Linear Regression Assignment

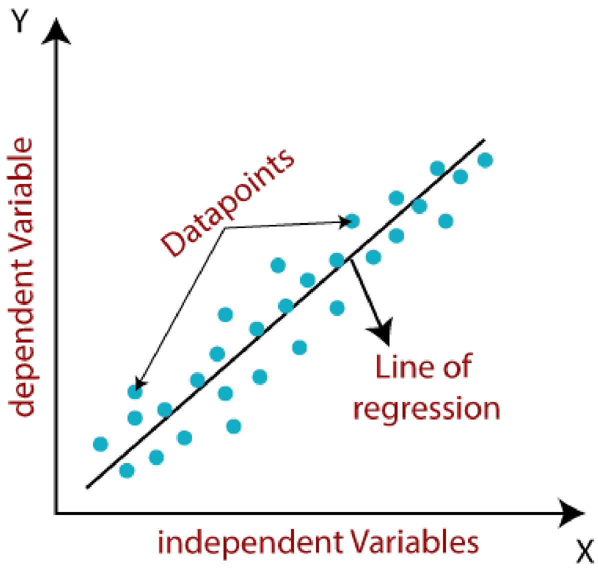
General Subjective Questions & Answers

MOHAMMED RAHAMATHULLA
iiitB and UpGrad Executive PG Programme
Machine Learning & Artificial Intelligence
mohammedrahmat@gmail.com

Q1. Explain the linear regression algorithm in detail.

Linear regression is one of the easiest and most popular Machine Learning algorithms. It is a statistical method that is used for predictive analysis. Linear regression makes predictions for continuous/real or numeric variables such as **sales, salary, age, product price,** etc.
Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables, hence called as linear regression. Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable

The linear regression model provides a sloped straight line representing the relationship between the variables. Consider the below image:



Mathematically, we can represent a linear regression as:

```
   1.   y= a0+a1x+ ε
```

Here,

Y= Dependent Variable (Target Variable)
X= Independent Variable (predictor Variable)
a0= intercept of the line (Gives an additional degree of freedom)
a1 = Linear regression coefficient (scale factor to each input value).
ε = random error

The values for x and y variables are training datasets for Linear Regression model representation.

Steps to implement Linear regression model

1. Initialize the parameters.
2. Predict the value of a dependent variable by given an independent variable.
3. Calculate the error in prediction for all data points.
4. Calculate partial derivative w.r.t a0 and a1.
5. Calculate the cost for each number and add them.
6. Implement use case of Linear regression with python code.

Linear regression can be further divided into two types of the algorithm:

- **Simple Linear Regression:**
  If a single independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Simple Linear Regression.

- **Multiple Linear regression:**
  If more than one independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Multiple Linear Regression

Summary

In Regression, we plot a graph between the variables which best fit the given data points. Linear regression shows the linear relationship between the independent variable (X-axis) and the dependent variable (Y-axis).To calculate best-fit line linear regression uses a traditional slope-intercept form. A regression line can be a Positive Linear Relationship or a Negative Linear Relationship.

The goal of the linear regression algorithm is to get the best values for a0 and a1 to find the best fit line and the best fit line should have the least error. In Linear Regression, **Mean Squared Error (MSE)** cost function is used, which helps to figure out the best possible values for a0 and a1, which provides the best fit line for the data points. Using the MSE function, we will change the values of a0 and a1 such that the MSE value settles at the minima. Gradient descent is a method of updating a0 and a1 to minimize the cost function (MSE).

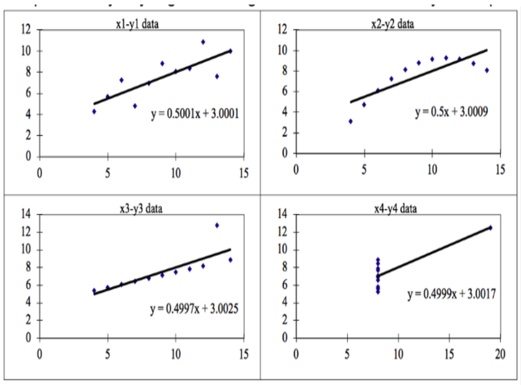
Q2. Explain the Anscombe’s quartet in detail.

