Multiple Linear Regression Model Assignment

for Prediction of demand for Shared Bikes in US

Assignment by **M Sreenu**

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)

Ans: Season, Year, Month, Holiday, Weekday and Weathersit are categorical variables in the dataset. And from the analysis, it can be inferred that,

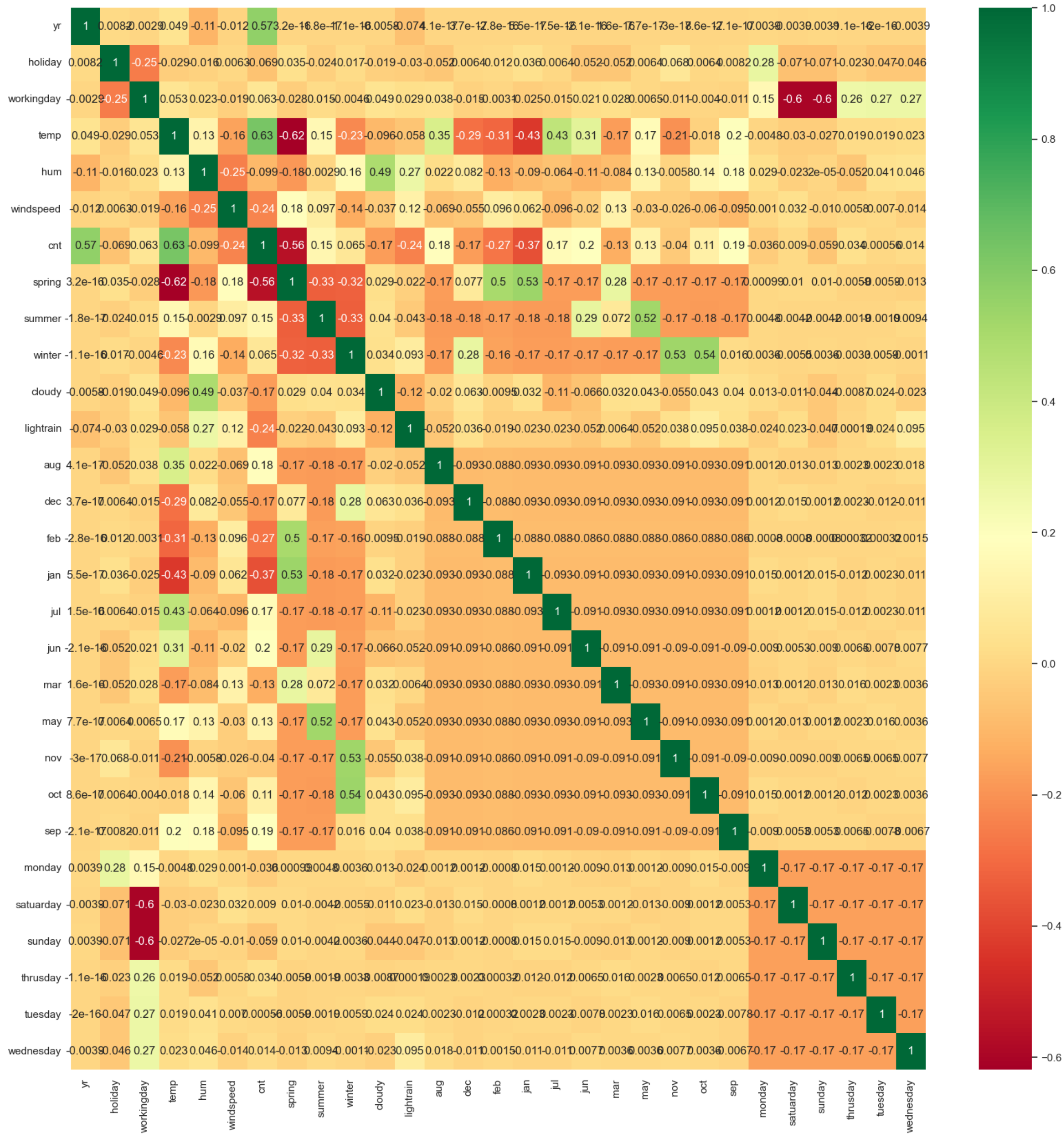
1. Season fall had maximum number of active customers (September). 2019 have seen more sales than 2018, by which we can infer that the service popularity or demand is increasing, potentially due to factors like effective marketing, positive word-of-mouth, or growth in consumer embracing the services.
2. Holidays impact the sales negatively, as decline in travel frequency and lack of regular transits may be due to consumers focusing more on family gatherings and leisure activities.
3. During heavy rain, there are no users whereas partly cloudy/clear sky saw reasonably high count of users.
4. Bookings have seen a high raise during the months of May through October. Trend seems to be increasing from the beginning of the year till mid of the year and from there it started decreasing as we approach the Yearend.
5. Clear weather days have more booking which seems reasonable.
6. Thursday through Sunday have elevated number of bookings when compared to the beginning of the week where the booking are low.
7. Apart from Holiday, bookings seemed to be identical on both working and non-working days, which can be inferred as the customer is using the bike-sharing service as a Convenience and treating it similarly on both weekends and weekdays.

