided by the Euclidean length of the feature vector.

$$x' = \frac{x}{||x||}$$

7. **You might have observed that sometimes the value of VIF is infinite. Why does this happen?**

   The **variance inflation factor** quantifies the extent of correlation between one predictor and the other predictors in a model. In other words, VIF is a measure of how much an independent variable can be explained by others within the same data set. The more it can be explained, the less it makes sense to use in the model. It is used for diagnosing collinearity / multicollinearity. Higher values signify that it is difficult to assess accurately the contribution of predictors to a model.

$$\text{VIF} = \frac{1}{1 - \text{R}^2}$$

   Where,

   R = coefficient of determination

   Note the when VIF = inf when R = 1.
   This basically symbolizes that there is perfect collinearity which technically means that the variable is redundant. Hence, the variable is completely non distinguishable in terms of its impact on the outcome when compared to some other variable.

8. **What is the Gauss-Markov theorem?**

   The Gauss–Markov theorem states that in a linear regression model in which the errors are uncorrelated, have equal variances and expectation value of zero, the best linear unbiased estimator (BLUE) of the coefficients is given by the ordinary least squares (OLS) estimator, provided it exists. In this context, the definition of “best” refers to the minimum variance or the narrowest sampling distribution.

Let us try to understand the above statement. We know that the regression equation is:

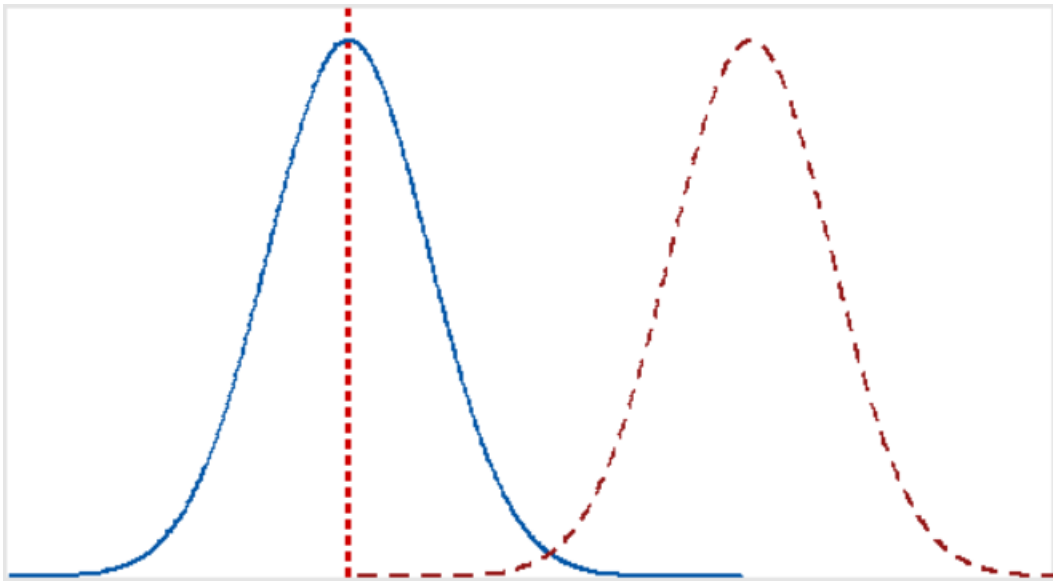
$$Y = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots + \hat{\beta}_k X_k + e$$

Where the beta coefficients are the indicators of the amount of impact of every independent variable on Y and epsilon is the residual error.

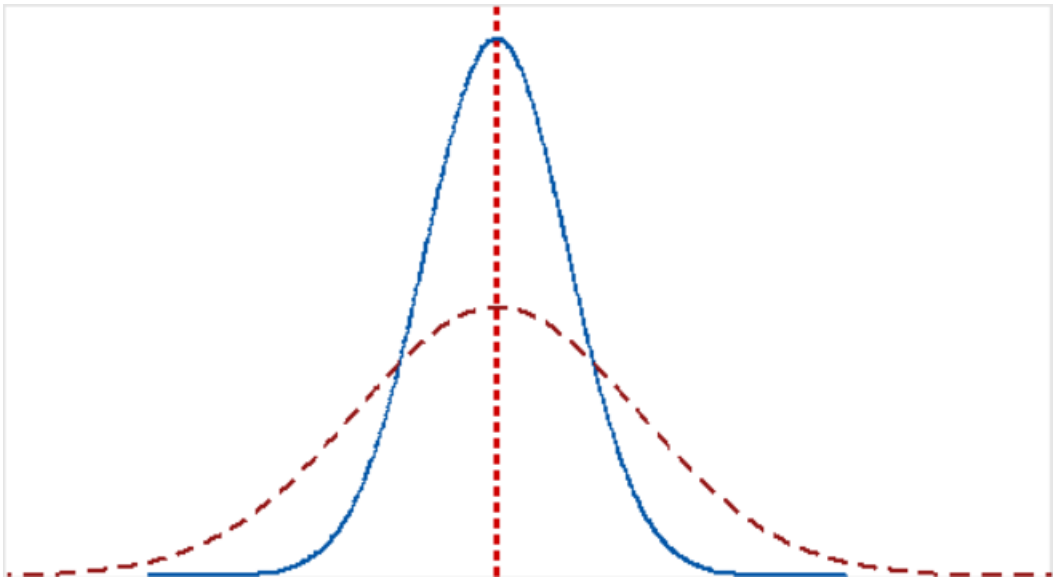
Let us say that we take many samples and estimate the values of the parameters in the regression equation. We end up getting a Sampling Distribution for each parameter. The Gauss-Markov theorem states that satisfying the OLS assumptions keeps the sampling distribution as tight as possible for unbiased estimates. What does this mean?

