GENERAL SUBJECTIVE QUESTIONS

by:
Mohammad Jawad Amina

1. EXPLAIN THE LINEAR REGRESSION ALGORITHM IN DETAIL

Linear Regression Algorithm is a machine learning algorithm based on supervised learning. Linear regression is a part of regression analysis. Regression analysis is a technique of predictive modelling that helps you to find out the relationship between Input and the target variable.

1. Here are the types of regressions:
    - Linear Regression
    - Multiple Linear Regression
    - Logistic Regression
    - Polynomial Regression

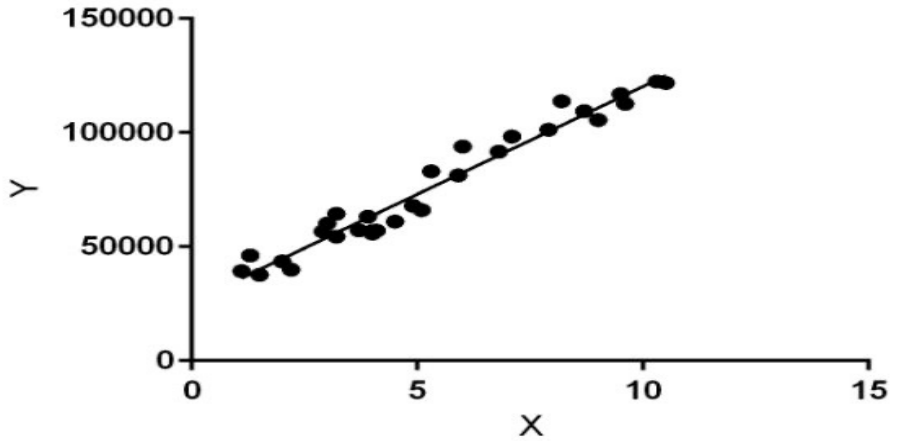
2. There are 3 Machine Learning techniques/algorithms:
    - Regression:
      The output variable to be predicted is a "Continuous Variable". (Ex: Scores of a Student)
    - Classification
      The output variable to be predicted is a "Categorical Variable". (Ex: Classifying incoming E-mails as SPAM or Ham)
    - Clustering
      No pre-defined notion of a label is allocated to groups/clusters formed. (Ex: Customer Segmentation).

3. Classifying Machine Learning models into two broad categories:
    - Supervised Learning Methods:
      Past data with labels is used for building the model.
      Regression and Classification algorithms fall under this categories.
    - Unsupervised Learning Methods:
      No pre-defined labels are assigned to past data.
      Clustering algorithms fall under this category.

4. Past data set is divided into two parts during "**Supervised Learning**" method:
    - Training data: Is used for model to learn during modelling.
    - Testing data: Is used by the trained model for prediction and model evaluation.

5. Linear regression models can be classified into two types depending upon the number of independent variables:
    - Simple Linear Regression: When the number of independent variable is 1.
    - Multiple Linear Regression: When the number of independent variable is more than 1.

6. The equation of the best fit regression line **$Y=\beta 0+\beta 1.x$** ($\beta 0$ – Intercept, $\beta 1$ – Slope), the *independent variable* is also known as the "**predictor variable**" and the *dependent variables* are also known as the "**output variable**".



7. The equation of the best fit regression line **$Y=\beta 0+\beta 1.x$** can be found by minimising the cost function (RSS in this case, using the ordinary least squares method), which is done using the following two methods:

    Differentiation & Gradient Descent Method

8. The strength of a Linear Regression model is mainly explained by R-Squared, where

    R-Squared = 1 – (RSS/TSS)

RSS: Residual Sum of Squares
TSS: Total Sum of Squares

2. EXPLAIN THE ANSCOMBE'S QUARTET IN DETAIL

Anscombe's Quartet can be defined as a group of four data sets which are **nearly identical in simple descriptive statistics**, but there are some peculiarities in the dataset that **fools the regression model** if built. They have very different distributions and **appear differently** when plotted on scatter plots.