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

Ans: “drop_first = True” is important to use, as not using it would make the dummy variables correlated to each other and hence, redundant, which is not expected of our analysis. 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? (1 mark)

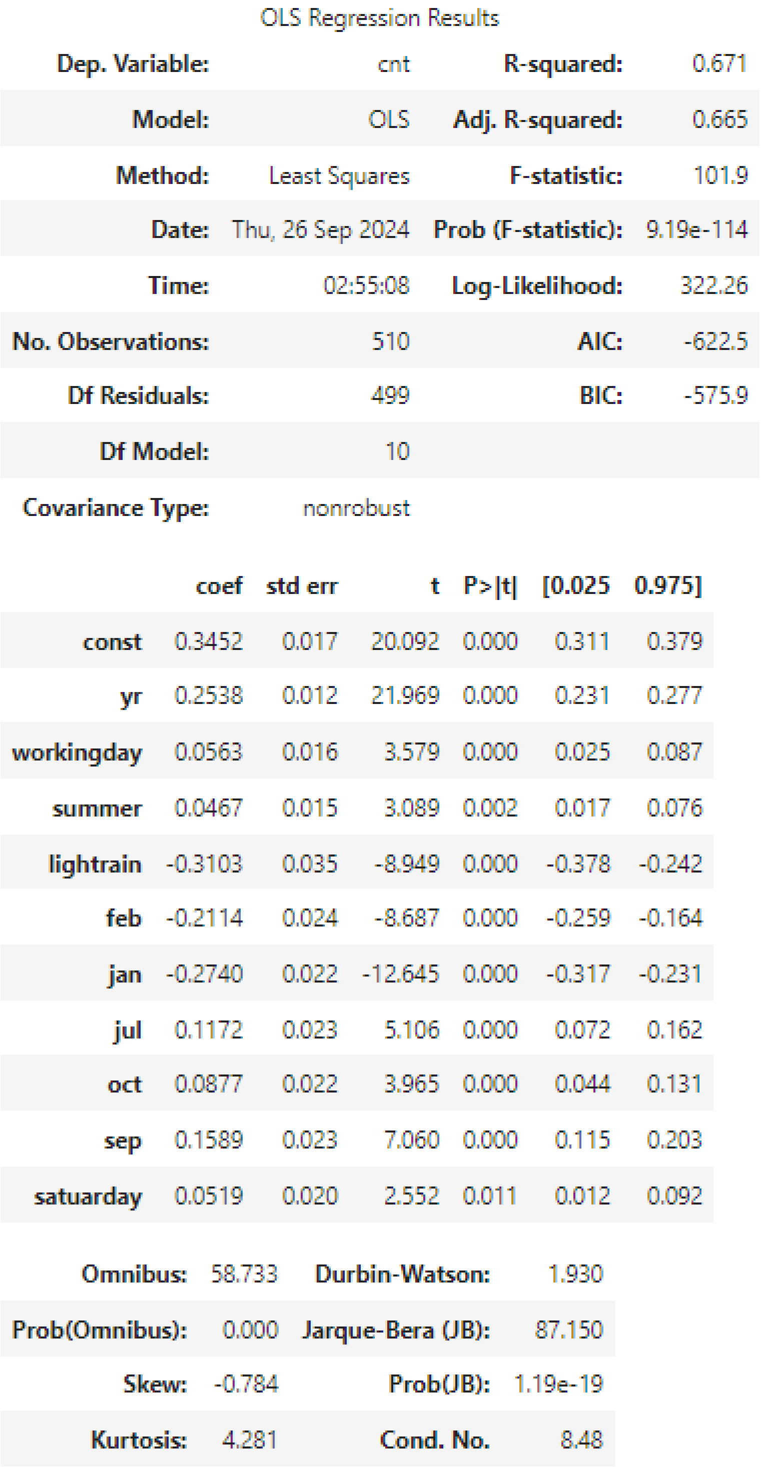
Ans: temp attribute is having the highest correlation with cnt (target variable) with 0.63 value (see below figure).



