Course 3
Module 2
Linear Regression Assignment

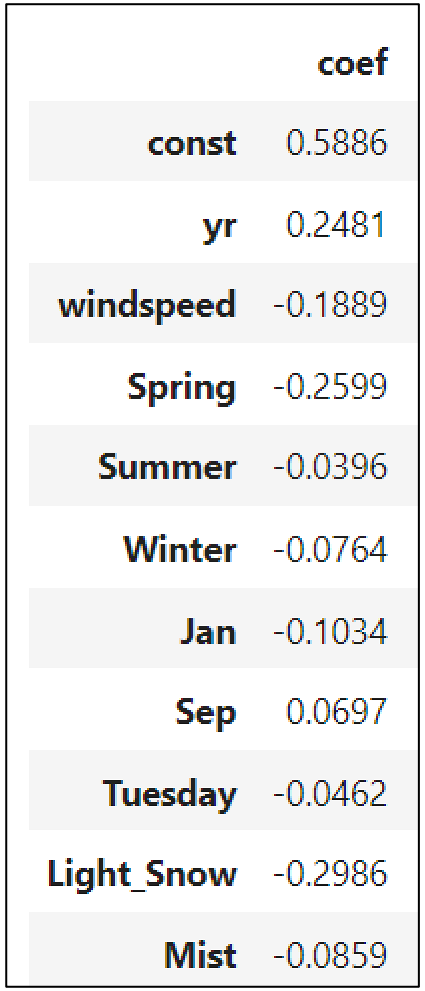
NAME: SANDEEP KUMAR

EMAIL: SANDEEP.NITA9@GMAIL.COM

Assignment Based Subjective Questions

Q1- From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variables?

Based on our Model Design we can derive below inferences about the categorical values



	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

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

Month : ‘January’ month is having negative coefficient; it indicates that Bike Sharing demand decreases during Jan month where ‘September’ month is having positive coefficient and have maximum demand of Bike sharing.

Weather Situation : ‘Light Snow with rain’ and ‘Mist’ weather having negative coefficient which indicates bike sharing demands decreases during these weather conditions.

Q2- Why it is important to use drop_first=True during dummy variable creation?

Machine learning model performs better when we use optimized and relevant number of independent variables during model building.

When we created dummy variables for Categorical Variable then removing one variable/column help us to keep less number of independent variable without loosing any important information.

Subject
English
Hindi
Math

We can understand this using below information
Assuming we are having one categorical variable which denotes Subject chosen by a Student. Student can have either 'English' or 'Hindi' or 'Math' as subject

While creating dummy variable for Subject we can have same information even in 2 dummy variables

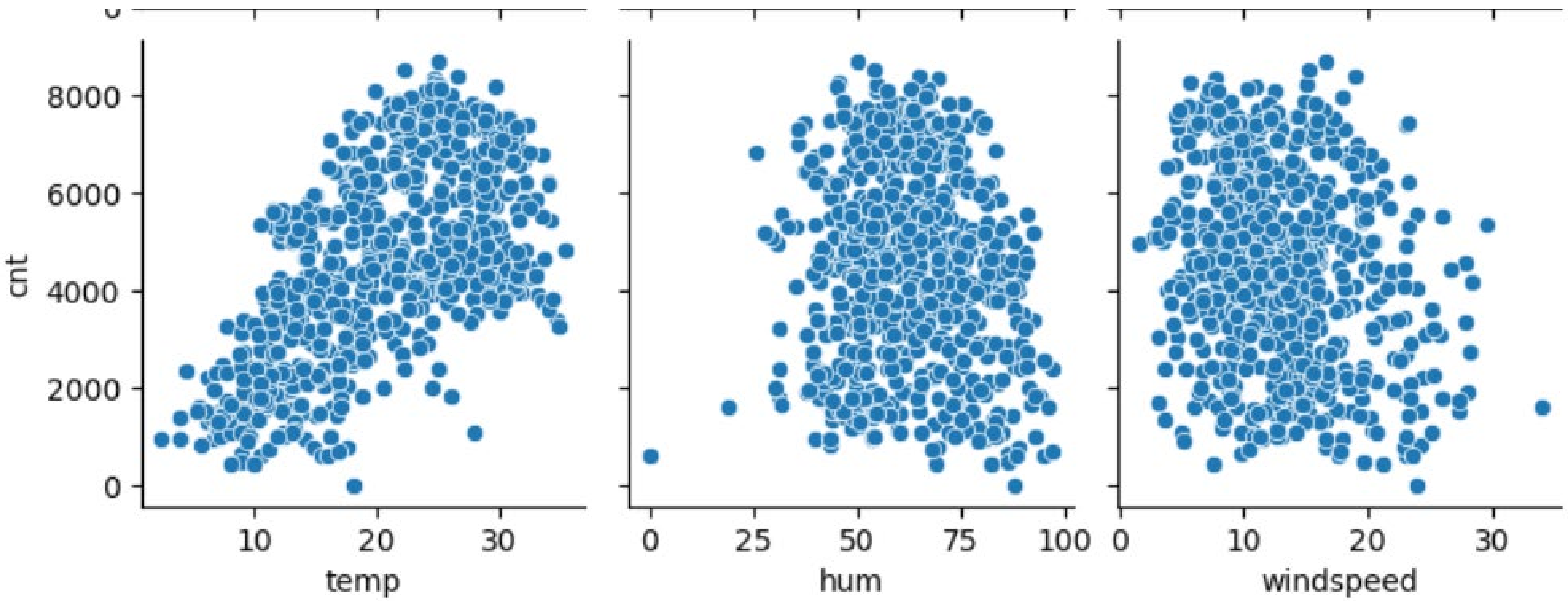
Subject_English	Subject_Hindi
1	0
0	1
0	0