It was constructed in 1973 by statistician **Francis Anscombe** to illustrate the **importance** of **plotting the graphs** before analysing and model building, and the effect of other **observations on statistical properties**.

There are these four data set plots which have nearly **same statistical observations**, which provides same statistical information that involves **variance**, and **mean** of all (x,y) points in all four datasets.

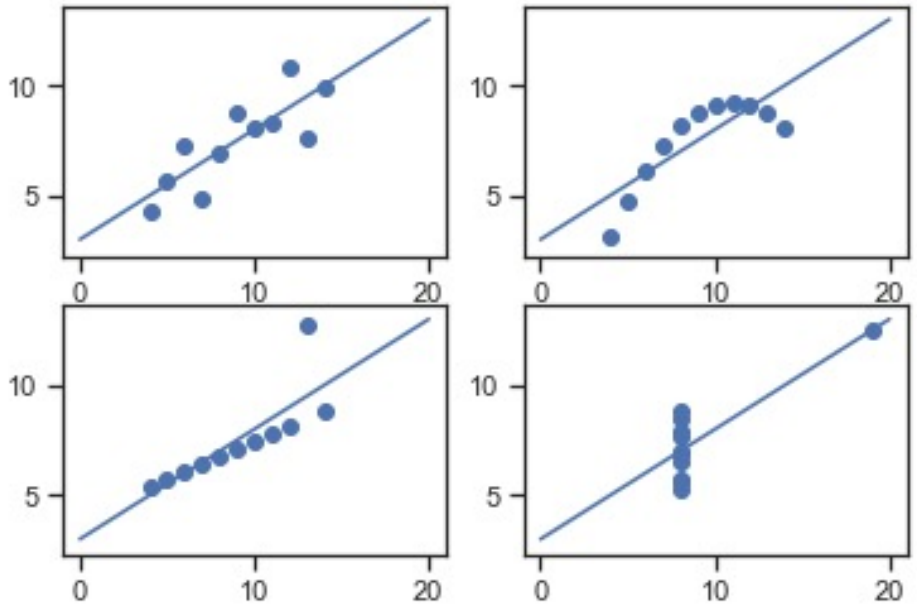
The basic thing to analyse about these data-sets is that they all share the same descriptive statistics(mean, variance, standard deviation etc) but different graphical representation. Each graph plot shows the different behaviour irrespective of statistical analysis.

x1	y1	x2	y2	x3	y3	x4	y4
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

Apply the statistical formula on the above data-set:

Average Value of x = 9
Average Value of y = 7.50
Variance of x = 11
Variance of y =4.12
Correlation Coefficient = 0.816
Linear Regression Equation : y = 0.5 x + 3

However, the statistical analysis of these four data-sets are pretty much similar. But when we plot these four data-sets across the x & y coordinate plane, we get the following results & each pictorial view represent the different behaviour.



<u>The four datasets can be described as</u>:

Dataset 1: this fits the linear regression model pretty well.

Dataset 2: this could not fit linear regression model on the data quite well as the data is non-linear.

Dataset 3: shows the outliers involved in the dataset which cannot be handled by linear regression model.

Dataset 4: shows the outliers involved in the dataset which cannot be handled by linear regression model.

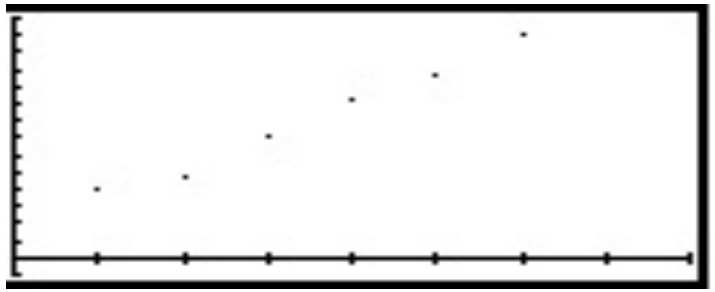
3. WHAT IS PEARSON'S R?