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

Ans: We can validate the model on the following scenarios.

1. Low p-value :- p-value should be low which is less than 0.05 is ideal.



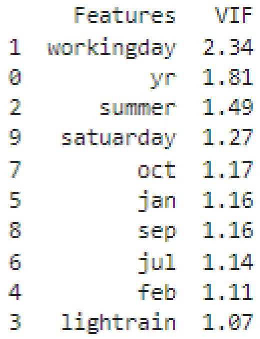
OLS Regression Results

Dep. Variable:	cnt	**R-squared:**	0.671
Model:	OLS	**Adj. R-squared:**	0.665
Method:	Least Squares	**F-statistic:**	101.9
Date:	Thu, 26 Sep 2024	**Prob (F-statistic):**	9.19e-114
Time:	02:55:08	**Log-Likelihood:**	322.26
No. Observations:	510	**AIC:**	-622.5
Df Residuals:	499	**BIC:**	-575.9
Df Model:	10		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	0.3452	0.017	20.092	0.000	0.311	0.379
yr	0.2538	0.012	21.969	0.000	0.231	0.277
workingday	0.0563	0.016	3.579	0.000	0.025	0.087
summer	0.0467	0.015	3.089	0.002	0.017	0.076
lightrain	-0.3103	0.035	-8.949	0.000	-0.378	-0.242
feb	-0.2114	0.024	-8.687	0.000	-0.259	-0.164
jan	-0.2740	0.022	-12.645	0.000	-0.317	-0.231
jul	0.1172	0.023	5.106	0.000	0.072	0.162
oct	0.0877	0.022	3.965	0.000	0.044	0.131
sep	0.1589	0.023	7.060	0.000	0.115	0.203
saturday	0.0519	0.020	2.552	0.011	0.012	0.092