First Entry show as Subject is 'English'
Second Entry shows Subject as 'Hindi'
Third Entry shoes Subject as 'Math'

For Machine learning algorithm these combination of '0' and '1' shows the various pattern. So by dropping one column we are reducing number of variable, keeping model simple and not loosing any information as well.

Q3- Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Pair plot generated for ‘Temperature’, ‘Humidity’, windspeed and ‘Count’



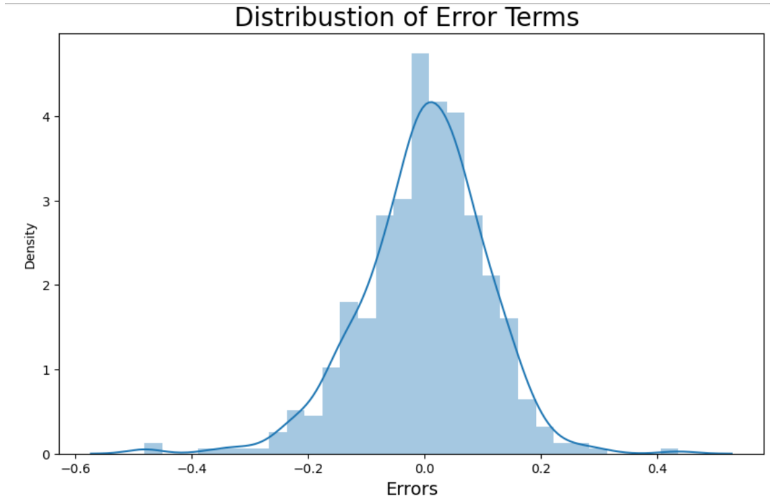
It clearly visible that ‘**Temp** (**Temperature**)’ column is having a strong correlation with Count (Bike Sharing Demand).

Above graph also showing a linear relationship with a positive slope between ‘Bike Sharing Demand (cnt) and Temp (Temperature)

Q4- How did you validate the assumptions of Linear Regression after building the Model on the training set?

Assumptions of Linear Regression

- Linear relationship between X and Y
- Error terms are normally distributed (not X, Y)
- Error terms are independent of each other
- Error terms have constant variance (homoscedasticity)



How we Validated:

1- By Creating a **Distribution graph of Residual Term** (Actual value of Target Variable – Predicted Value of Target Variable) is one of the strond validator that our assumptions are correct. If center of this graph is around ZERO then it considered a well-balanced model.

2- Also by creating scatter plot betweeb X and Y variable we can observe if there is any visible linear relationship exist or not.

Q5- Based on the final model, which are the top 3 feature contributing significantly towards explaining the demand of the shared bikes?

Final Model Equation

count= 0.5886+(0.2481*yr)-(0.1889*windspeed)-(0.2599*Spring)-(0.0396*Summer)-(0.0764*Winter)-(0.1034*Jan)+(0.0697*Sep)-(0.0462*Tuesday)-(0.2986*Light_Snow)-(0.0859*Mist)

Top 3 Factor that are affecting Bike Sharing Demand:

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

Month : 'January' month is having negative coefficient; it indicates that Bike Sharing demand decreases during Jan month where 'September' month is having positive coefficient and have maximum demand of Bike sharing.

Weather Situation : 'Light Snow with rain' and 'Mist' weather having negative coefficient which indicates bike sharing demands decreases during these weather conditions.

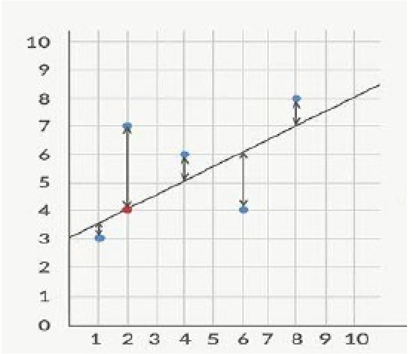
General Subjective Questions

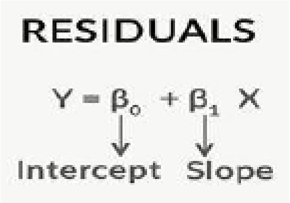
Q1- Explain Linear Regression Algorithm in Details.

Linear Regression Model : It is form 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.