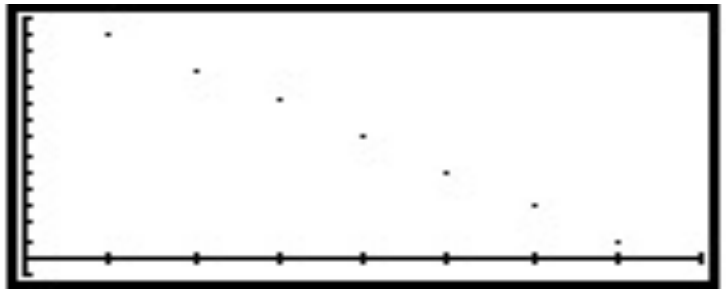
The **Pearson correlation coefficient** (**PCC**) — also known as **Pearson's *R***, the **Pearson product-moment correlation coefficient** (**PPMCC**), the **bivariate correlation**, or colloquially simply as **the correlation coefficient.**

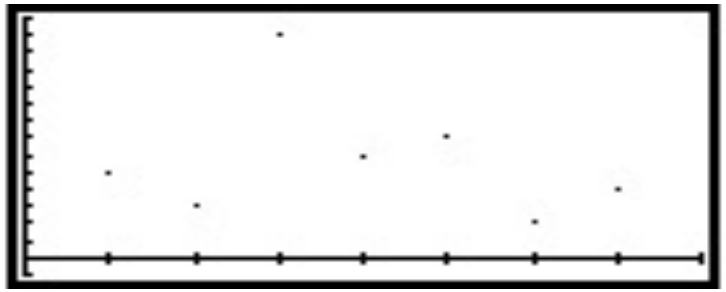
The Pearson's correlation coefficient varies between -1 and +1 where:

- r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)
- r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)
- r = 0 means there is no linear association
- r > 0 < 5 means there is a weak association
- r > 5 < 8 means there is a moderate association
- r > 8 means there is a strong association

The figure below shows some data sets and their correlation coefficients. The first data set has an r=0.996, the second has an r = -0.999 and the third has an r= -0.233.







Formula

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Here,

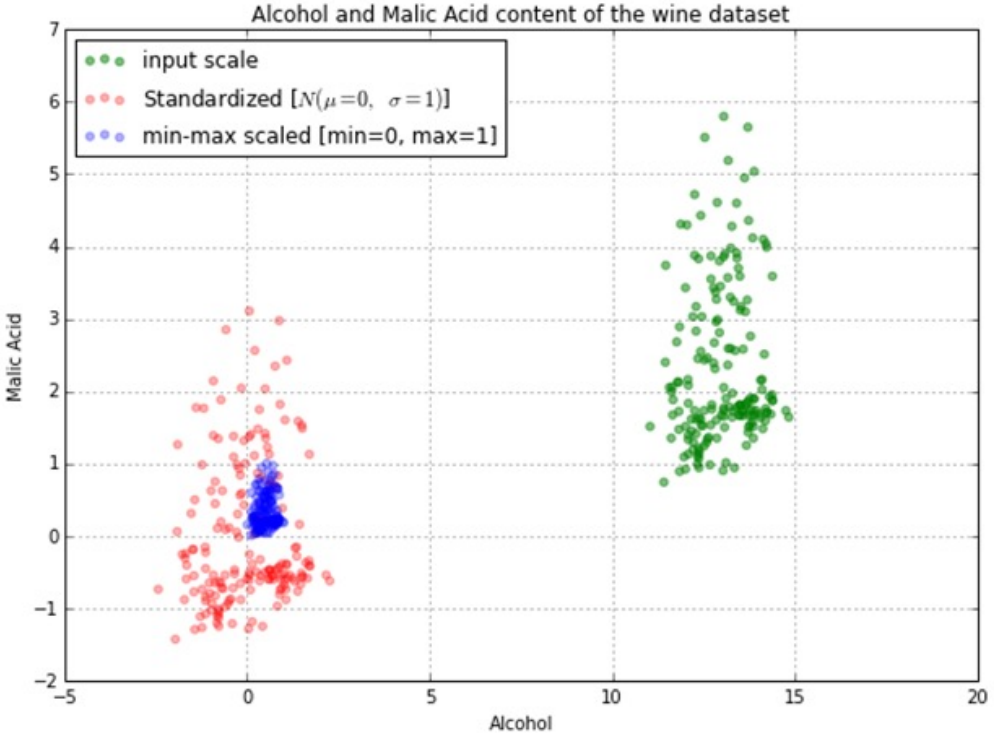
- r =correlation coefficient
- x_i =values of the x-variable in a sample
- $\bar{x}$ =mean of the values of the x-variable
- y_i =values of the y-variable in a sample
- $\bar{y}$ =mean of the values of the y-variable

