ASSIGNMENT: LINEAR REGRESSION

V S S Anirudh Sharma

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? (3 marks)

Categorical Factor in model	Effect on dependent variable
year (yr)	+ ve
holiday ($holiday$)	- ve
summer ($season$ = 2)	+ ve
winter ($season$ = 4)	+ ve
August ($mnth$ = 8)	+ ve
September ($mnth$ = 9)	+ ve
Bad weather ($weathersit$ = 3)	- ve

All the above impacts are understandable and explainable:

1. The company would have increased its reach by the year so 2019 in general would have more customers than 2018
2. For a set of customers, commuting is mainly for work so holidays would mean these people will not commute.
3. Summer and winter vacations will attract student commuters, either for work or fun.
4. Bad weather, obviously, will reduce a lot of bike usage.

The interesting observation here would be the special positive impact in the months of August and September. Why do we see increase in these months?

American college year starts in August-September (Fall Semester) and hence, we would have a lot of new students coming to campuses for the first time, using these bike services for commute until they purchase their own bike. This could probably be the reason behind this August-September spike. We need further data on the user demographics to test this hypothesis.

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

When converting categorical data (on n unique values) into one-hot-key encoded dummy variables, we get an n -size vector. These n values are dependent and any one of the values can be expressed as a linear combination of the other n-1.

Thus, since our basic requirement for the predictors is independence, we eliminate one of these n factors, so that the system of remaining n-1 becomes an independent set.

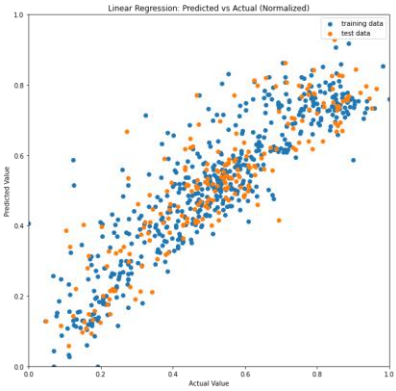
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Temperature (*temp*) and Feeling Temperature (*atemp*), with Pearson correlation of 0.63 with 'cnt'

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Following are the assumptions for Linear Regression:

1. Linearity: The relationship between X and the mean of Y is linear.
2. Homoscedasticity: The variance of residual is the same for any value of X.
3. Independence: Observations are independent of each other.
4. Normality: For any fixed value of X, Y is normally distributed about 0.

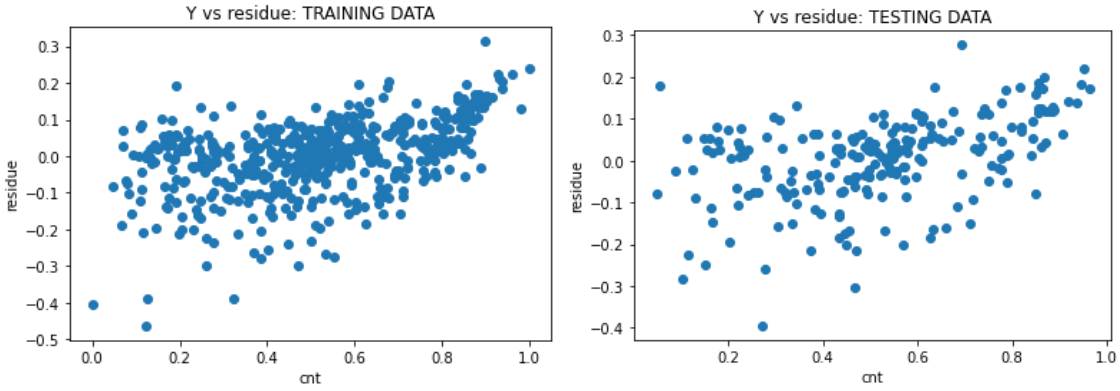


Let's look at each assumption

Linearity is validated by the final model performance. The high R-squared values of test and training data validate this assumption:

```
R-squared for training =  0.7949565194804429
R-squared for testing =  0.7775633717518898
```