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.

This tells us about the importance of visualising the data before applying various algorithms out there to build models out of them which suggests that the data features must be plotted in order to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data, etc. Also, the Linear Regression can be only be considered a fit for the data with linear relationships and is incapable of handling any other kind of datasets. These four plots can be defined as follows:

When these models are plotted on a scatter plot, all datasets generates a different kind of plot that is not interpretable by any regression algorithm which is fooled by these peculiarities and can be seen as follows:



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

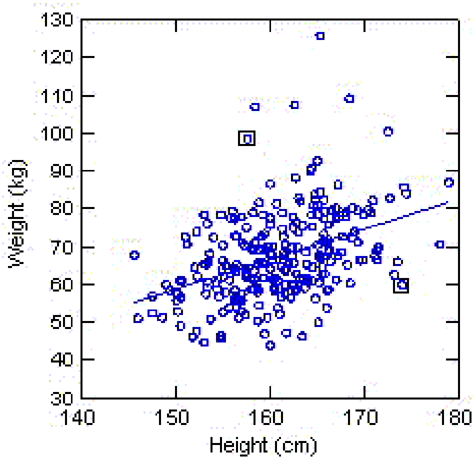
Dataset 4: shows the **outliers** involved in the dataset which **cannot be handled** by linear regression model

Conclusion:

We have described the four datasets that were intentionally created to describe the importance of data visualisation and how any regression algorithm can be fooled by the same. Hence, all the important features in the dataset must be visualised before implementing any machine learning algorithm on them which will help to make a good fit model

Q3: What is Pearson’s R?

Pearson's r is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.



"Tends to" means the association holds "on average", not for any arbitrary pair of observations, as the following scatterplot of weight against height for a sample of older women shows. The correlation coefficient is positive and height and weight tend to go up and down together. Yet, it is easy to find pairs of people where the taller individual weighs less, as the points in the two boxes illustrate.

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

Q4. What is scaling?
Why is scaling performed?
What is the difference between normalized scaling and standardized scaling?

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.

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.

Therefore, we need to scale features because of two reasons:

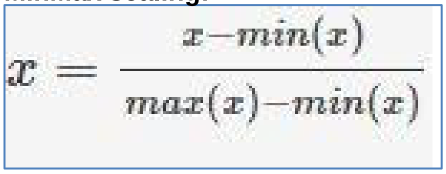
1. Ease of interpretation
2. Faster convergence for gradient descent methods.

NOTE: 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?

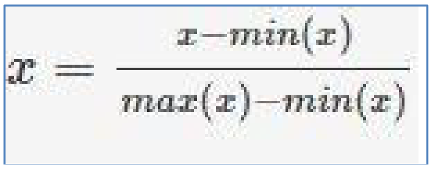
Normalization/Min-Max Scaling:
The variables are scaled in such a way that all the values lie between zero and one using the maximum and the minimum values in the data. Sklearn -> preprocessing -> MinMaxScaler helps to implement normalization in python.

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). 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.



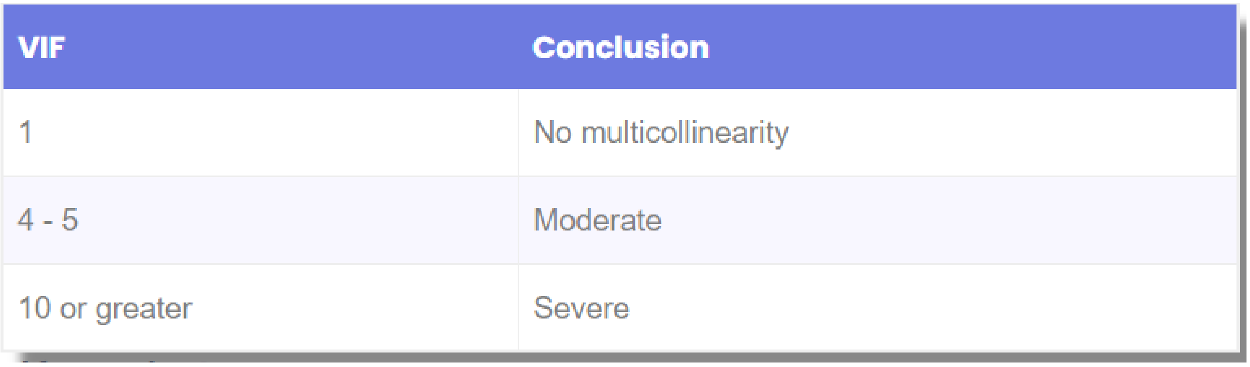
$$x = \frac{x - min(x)}{max(x) - min(x)}$$