4. WHAT IS SCALING?

- It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

- Feature scaling is a method used to normalize the range of independent variables or features of data. In data processing, it is also known as data normalization and is generally performed during the data pre-processing step. Just to give you an example — if you have multiple independent variables like age, salary, and height; With their range as (18–100 Years), (25,000–75,000 Euros), and (1–2 Meters) respectively, feature scaling would help them all to be in the same range, for example- centred around 0 or in the range (0,1) depending on the scaling technique.

- In order to visualize the above, let us take an example of the independent variables of alcohol and Malic Acid content in the wine dataset from the "Wine Dataset" that is deposited on the UCI machine learning repository. Below you can see the impact of the two most common scaling techniques (Normalization and Standardization) on the dataset.



WHY IS SCALING PERFORMED?

Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.
It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

WHAT IS THE DIFFERENCE BETWEEN NORMALIZED SCALING AND STANDARDIZED SCALING?

Normalization/Min-Max Scaling:

It brings all of the data in the range of 0 and 1.
sklearn.preprocessing.MinMaxScaler *helps to implement normalization in python.*

$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

Standardization Scaling:

Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).

$$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$$

sklearn.preprocessing.scale helps to implement standardization in python.
One disadvantage of normalization over standardization is that it **loses** some information in the data, especially about **outliers**.

One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers.

MinMax scaling - Method used to make sure that **data** is internally consistent.
Standardisation - Method used to make sure that **data** is internally consistent.

5. YOU MIGHT HAVE OBSERVED THAT SOMETIMES THE VALUE OF VIF IS INFINITE. WHY DOES THIS HAPPEN?

- VIF is an index that provides a measure of how much the variance of an estimated regression coefficient increases due to collinearity. In order to determine VIF, we fit a regression model between the independent variables.

- If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

- If all the independent variables are orthogonal to each other, then VIF = 1.0. If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables.

- If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity. This would mean that that standard error of this coefficient is inflated by a factor of 2 (square root of variance is the standard deviation).

- The standard error of the coefficient determines the confidence interval of the model coefficients. If the standard error is large, then the confidence intervals may be large, and the model coefficient may come out to be non-significant due to the presence of multicollinearity.

VIF	Conclusion
1	No multicollinearity
4 - 5	Moderate
10 or greater	Severe

6. WHAT IS A Q-Q PLOT?

Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.

EXPLAIN THE USE AND IMPORTANCE OF A Q-Q PLOT IN LINEAR-REGRESSION.

This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.

