Linear Regression Assignment

Assignment-based Subjective Questions:

1A.

- Weather wise impact is that, boombikes are most often used when weather is clear, demand is also good when there is mist or cloudy. But demands drops when there is light rain or heavy rain.
- Most demand for bikes comes during Fall followed by Summer. Spring has least demand.
- Demand for bikes have increased in 2019 compared to 2018
- Whether is working day or not, the demand remains the same

2A.

It is necessary to drop_first=True, because it causes Dummy variable trap, which might lead to wrong predictions.

3A. Temp (and aTemp) has highest correlation with target variable

---

4A. To validate Linear regression after model building, below steps can be ensured:

- Make sure residual theoretical and actual plot is following same pattern
- Make sure R2 and adj R2 are nearing to 100
- Make sure pvalue is less than 0.05

5A. Temp, weather situation and season are 3 top features contributing to explaining demand.

General Subjective Questions Answers:

1A.

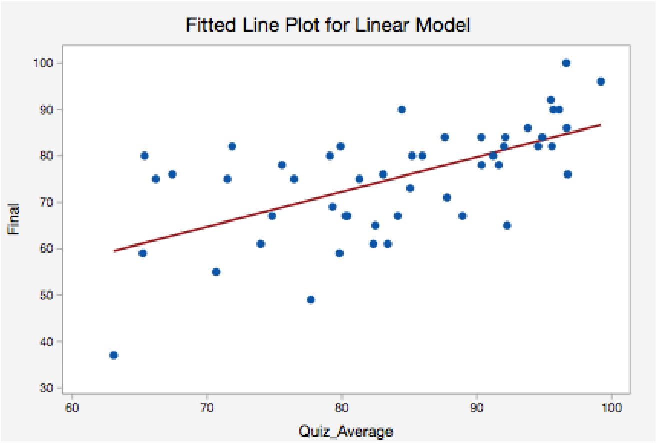
Linear Regression is a method used to define a relationship between a dependent variable (Y) and independent variable (X). Which is simply written as :

$$Y=mx+c$$

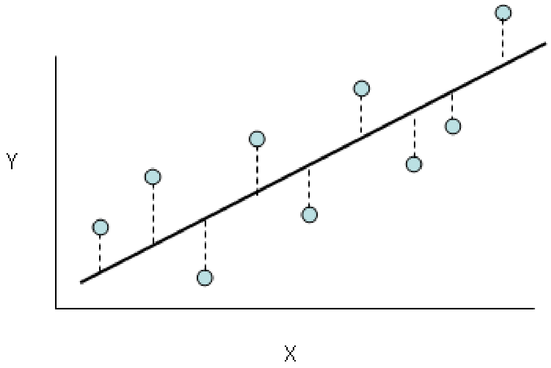
Where y is the dependent variable, m is the scale factor or coefficient, b being the bias coefficient and X being the independent variable. The bias coefficient gives an extra degree of freedom to this model. The goal is to draw the line of best fit between X and Y which estimates the relationship between X and Y.

Ordinary Least Mean Square

Earlier we discussed estimating the relationship between X and Y to a line. For example, we get sample inputs and outputs and we plot these scatter point on a 2d graph, we something similar to the graph below :



The line seen in the graph is the actual relationship we going to accomplish, And we want to minimize the error of our model. This line is the best fit that passes through most of the scatter points and also reduces error which is the distance from the point to the line itself as illustrated below.



And the total error of the linear model is the sum of the error of each point. I.e. ,

$$\sum_{i=1}^{n} r_i^{\ 2}$$

ri = Distance between the line and ith point.

n =Total number of points.

We are squaring each of the distance's because some points would be above the line and some below. We can minimize the error of our linear model by minimizing r thus we have

$$\beta_i = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \overline{\beta}_1 \bar{x}$$

where x¯ is the mean of the input variable X and y¯ being the mean of the output variable Y.

2A.

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 analyzing 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.

3A.

The Pearson correlation coefficient (r) is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables. When one variable changes, the other variable changes in the same direction.

4A.

Feature scaling is one of the most important data preprocessing step in machine learning. Algorithms that compute the distance between the features are biased towards numerically larger values if the data is not scaled.

Tree-based algorithms are fairly insensitive to the scale of the features. Also, feature scaling helps machine learning, and deep learning algorithms train and converge faster.

There are some feature scaling techniques such as Normalization and Standardization that are the most popular and at the same time, the most confusing ones.

Let’s resolve that confusion.

Normalization or Min-Max Scaling is used to transform features to be on a similar scale. The new point is calculated as:

X_new = (X - X_min)/(X_max - X_min)

This scales the range to [0, 1] or sometimes [-1, 1]. Geometrically speaking, transformation squishes the n-dimensional data into an n-dimensional unit hypercube. Normalization is useful when there are no outliers as it cannot cope up with them. Usually, we would scale age and not incomes because only a few people have high incomes but the age is close to uniform.

Standardization or Z-Score Normalization is the transformation of features by subtracting from mean and dividing by standard deviation. This is often called as Z-score.

X_new = (X - mean)/Std

Standardization can be helpful in cases where the data follows a Gaussian distribution. However, this does not have to be necessarily true. Geometrically speaking, it translates the data to the mean vector of original data to the origin and squishes or expands the points if std is 1 respectively. We can see that we are just changing mean and standard deviation to a standard normal distribution which is still normal thus the shape of the distribution is not affected.

Standardization does not get affected by outliers because there is no predefined range of transformed features.

Difference between Normalization and Standardization

S.NO.	Normalization	Standardization
1.	Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
2.	It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
3.	Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
4.	It is really affected by outliers.	It is much less affected by outliers.

S.NO.	Normalization	Standardization
5.	Scikit-Learn provides a transformer called MinMaxScaler for Normalization.	Scikit-Learn provides a transformer called StandardScaler for standardization.
6.	This transformation squishes the n-dimensional data into an n-dimensional unit hypercube.	It translates the data to the mean vector of original data to the origin and squishes or expands.
7.	It is useful when we don’t know about the distribution	It is useful when the feature distribution is Normal or Gaussian.
8.	It is a often called as Scaling Normalization	It is a often called as Z-Score Normalization.

5A.

Variation Inflation Factor would be infinite when there is perfect correlation. A large value of VIF indicates that there is a correlation between the variables.

6A.

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.

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