Q5: You might have observed that sometimes the value of VIF is infinite. Why does this happen?

If all the independent variables are orthogonal to each other, then VIF = 1.0. If there is perfect correlation, then VIF = infinity. 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. An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

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.

A general rule of thumb is that if VIF > 10 then there is multicollinearity. Note that this is a rough rule of thumb, in some cases we might choose to live with high VIF values if it does not affect our model results such as when we are fitting a quadratic or cubic model or depending on the sample size a large value of VIF may not necessarily indicate a poor model.



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

Q6: What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression

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.

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 behavior

Interpretation:

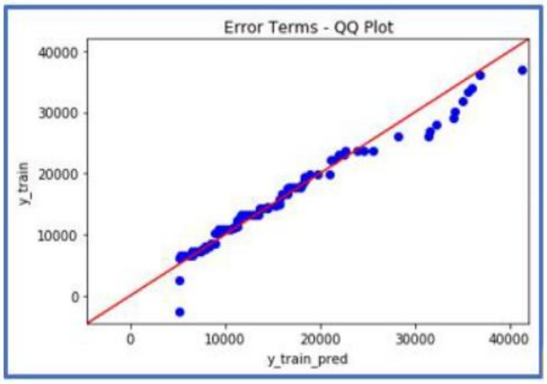
A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.
Below are the possible interpretations for two data sets.

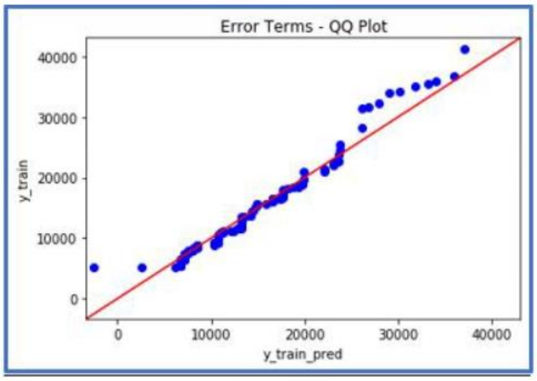
*a) **Similar distribution**: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x -axis*

*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*





Linear Regression Assignment

Assignment-based Subjective Q&A

MOHAMMED RAHAMATHULLA
iiitB and UpGrad Executive PG Programme
Machine Learning & Artificial Intelligence
mohammedrahmat@gmail.com

1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?**

The demad of bike is less in the month of spring when compared with other seasons. The demand bike increased in the year 2019 when compared with year 2018.

2. **Why is it important to use drop_first=True during dummy variable creation?**

drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables

3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**

The numerical variable 'registered' has the highest correlation with the target variable 'cnt' , if we consider all the features.

4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**

- There should be a linear and additive relationship between dependent (response) variable and independent (predictor) variable(s). A linear relationship suggests that a change in response Y due to one unit change in X^1 is constant, regardless of the value of X^1. An additive relationship suggests that the effect of X^1 on Y is independent of other variables.
- There should be no correlation between the residual (error) terms. Absence of this phenomenon is known as Autocorrelation.
- The independent variables should not be correlated. Absence of this phenomenon is known as multicollinearity.
- The error terms must have constant variance. This phenomenon is known as homoskedasticity. The presence of non-constant variance is referred to heteroskedasticity.
- The error terms must be normally distributed.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

1. **Temperature**
2. **weathersit**: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
3. **year**