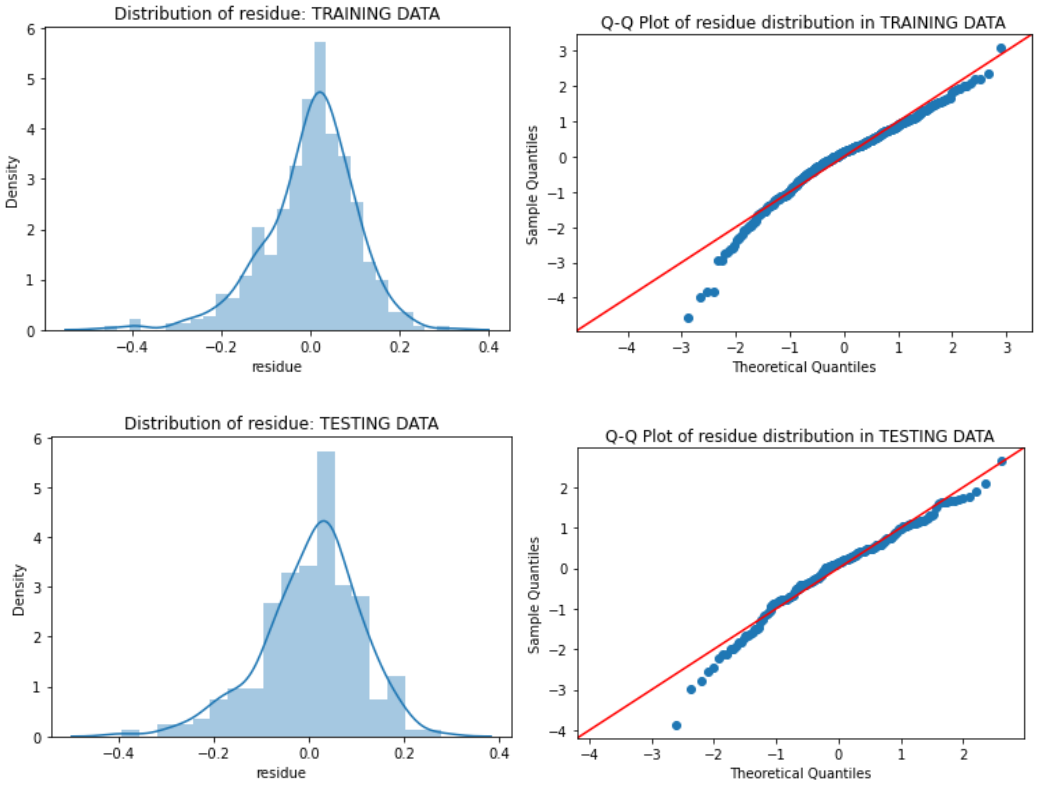
Homoscedasticity is validated by the plot of residue vs Y. As we see below, the variance in residue is independent of 'cnt'. It is uniform across.



Independence is validated by low VIF values between predictors (threshold 5):

Features	VIF
temp	3.82
Yr	1.93
season = 2	1.75
mnth = 8	1.55
season = 4	1.41
mnth = 9	1.28
weathersit = 3	1.05
holiday	1.03

Normality is validated by the distribution of residue for training and testing data and corresponding QQ plots. Clearly, both are reasonably close to a normal distribution around 0.



5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Our model is described as follows:

$$cnt_{normalized} = 0.0326 + 0.2320 \times yr - 0.0880 \times holiday + 0.0865 \times season_2 + 0.1392 \times season_4 + 0.0437 \times mnth_8 + 0.1076 \times mnth_9 - 0.2664 \times weathersit_3 + 0.5666 \times temp_{normalized}$$

Since all our data is normalized between 0 and 1, we can say that contribution of a feature increases with the magnitude of this normalized coefficient.

Thus, *temp*, weathersit (if weathersit = 3), and yr are top contributing features.

