Linear Regression Assignment

- Ashok Vashist

Agenda

- Assignment Questions
- Answers

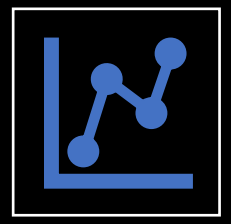
Question 1

From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? 3 Marks

Inference on Dependent variables

	coef
const	0.5886
yr	0.2481
windspeed	-0.1889
Spring	-0.2599
Summer	-0.0396
Winter	-0.0764
Jan	-0.1034
Sep	0.0697
Tuesday	-0.0462
Light_Snow	-0.2986
Mist	-0.0859

Weather Situation : As ‘windspeed’ ,‘Light Snow with rain’ & ‘Mist’ has negative correlation so it can be inferred that the bike sharing demand would decrease in these weather conditions.



Month : As ‘January’ month has negative correlation and ‘September’ month has positive correlation ; it can be inferred bike sharing demand would decrease in January and increase in September month .

Season : ‘Spring’, ‘Summer’ & ‘Winter’ is having negative coefficient with Bike Sharing demand, it indicates that these season have low demand of Bike sharing.

Question 2

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

Use **drop_first=True**

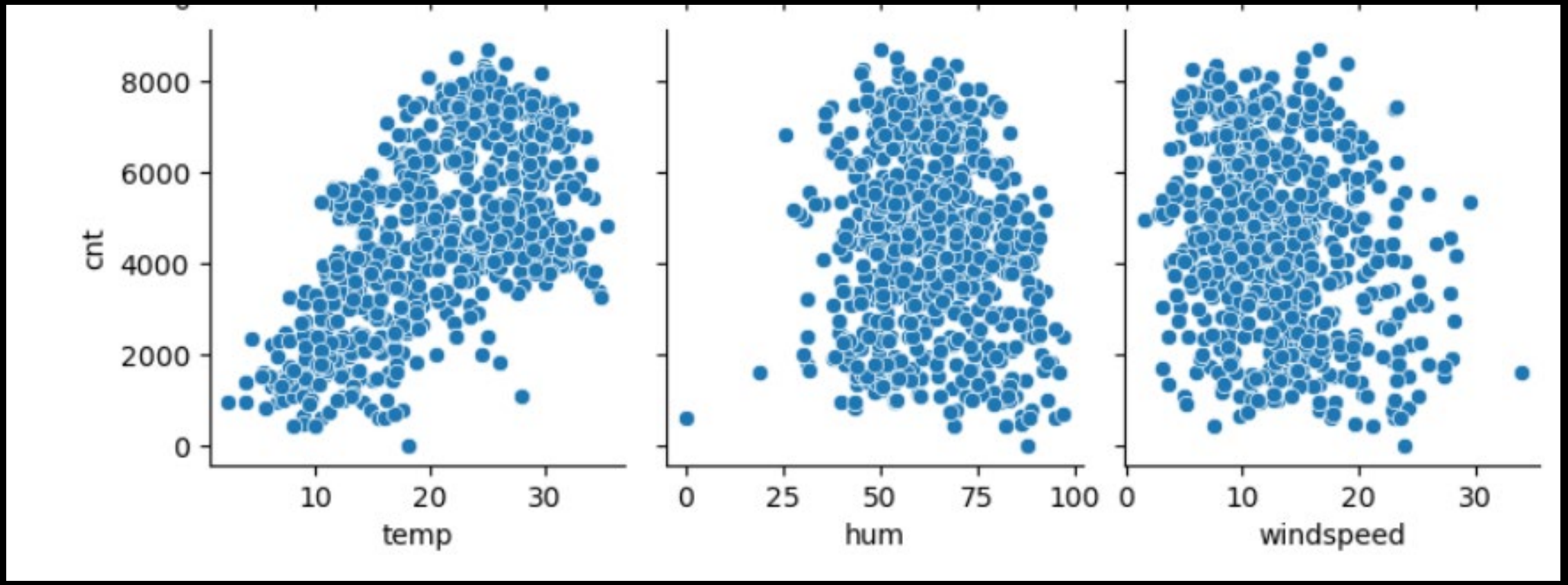
- Machine learning model performs better when we use optimized and relevant independent variables during model building.
- When we create dummy variables for Categorical Variable then removing one variable/column help us to keep less number of independent variables without loosing any important information.
- drop_first = True , helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