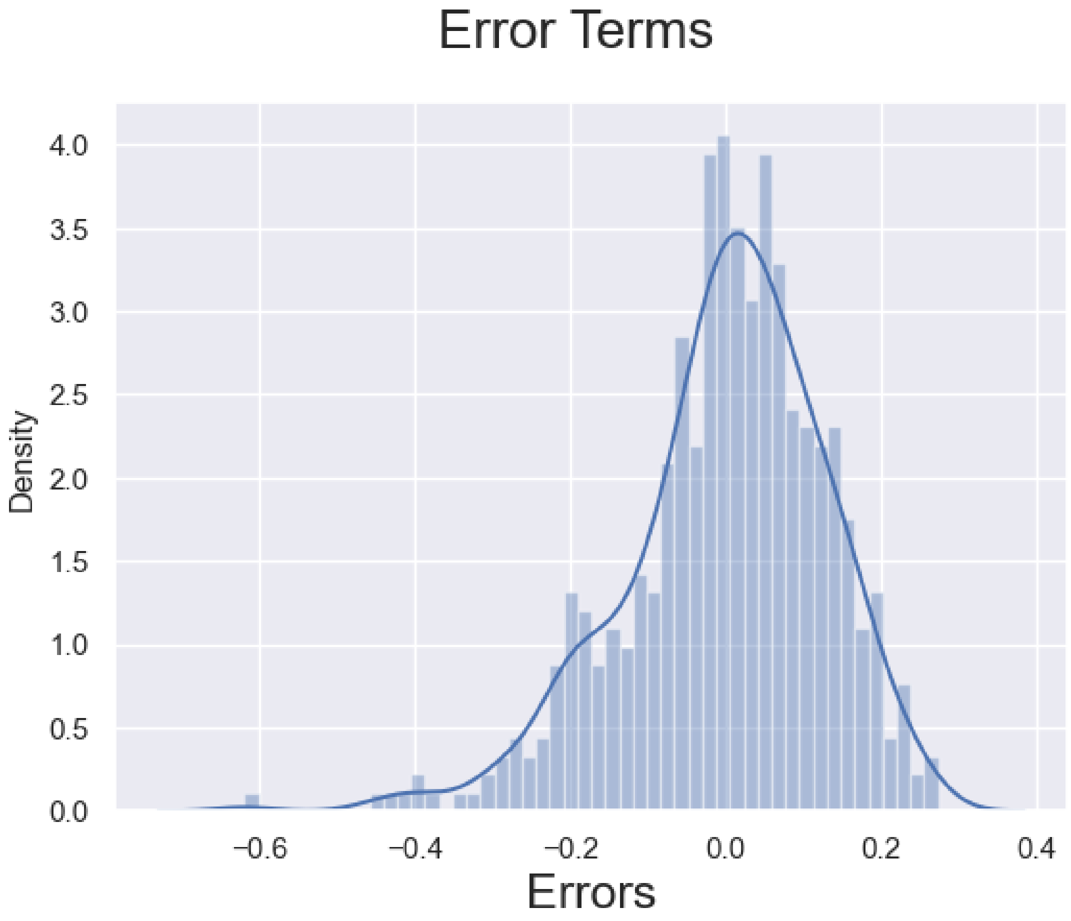
Omnibus:	58.733	**Durbin-Watson:**	1.930
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	87.150
Skew:	-0.784	**Prob(JB):**	1.19e-19
Kurtosis:	4.281	**Cond. No.**	8.48

2. Variance inflation factor (VIF): - VIF should ideally be lesser then 5 (VIF = 1/(1-R-square))



```
      Features   VIF
1   workingday  2.34
0           yr  1.81
2       summer  1.49
9     saturday  1.27
7          oct  1.17
5          jan  1.16
8          sep  1.16
6          jul  1.14
4          feb  1.11
3    lightrain  1.07
```

3. Error Rate: Normalised Error rate with centralized at zero.



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

Ans: Top 3 features contributing significantly towards explaining the demand of the shared bikes are Const, lightrain and Jan. lightrain and jan have negative relationship between the feature and the target variable (demand for shared bikes) though.

General Subjective Questions

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

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 catagories (e.g. cat, dog). There are two main types:

1. Simple regression: Simple linear regression uses traditional slope-intercept form, where m and b are the variables our algorithm will try to "learn" to produce the most accurate predictions. X represents our input data and y represents our prediction.

$$y=mx+b$$

2. Multivariable regression: A more complex, multi-variable linear equation might look like this, where w represents the coefficients, or weights, our model will try to learn.

