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: If the categorical is dependent variable here refers to as a binary, ordinal, nominal or event count variable. When the dependent variable is categorical, the ordinary least squares (OLS) method can no longer produce the best linear unbiased estimator ,that is, the OLS is biased and inefficient.

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

Ans: 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?

Ans: The variables considered are: 'temp', 'atemp', 'hum', 'windspeed' and the target variable considered is 'cnt'. The highest correlation with the target variable is 'hum'.

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

Ans: The assumptions of linear regression are:

- Linear Relationship between the features and target
- Little or no Multicollinearity between the features
- Homoscedasticity Assumption
- Normal distribution of error terms
- Little or No autocorrelation in the residuals

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

Ans: The top 3 features contributing significantly towards explaining the demand of the shared bikes are:

- temp
- working day
- windspeed

General Subjective Questions

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

Ans: There are types of regressions they are:

- Linear Regression
- Multiple Linear Regression
- Logistic Regression
- Polynomial Regression

Linear Regression (LR):

- Linear regression is a machine learning technique where a model is used to predict the given data based on some variables. Mathematically it is represented as Y=mX+c.
- There are 2 types of linear regression they are:

1. Simple linear regression
2. Multiple linear regression

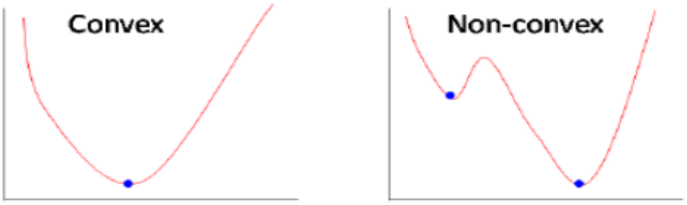
- Regression techniques mostly differ based on the number of independent variables and the type of relationship between the independent and dependent variables.y = b0 + b1 * x, linear regression algorithm is to find the best values for b0 and b1.

- Two important concepts you must know to better understand linear regression.

1. Cost Function
2. Gradient Descent

- **Cost Function**: The cost function help us to find out possible values for b0 and b1 which provide the best fit line for data points. Since for the best values for b0 and b1, convert the search problem into a minimization problem where we minimize the error between the predicted value and given value.The difference between predicted values and ground measures results the error difference. By squaring the error difference and sum of all data points and divide the value by the total no.of data points. Then it provides the average squared error over all data points. This cost function is also known as the Mean Squared Error(MSE) function.

- **Gradient Descent**: This is a method of updating b0 and b1 to reduce cost function.This method helps us on how to change the values.In this algorithm, the no.of steps we consider is the learning rate. This decides on how fast the algorithm converges to the minima.

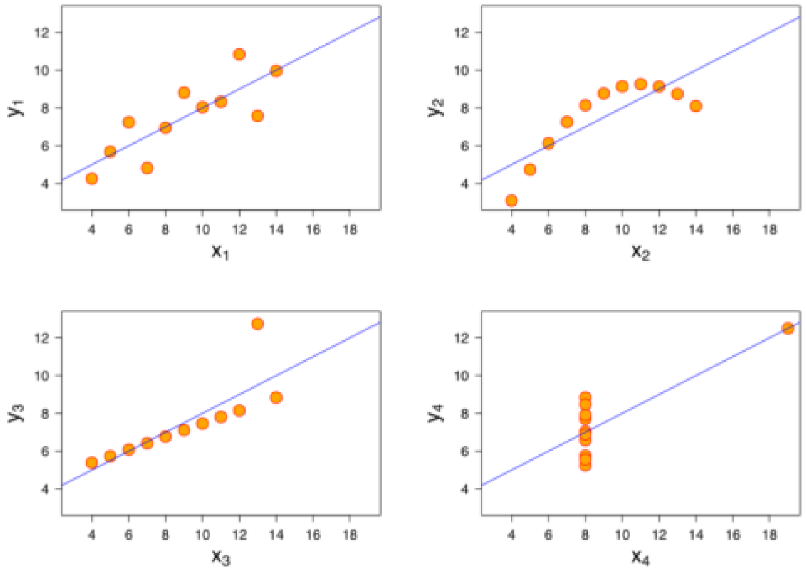


- Sometimes the cost function can be a non-convex function where local minima is for linear regression, it is always a convex function. How to use gradient descent to update b0 and b1. To update b0 and b1, we take gradients from the cost function. To find these gradients, we take partial derivatives with respect to b0 and b1.