***Few advantages**:*

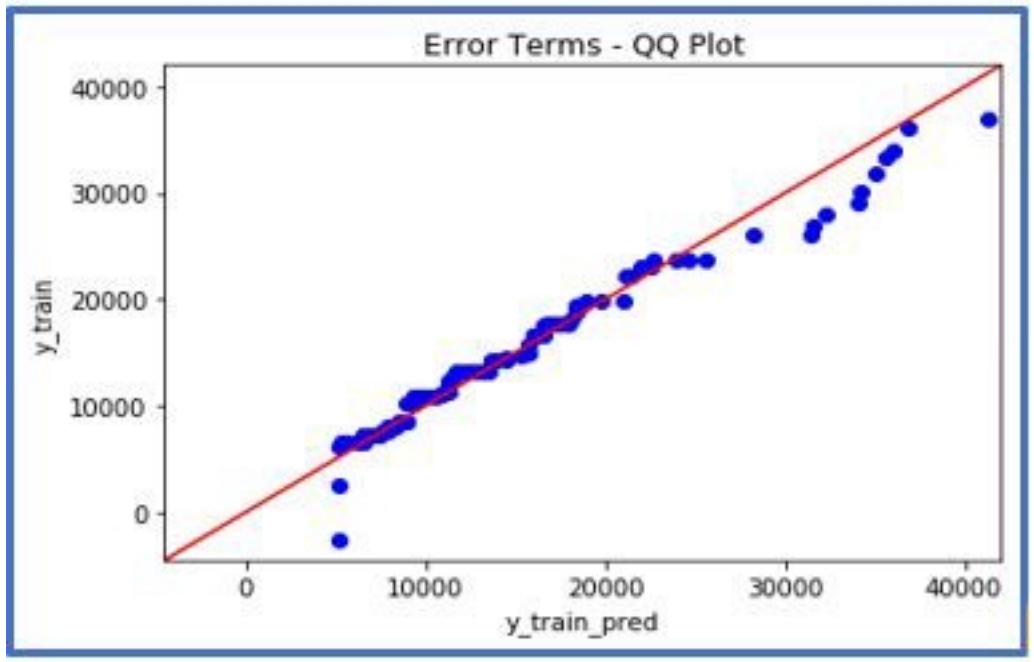
a) It can be used with sample sizes also
b) Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.

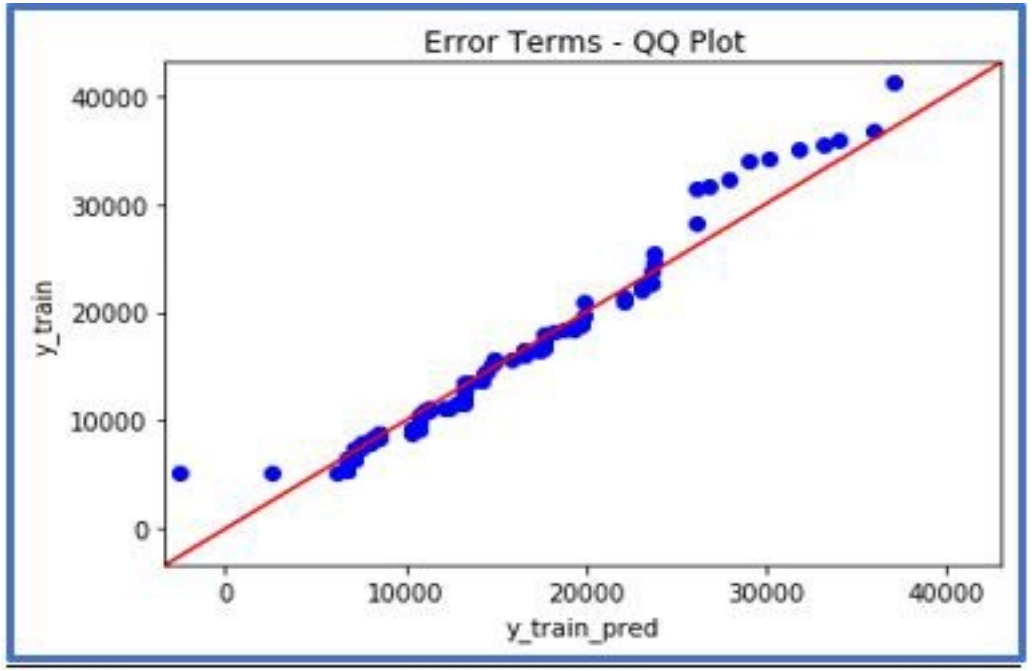
It is used to check following scenarios:
If two data sets —
i. Come from populations with a common distribution
ii. have common location and scale
iii. have similar distributional shapes
iv. have similar tail behaviour

Interpretation:

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.

Below are the possible interpretations for two data sets.
a) Similar distribution: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x -axis
b) Y-values < X-values: If y-quantiles are lower than the x-quantiles.





c) X-values < Y-values: If x-quantiles are lower than the y-quantiles.

d) Different distribution: If all point of quantiles lies away from the straight line at an angle of 45 degree from x -axis


Thank you