Linear Regression Assignment Questions

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?**

In my analysis I found 4 Categorical Variables

- **yr**
- **season**
- **holiday**
- **weathersit**

Out of which yr and season had very good corelation with the dependent variable (**cnt**), while weathersit had a little less correlation and holiday had very less correlation. Which makes perfect sense as bikes will be rented very frequently around certain seasons and in good weather. And according to the dataset we can remember that the renting/sales around one of the years were high which could explain the high correlation between yr and the target field.

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

While creating dummy variables from a categorical variable we can easily say that let's say for 3 categories if 2 of the categories are not met then the third is definitely met. And following this logic we can surely remove one of the dummy variables as it can be inferred using this logic. And `drop_first=True` does exactly this in pandas pd.get_dummies function.

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

`registered` variable had the highest correlation with the target variable, it almost appears like a line chart itself while plotting scatter plot between registered and target field `cnt`

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

We do Residual Analysis on the training set compared with the predicted target values of training set. If the gathered error terms between actual training target values and predicted target values appears to be a normalized distribution curve then we can say that the model is fit and it is not predicting correctly by chance. Also we confirm that this error terms scatter plot against target values have no pattern which reinforces the model fitness even more.

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

Top 3 features that correlate to the target variable the most and represents the linear regressions line very likely are as follows.

- **Intercept (const)**
- **casual**
- **atemp**

General Subjective Questions

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

Linear Regression is a supervised machine learning algorithm that computes the *linear relationship between a dependent variable and one or more independent feature (relational) variables*. It is primarily used for predicting, forecasting, time-series modelling or in other words determining and **cause-effect relationship between variables of a dataset**.

2. **Explain the Anscombe's quartet in detail.**

Anscombe's Quartet is a set of four dataset that are having identical statistical properties like means, variance, Residual Squared and linear regression lines but have very different representations when we plot a scatter plot for all those four datasets using any plotting method.

The datasets were created by Francis Anscombe to demonstrate the importance of visualizing data and to determine that summary statistics alone can be very misleading.

The four datasets here are

- *Simple linear relation*, where 2 variables form a very evident linear relationship on scatter plot and also have corresponding statistical features
- *Non linear relation,* here the 2 variables have some relationship but the relation is not linear in nature and still the same statistical features are formed
- *Linear relation with outlier,* here the 2 variables have simple linear relationship but also have a outlier in the dataset, this outlier is distanced enough that it can cause significant offset to the calculated regression line and lowers the correlation coef significantly. (by about 1 to 0.816)
- *One high leverage point,* here the 2 variables have no linear relation at all but the dataset have a high leverage point and it is causing the statistical features to form the same observations about the dataset while there is no evident relation between the variables at all.

3. **What is Pearson's R?**

Pearson's R or Pearson Correlation Coefficient, is a normalized measure of covariance between 2 variables and can have values between -1 to 1. This measure helps in determining if any variable has good enough correlation to our target variable of a linear relationship graph of not.

Mathematically the Pearson's R is covariance ratio of 2 variables divided by the product of their standard deviations.

For example, 2 variables like height and age of a student in high school will have significantly positive Pearson correlation coefficient.

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

Scaling or Feature Scaling is the method of standardizing the independent variables of a multi-feature dataset to a fixed range. As the range of different features in a data can be very different, the correlation of those variables can have biased effects on the target/dependent variable as some of the features can be very large in value while other can be minuscule. This difference in range can create bias in a linear regression model.

We perform this method so as to achieve less feature range bias in the model itself. This will help the linear relationship to be much precise in nature and have better fit to the data.

Normalized Scaling *scales the feature using minimum and maximum values*, while **Standardized Scaling** uses *features mean and standard deviation to achieve scaling*. Normalized scaling is beneficial when the *feature distribution is unknown to the dataset*, while standardized scaling is helpful when the *feature distribution is known and is consistent all the way*. ***Standardized scaling is also referred as z-score scaling*** as it works on normalized consistently distributed feature dataset.

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

Infinite VIF or Variance Inflation Factor indicates that the variables under observation have *perfect* correlation between them. This means that variables are orthogonal to each other and your feature set has multicollinearity.
Mathematically this occurs when the R-Squared value of the variable in observation is 1 and causing the VIF to become infinite.

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

A Q-Q Plot is a graphical plotting of quantile of sample distribution versus the quantile of theoretical distribution of data. It is used to determine if the dataset in observation follow any particular probability distribution graph like normalized probability distribution. This can help in linear regression model analysis stage. While we can plot Q-Q plot of original target variable data and predicted target variable data and figure out if there is Residual relation between them or not. We can figure out that if the *error terms on these to datasets have a normalized distribution of not which is one of the major assumptions that is made in a Linear Regression Model*
It has other uses as well, where it can be used to determine if 2 sample datasets are from the same population or not, if the Q-Q plot of the 2 samples have a linear relation then the samples are from the same population and can be treated as such.