Linear Regression Assignment Subjective Questions

-By Uttkarsh Mishra

(uttkarsh.mishra.ut@gmail.com)

Assignment-based Subjective Questions

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

Ans. Categorical data which is mostly nominal data, if not converted to dummy variables would affect the result of dependent variable.

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

Ans. If we have categorical variable with n-levels, then we need to use n-1 columns to represent the dummy variables.

It is important to use drop_first=True, 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?

Ans. Looking at the pair-plot among numerical variables, 'registered' column having highest correlation with trget variable 'cnt'.

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

Ans. After Building the model on training set, we de residual analysis to check if the error terms are also normally distributed (which is infact, one of the major assumptions of linear regression)

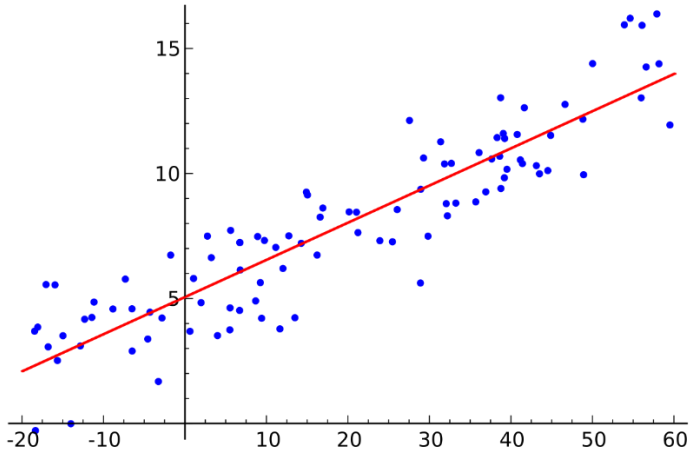
5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Ans. Based on the final model, top 3 features contributing significantly towards explaining the demand of shared bikes are 'yr', 'temp', 'hum'.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans. Linear Regression is a supervised machine learning algorithm where the predicted output is continuous and has a constant slope. It's used to predict values within a continuous range, (e.g. sales, price) rather than trying to classify them into categories (e.g. cat, dog).



Linear Regression

Simple Regression:

Simple linear regression is a type of regression analysis where the number of independent variables is one and there is a linear relationship between the independent(x) and dependent(y) variable. The red line in the above graph is referred to as the best fit straight line. Based on the given data points, we try to plot a line that models the points the best. The line can be modelled based on the linear equation shown below.

Equation of simple linear equation

$y = \beta 1 * x + \beta 0$

$\beta 1$ = Slope

$\beta 0$ = Intercept

Assumptions in Linear Regression:

- The relation between the dependent and independent variables should be almost linear.
- The data is homoscedastic, meaning the variance between the results should not be too much.

- The results obtained from an observation should not be influenced by the results obtained from the previous observation.
- The residuals should be normally distributed. This assumption means that the probability density function of the residual values is normally distributed at each independent value.

Model Performance:

R – Square (R^2)

$$R^2 = \frac{TSS - RSS}{TSS}$$

- R^2 always ranges between 0 to 1.
- R^2 of 0 means that there is no correlation between the dependent and the independent variable.
- R^2 of 1 means the dependent variable can be predicted from the independent variable without any error.

Root Mean Square Error (RMSE):

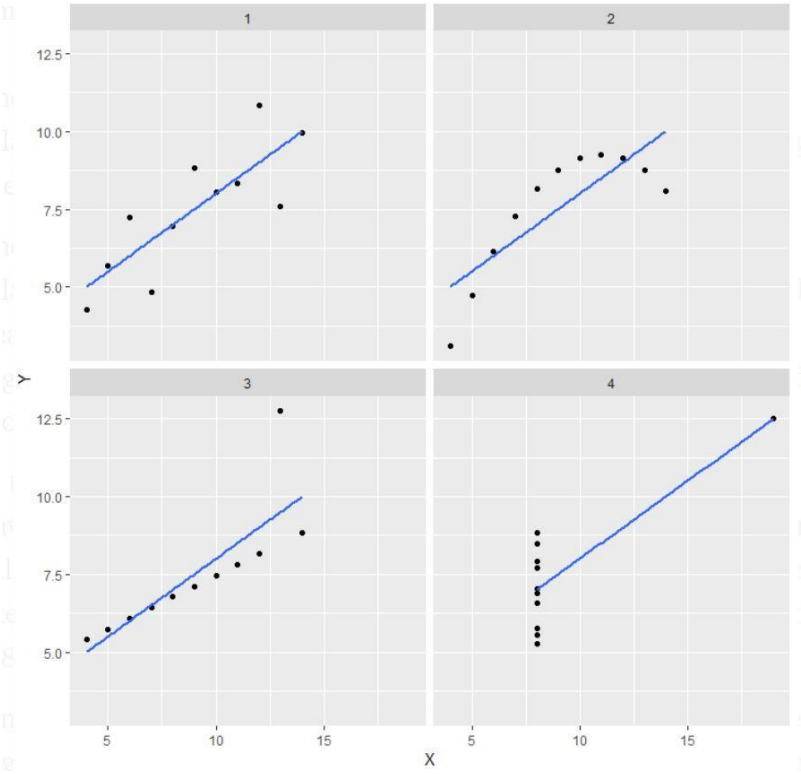
Root Mean Square Error (RMSE) is the standard deviation of the residuals (prediction errors). The formula for calculating RMSE is:

$$R^2 = \{(1/N) * \sum [(x_i - mean(x)) * (y_j - mean(y))] / (\sigma_x * \sigma_y) \}^2$$

Where N : Total number of observations

2. Explain the Anscombe's quartet in detail.

Ans. Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset consists of eleven (x, y) points.