2) Explain the Anscombe's quartet in detail?

Ans: Anscombe's Quartet can be defined as group of 4 datasets which are identical in simple descriptive statistics. There are 4 dataset plots which have same statistical observations, also provides same statistical information which involves variance, and mean of all x, y points in all 4 datasets.

The importance of visualising data before applying various algorithms is to build models out of them where the data features are plotted in order to see the distribution of samples that can help us to identify the various anomalies present in the data like outliers, diversity of the data etc. The Linear Regression can also be considered as fit for the data and is incapable of handling any other kind of datasets.



The four datasets are:

X1: This fits the linear regression model pretty well.

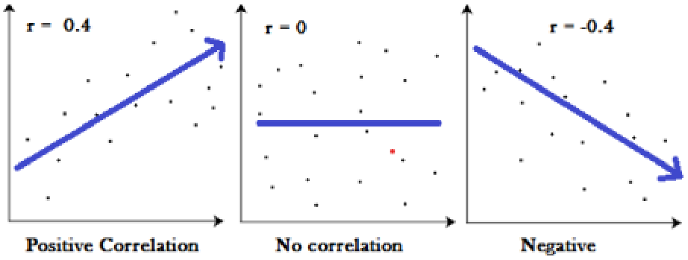
X2: This could not fit linear regression model on the data quite well as the data is non-linear.

X3: Shows the outliers involved in the dataset which cannot be handled by linear regression model.

X4: shows the outliers involved in the dataset which cannot be handled by linear regression model.

3) What is Pearson's R?

Ans: There are several types of correlation coefficient, one of them is Pearson's R. Pearson's correlation is a correlation coefficient commonly used in linear regression. Correlation coefficient formulas are used to find how strong a relationship is between data. The formulas return a value between -1 and 1:



- 1 indicates a strong positive relationship.
- -1 indicates a strong negative relationship.
- A result of zero indicates no relationship at all.

The absolute value of the correlation coefficient gives us the relationship strength. The larger the number, the stronger the relationship. The most commonly used formula is Pearson's correlation coefficient formula is

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[\,n\Sigma x^2 - (\Sigma x)^2\,]\,[\,n\Sigma y^2 - (\Sigma y)^2\,]}}$$

There are other correlation coefficients they are:
1. Sample correlation coefficient and
2. The population correlation coefficient.

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

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

Why Scaling: If scaling is not applied 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.

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.

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Standardization Scaling: Standardization is a scaling technique where values are centered around the mean with unit standard deviation. This means that the mean attribute becomes zero and the resultant distribution has a unit standard deviation.

$$X' = \frac{X - \mu}{\sigma}$$

μ is the mean of the feature values and σ is the standard deviation of the feature values. Note that in this case, the values are not restricted to a particular range.

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 ratio of the variance of all a given model's betas divide by the variance 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. 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 multi collinearity.

Infinite VIF mean:

The user has to select the variables to be included by ticking off the corresponding check boxes. 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).

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

Ans: Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data came from some theoretical distribution such as a Normal, exponential. 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.

The q-q plot is formed by:
Vertical axis: Estimated quantiles from data set 1
Horizontal axis: Estimated quantiles from data set 2

Both axes are in units of their respective data sets. That is, the actual quantile level is not plotted.

If the data sets have the same size, the q-q plot is essentially a plot of sorted data set 1 against sorted data set 2. If the data sets are not of equal size, the quantiles are usually picked to correspond to the sorted values from the smaller data set and then the quantiles for the larger data set are interpolated