Question 3

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

Temperature has highest correlation with count

- From the below pair plot it is evident that temperature has highest correlation with target variable count (bike sharing)



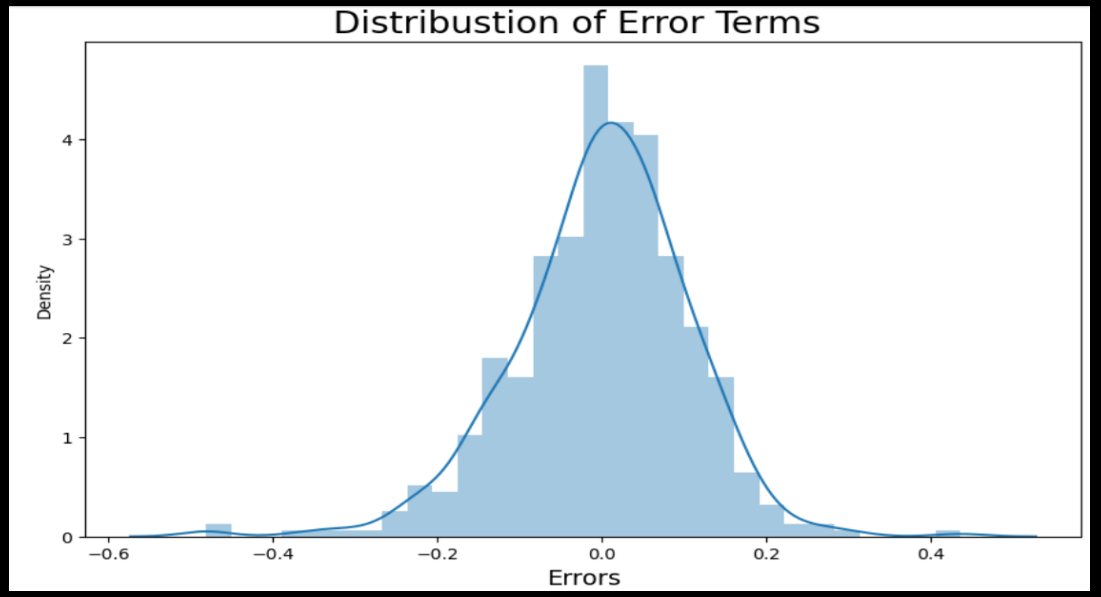
Question 4

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

Assumptions and validations for Linear Regression Model

Assumptions :

- Linear relationship.
- Multivariate normality.
- No or little multicollinearity.
- No auto-correlation.
- Homoscedasticity.



Validations

- Assumption can be validated by creating a distribution graph of Residual Term (Actual value of Target Variable – Predicted Value of Target Variable) . If Center of this graph is around ZERO then it is a well-balanced model.

- Creating scatter plot between X and Y

Question 5

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

Top 3 Contributing factors

1) Weather Situation - Negative correlation is indicated by

- Wind Speed
- Light Snow with rain
- Mist

Bike sharing demands decreases during the weather conditions.

2) Month - January has -Ve correlation

September has +Ve correlation

indicating bike sharing demands decreases in January and increases in September

3) Season - Winter , Spring and Summer has negative correlation indicating bike sharing demands decreases during these seasons

General Subjective Questions

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

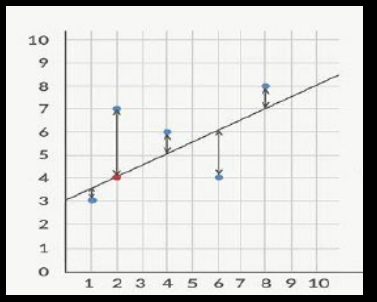
Regression Algorithms

Linear Regression Model :

It is type of predictive modelling technique which describe the relationship between the dependent (Target variable) and independent variables (predictors).