In simple words, the following are the top three things affecting usage of bike sharing system:

1. Temperature
2. If the weather is bad
3. Year

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

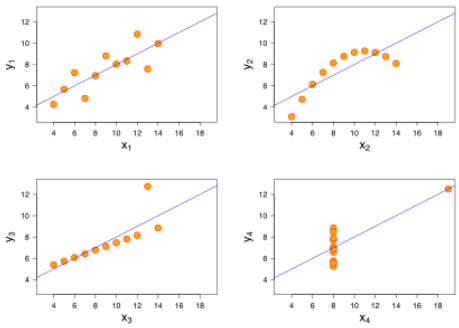
Linear Regression is a supervised learning algorithm used to predict a continuous dependent variable from a set of independent variables. It models the relationship between the dependent and independent variables as a linear equation. The goal is to find the line of best that minimizes the sum of the squared differences between the observed and predicted values. Let's understand all the elements of linear regression, step by step:

1. **Model**: Linear regression models the relationship between the dependent variable (y) and the independent variable(s) (x) as an equation of the form $y = b_0 + b_1x_1 + b_2x_2 + \cdots + b_nx_n$
2. **Cost Function:** To determine the best line of fit, a cost function is used. The most common cost function for linear regression is the mean squared error (MSE) given by $MSE = \frac{\sum(y_i - \hat{y}_i)}{n}$
3. **Optimization:** The optimization process is used to determine the best coefficients that minimize the cost function. This is typically done using gradient descent, which iteratively updates the coefficients until the cost function is minimized.
4. **Evaluation:** The performance of the linear regression model is evaluated by its coefficient of determination (R^2) which measures the proportion of variance in the dependent variable that is explained by the independent variables.
5. **Deployment**: Once the best coefficients are found, the linear regression model can be used to make predictions by plugging in new independent variable values into the equation.

2. Explain the Anscombe's quartet in detail. (3 marks)

Anscombe's quartet is a set of four datasets that have nearly identical statistical properties (mean, variance, correlation) yet appear quite different when plotted. The quartet was created by Francis Anscombe in 1973 to demonstrate the importance of visualizing data before analyzing it. The four datasets are:

1. Dataset 1: A set of 11 pairs of (x,y) values that form a roughly linear pattern.
2. Dataset 2: A set of 11 pairs of (x,y) values that form a perfect linear pattern.
3. Dataset 3: A set of 11 pairs of (x,y) values that form a curvilinear pattern.
4. Dataset 4: A set of 11 pairs of (x,y) values that form an outlier-influenced pattern.



The datasets show that summary statistics alone can be misleading and that it is important to visualize data to gain a full understanding of it. The quartet is often used in statistics and data analysis courses to emphasize the importance of visualizing data and the dangers of relying solely on summary statistics.

3. What is Pearson's R? (3 marks)

Pearson's R, also known as Pearson's correlation coefficient, is a measure of the linear association between two continuous variables. It ranges from -1 to 1, with -1 indicating a perfect negative linear relationship, 0 indicating no linear relationship, and 1 indicating a perfect positive linear relationship.

It is calculated as follows:

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

It is important to note that Pearson's R only measures linear relationships and cannot capture non-linear relationships between variables.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Scaling is a preprocessing step in machine learning and statistics that transforms the values of features (predictors, independent variables) to have a similar scale.

Scaling is performed to ensure that features are on a similar scale, which can improve the performance and stability of certain algorithms, and to help with algorithmic convergence.

Difference between normalized scaling and standardized scaling is as follows:

Normalization	**Standardization**
Also known as Min-Max scaling	Also known as Z-score normalization
Scales the values of each feature to the range [0, 1].	Scales the values of each feature to have a mean of 0 and a standard deviation of 1
$X_{\mathrm{i},scaled} = \frac{X_i - X_{min}}{X_{max} - X_{min}}$	$X_{i,scaled} = \frac{\left(X_i - mean(X)\right)}{stddev(X)}$

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

Infinite VIF would mean that the R_i^2 for a given variable w.r.t others is 1. This means that this variable can be perfectly expressed as a linear combination of some of the other variables with non-zero coefficients.