When all the OLS assumptions are satisfied, you obtain a sampling distribution curve which is very close to the curve which would have represented the actual values. In other words, you get Best Linear Unbiased Estimator (BLUE). Hence, the curve is centered at the true beta (actual value) and the spread of the curve is close to the actual curve (tight sampling).

Lets us make the concept clearer with examples.



(There is a high positive bias)



(The variance of the curve is much more than expected)

Such scenarios can be countered by adhering to OLS assumptions.

9. **Explain the gradient descent algorithm in detail.**

Gradient descent is a first order derivative optimization algorithm for finding the local minimum of a function. In mathematics we take the first order derivative to find the local minima or maxima. We equate this derivative to zero which basically means that at this point the gradient "will change". The sign of the second order derivative usually tells us if it's a maxima or minima.

In other words, the gradient decent algorithm is used to find the values of parameters of a function that minimizes a cost function. Doing the same using linear algebra would be extremely time consuming.

Steps for gradient descent:

- Select initial values of the coefficients.
- Calculate the cost of the coefficients by substituting the coefficient in the function.

$$\text{cost} = \text{function}(\text{coefficient})$$

- Calculate the derivative of the cost.

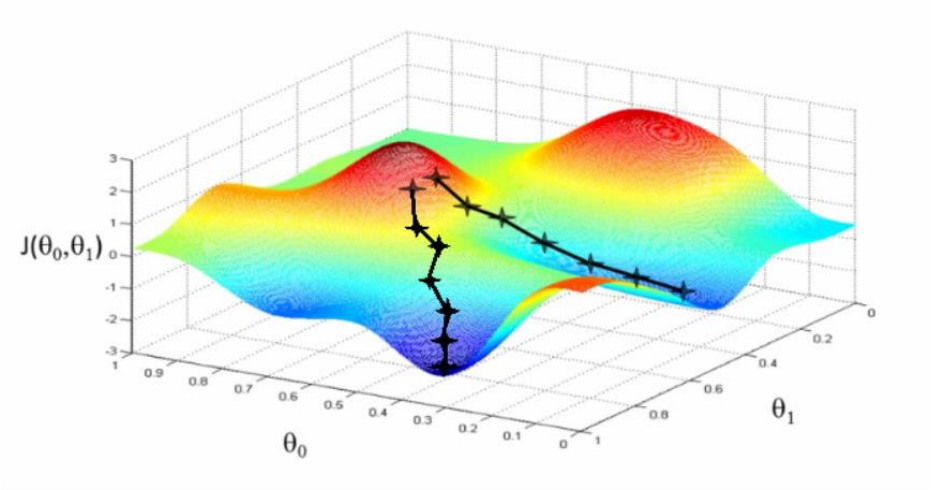
$$\text{delta} = \text{derivative}(\text{cost})$$

- Using the derivative, determine the direction in which to move to get closer to the local minima.
- Update the coefficient for the next iteration using the learning rate parameter (α).

coefficient = coefficient – (alpha * delta)

The different types of gradient descent algorithms are:

- Batch gradient descent.
- Stochastic gradient descent.



10. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it.

The Q-Q plot is created in order to determine whether the two sets of data come from the same distributions. Hence, the quantiles of the first dataset are plotted against those of the second dataset. A 45-degree reference line

is also plotted and if the points fall on this line, then the two sets of data are expected to belong to the same distribution.

Importance of Q-Q plots:

Q-Q plots are used to answer several questions like, comparing the two datasets in terms of location, scale, origin, behavior, shape etc. Answers to these questions can be extremely useful while doing linear regression.

Q-Q plots are also used to estimate the quantiles within the same data set. We plot the quantile values against standardized quantile values. The goal is to minimize the residuals which helps us to get a better line of regression. The closer the quantile points lie on a straight line the better. Q-Q plots can be used to determine the data distribution (is it normal, uniform etc.) which is helpful in linear regression.