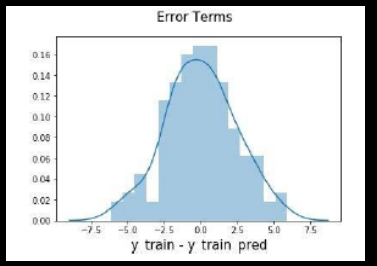
- Simple Linear Regression :
  - It is the simplest form of Linear regression, in which we try to find out linear relationship between one dependent and one independent variable.
- Multiple Linear Regression :
  - It is the complex form of Linear regression, in which we try to find out linear relationship between one dependent and multiple independent variables.

- Using Linear Regression Algorithm, we find the coefficient for independent variable in Best Fit Line which has minimum Residual $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$



- Gradient Decent Process:

To find best optimized coefficient for independent variables we use Gradient Decent Method.



Question -2

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

Anscombe’s Quartet

Anscombe’s Quartet :

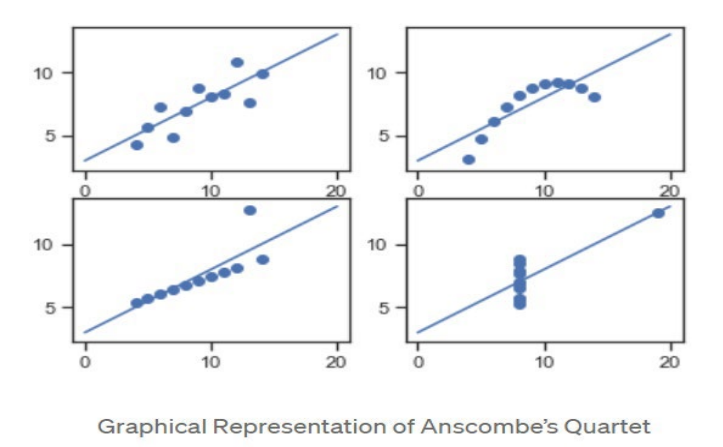
- It is the modal example to demonstrate the importance of data visualization which was developed by the statistician Francis Anscombein 1973 to signify both the importance of plotting data before analyzing it with statistical properties.

Example:

- Below data set gives and impression that if we do a statistical analysis between x1,y1 or x2,y2 or x3,y3 or x4,y4 then we will get similar kind of infer out of it.
- But if we follow the Anscombe’s quartet guideline and try to visualize these relationship then it will realize that it is not true.

x1	y1	x2	y2	x3	y3	x4	y4
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

If we to statistical analysis of four datasets
Average Value of x = 9
Average Value of y = 7.50
Variance of x = 11
Variance of y =4.12
Correlation Coefficient = 0.816
Linear Regression Equation : y = 0.5 x + 3



Graphical Representation of Anscombe’s Quartet

Quartet: It clearly shows that not all four dataset having linear relationship between x and y variable.
shows that visual analysis is very important while doing data analysis.
Source: https://builtin.com/data-science/anscombes-quartet

Question 3

What is Pearson’s R? (3 Marks)

Pearson’s R

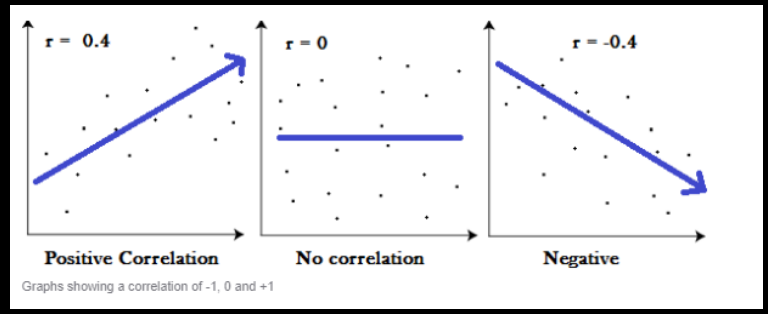
- Pearson's r is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative. The Pearson correlation coefficient, r, can take a range of values from +1 to -1.