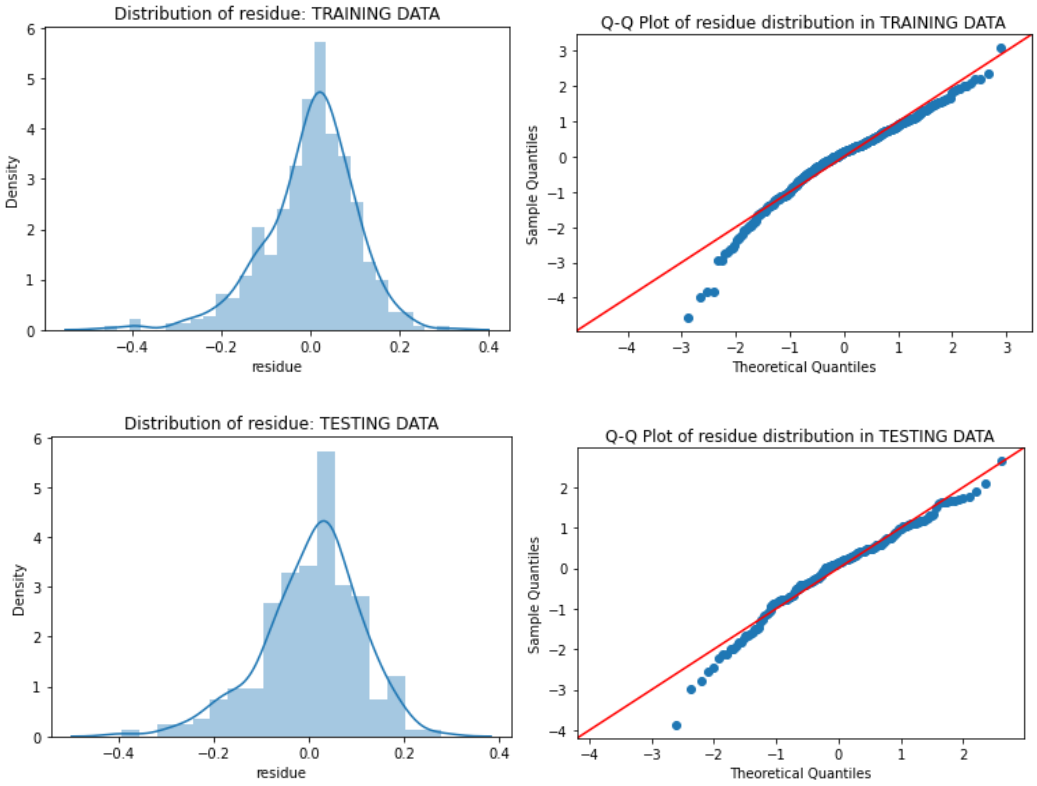
Though there were no factors with infinite VIF in the models explored in our project, such variables need to be eliminated to get rid of multicollinearity problem.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

A Q-Q (Quantile-Quantile) plot is a graphical tool used to assess the distributional assumptions of a set of data. In a Q-Q plot, the observed data is plotted against a theoretical distribution (e.g. a normal distribution) to visually assess if the observed data follows the same distribution as the theoretical distribution.

In the context of linear regression, a Q-Q plot is used to assess the assumption of normally distributed residuals. If the residuals are normally distributed, then the observed residuals will closely follow the diagonal line in the Q-Q plot.

The importance of a Q-Q plot in linear regression is that it provides a visual tool to assess the normality assumption of the residuals. If the residuals are not normally distributed, then the linear regression model may not be appropriate, and alternative models or data transformations should be considered.



The above plots are generated from our final model in the project. The residuals closely align with the diagonal line, indicating that the residuals are close to normally distributed.