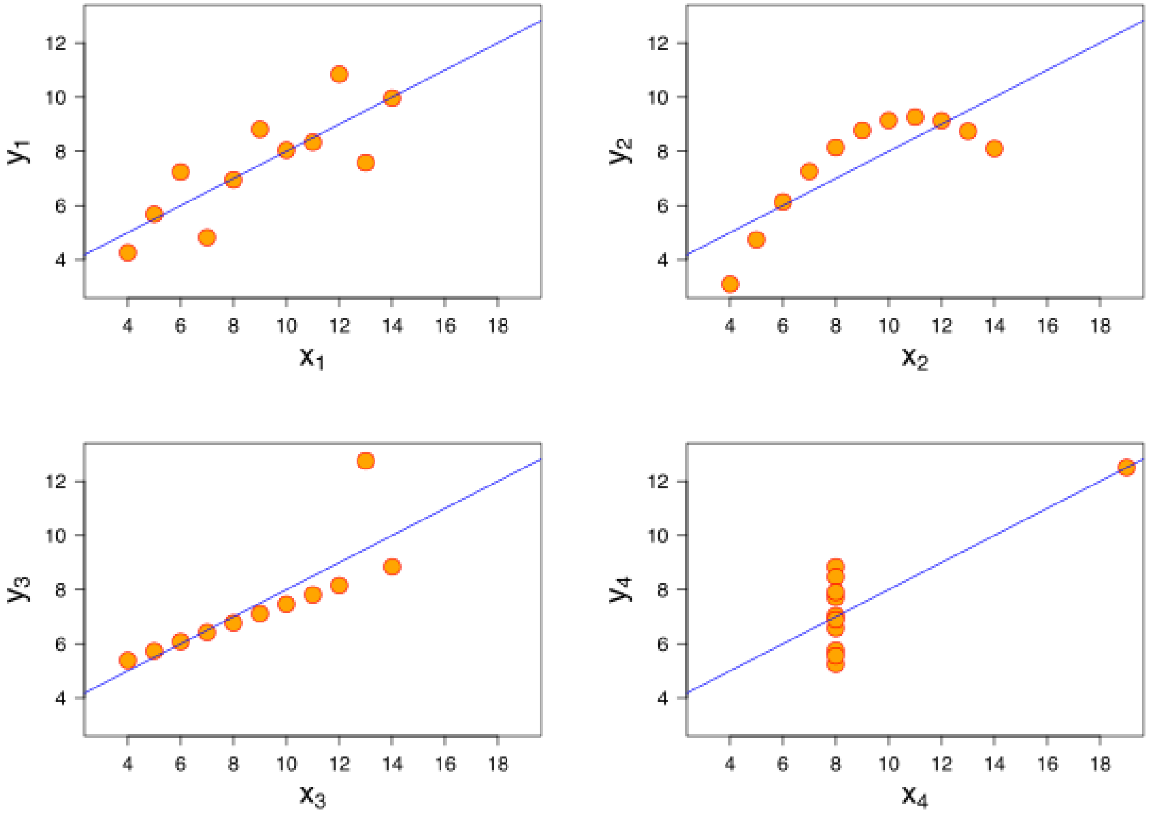
$$f(x,y,z)=w1x+w2y+w3z$$

For example: For sales predictions, these attributes might include a company's advertising spend on radio, TV, and newspapers.

$$Sales=w1Radio+w2TV+w3News$$

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

Ans: Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed.



These above datasets have same Mean(x), std(x), Mean(y), std(y) and Cor (x,y) but their prediction are completely different.

Conclusion: It is strongly recommended to look at data first then start performing linear regression or any other analysis.

3. What is Pearson’s R? (3 marks)

Ans: This is the test statistics that measures the statistical relationship, or association, between two continuous variables. It is known as the best method of measuring the association between variables of interest because it is based on the method of covariance. It gives information about the magnitude of the association, or correlation, as well as the direction of the relationship. Coefficient values can range from +1 to -1, where +1 indicates a perfect positive relationship, -1 indicates a perfect negative relationship, and a 0 indicates no relationship exists. Its formula is expressed as

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y}$$

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

Ans: Feature 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.

Normalized scaling: This is also called Min-Max scaling, This technique re-scales a feature or observation value with distribution value between 0 and 1.

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

Standardized saling: It is a very effective technique which re-scales a feature value so that it has distribution with 0 mean value and variance equals to 1.

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

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

Ans: Formula for VIF is 1/(1-R-square). If VIF is infinite means denominator is zero, R-square is equal to 1 or close to 1. Causes of infinite VIF are:

1. Linear dependence: A feature can be exactly predicted by a combination of other features.
2. Zero variance: A feature has no variation, making it impossible to estimate its effect.
3. Numerical instability: Computational issues, such as:
    - Rounding errors.
    - Overflow.

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

Ans: A Q-Q plot (Quantile-Quantile plot) is a graphical tool used to compare distributions of two datasets or check assumptions of linear regression. In linear regression, Q-Q plots helps to check if residuals follow a normal distribution, identify outliers and detect non-linear relationships. In linear regression it is important because it helps us ensures the robustness of regression results, also helps us identify potential issues in the data like Non-normal residuals, Outliers, Non-linear relationships, etc., It also helps us to improve models with transformation of variables and alternative models.