- Formula to calculate Pearson R

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

Pearson’s R values range 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.



Source : https://www.statisticshowto.com/probability-and-statistics/correlation-coefficient-formula/

Question 4

- What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 Marks)

What is Scaling , Why It is performed & Difference between Normalization and standardization

What is Scaling :

Feature scaling (or scaling of variable values) is a technique to transform the feature values on a common scale. It is part of data preprocessing step in Machine Learning Model building.

Why is Scaling Required :

Scaling helps to bring values of all variable within a specified min/max range, which help algorithm (e.g. Gradient Decent) to design a Model where each variable have equal contribution.

Feature scaling become important when different variables/independent-variable have values which has a huge different in min/max values for column values.

Normalization	Standardization
Rescales values to a range between 0 and 1	Centers data around the mean and scales to a standard deviation of 1
Useful when the distribution of the data is unknown or not Gaussian	Useful when the distribution of the data is Gaussian or unknown
Sensitive to outliers	Less sensitive to outliers
Retains the shape of the original distribution	Changes the shape of the original distribution
May not preserve the relationships between the data points	Preserves the relationships between the data points
Equation: (x – min)/(max – min)	Equation: (x – mean)/standard deviation

Source:
https://www.analyticsvidhya.com/blog/2020/04/feature-scaling-machine-learning-normalization-standardization/

Question 5

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

VIF:

Feature scaling (or scaling of variable values) is used to find a correlation (Multicollinearity) between two independent variables

Value of VIF starts from 1 and has no upper limit.

If VIF is infinite that means there is a strong multicollinearity between those two independent variables, and it is not good for Model Building

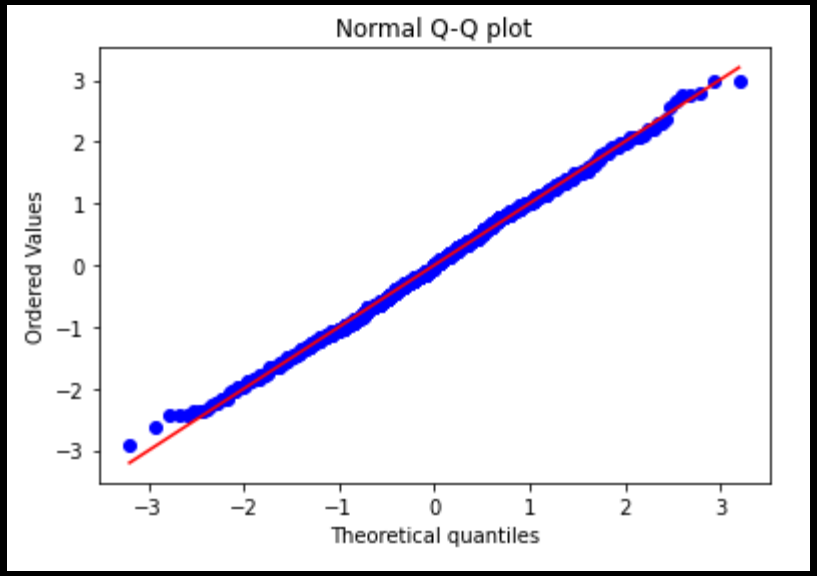
In the case of perfect correlation, we get R-squared (R2) =1, which lead to 1/ (1-R2) infinity. To solve this we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

Question 6

Explain the use and importance of a Q-Q plot in linear regression

QQ Plot

Q Plot: The quantile-quantile plot is a graphical method for determining whether two samples of data came from the same population or not.



We can find below data inferences from Q-Q plot

- Determine whether two samples are from the same population.
- Whether two samples have the same tail
- Whether two samples have the same distribution shape.
- Whether two samples have common location behavior.

Importance of Q-Q plot

- As Q-Q plot is like probability plot. So, while comparing two datasets the sample size need not to be equal.
- As we need to normalize the dataset, so we don’t need to care about the dimensions of values.

Source : https://www.geeksforgeeks.org/quantile-quantile-plots/