Explanation of this output:

- In the first one (top left) if you look at the scatter plot you will see that there seems to be a linear relationship between x and y.
- In the second one (top right) if you look at this figure you can conclude that there is a non-linear relationship between x and y.
- In the third one (bottom left) you can say when there is a perfect linear relationship for all the data points except one which seems to be an outlier which is indicated be far away from that line.
- Finally, the fourth one (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient.

Application:

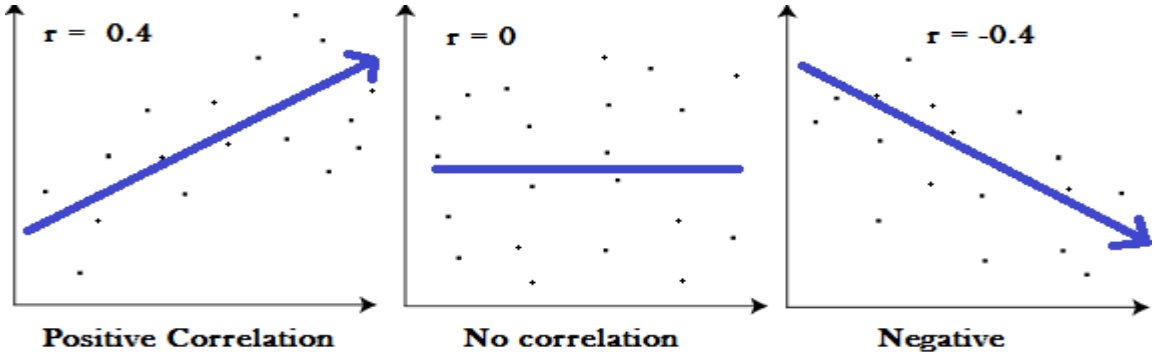
The quartet is still often used to illustrate the importance of looking at a set of data graphically before starting to analyse according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets.

3. What is Pearson's R?

Ans. The **Pearson correlation coefficient** is a statistical formula that measures the strength between variables and relationships. In the field of statistics, this formula is often referred to as the **Pearson R test**. When conducting a statistical test between two variables, it is a good idea to conduct a Pearson correlation coefficient value to determine just how strong that relationship is between those two variables.

In order to determine how strong the relationship is between two variables, a formula must be followed to produce what is referred to as the **coefficient value**. The coefficient value can range between -1.00 and 1.00. If the coefficient value is in the negative range, then that means the relationship between the variables is negatively correlated, or as one value increases, the other

decreases. If the value is in the positive range, then that means the relationship between the variables is positively correlated, or both values increase or decrease together.



Formula:

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

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

Ans. Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.

We need to scale feature for:

a. Ease of interpretation

b. Faster convergence of Gradient descent Method.

Eg: If an algorithm is not using feature scaling method then it can consider the value 3000 meter to be greater than 5 km but that's actually not true and in this case, the algorithm will give wrong predictions. So, we use Feature Scaling to bring all values to same magnitudes and thus, tackle this issue.

Scaling Methods:

a. MinMax Scaling: It brings all the data in range 0-1, where 0 will be minimum value and 1 being maximum.

$$X_{new} = \frac{X_i - \min(X)}{\max(x) - \min(X)}$$

b. Standardization: It brings all the data into a standard normal distribution with mean= 0 and standard deviation = 1.

$$X_{new} = \frac{X_i - X_{mean}}{\text{Standard Deviation}}$$

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

Ans. The Variance Inflation Factor (VIF) is a measure of colinearity among predictor variables within a multiple regression. It is calculated by taking the the ratio of the variance of all a given model's betas divide by the variane of a single beta if it were fit alone.

If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables.

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables.

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

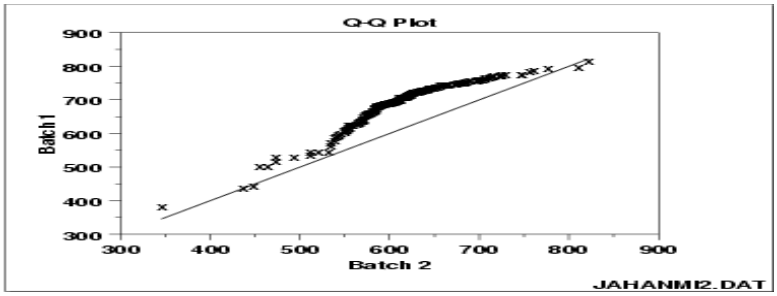
Ans. The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution.

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value.

A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions.

The advantages of the q-q plot are:

1. The sample sizes do not need to be equal.
2. Many distributional aspects can be simultaneously tested. For example, shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot. For example, if the two data sets come from populations whose distributions differ only by a shift in location, the points should lie along a straight line that is displaced either up or down from the 45-degree reference line.


Q-Q Plot
900
800
700
600
500
400
300
Batch 1
300
400
500
600
700
800
900
Batch 2
JAHANMI2.DAT