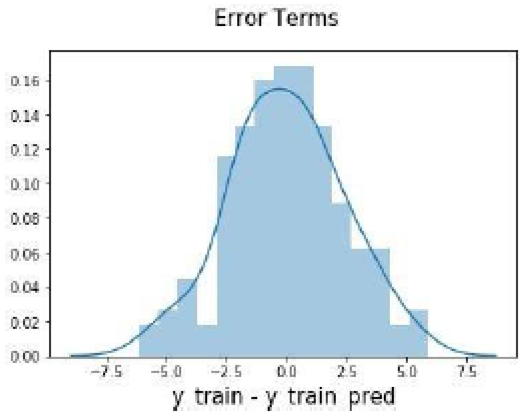



RESIDUALS

$$Y = \underbrace{\beta_0}_{\text{Intercept}} + \underbrace{\beta_1}_{\text{Slope}} X$$

Best Fit Line: Using Linear Regression Algorithm, we try to find the coefficient for independent variable in Best Fit Line which have minimum Residual (Error Term)

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$



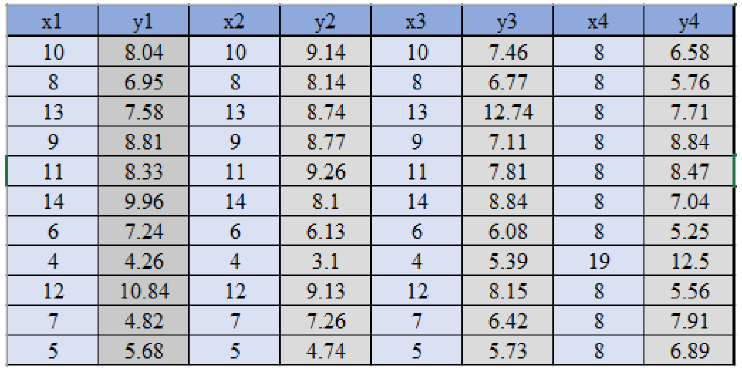
Gradient Decent Process: To find best optimized coefficient for independent variables we use Gradient Decent Method.

Q2- Explain Anscombe’s quartet in Details.

Anscombe’s Quartet : It is the modal example to demonstrate the importance of data visualization which was developed by the statistician Francis Anscombe in 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

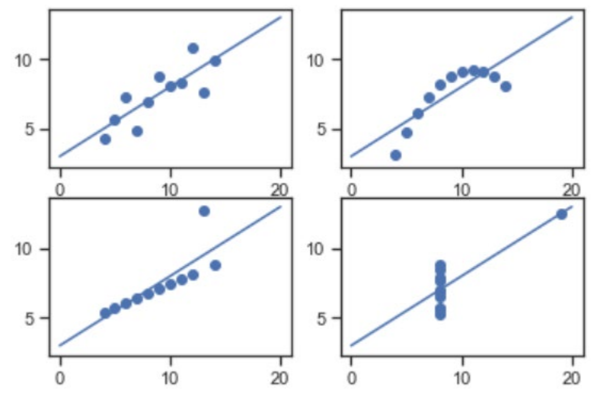
Average Value of y = 7.50

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

Visual Representation of Anscombe’s Quartet: It clearly shows that not all four dataset having linear relationship between x and y variable.

It shows that visual analysis is very important while doing data analysis.

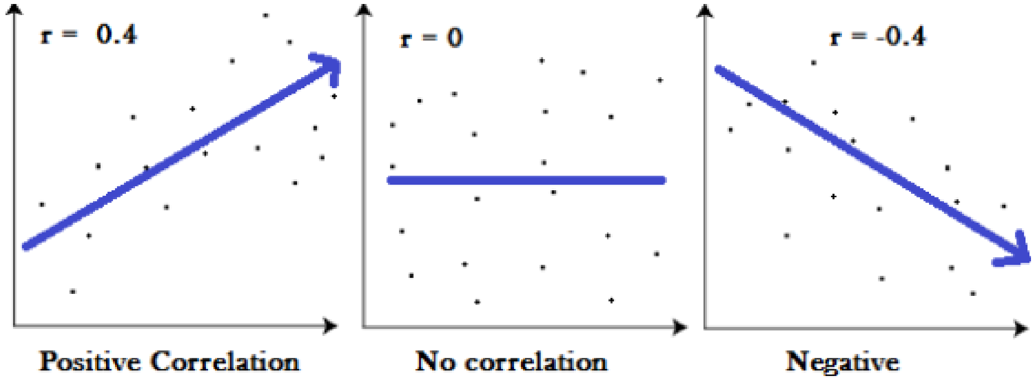
Credit : Source

Q3- What is Pearson’s R?

Pearson’s R: It is a common correlation coefficient which is used to find a correlation between two variables. It is also known as **Pearson’s correlation**.

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.



Positive Correlation

No correlation

Negative

Graphs showing a correlation of -1, 0 and +1

Formula to calculate R (coefficient r) value

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

Credit : Source

Q4- What is Scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

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..

What is Scaling : 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
Formula: (x – min)/(max – min)	Formula: (x – mean)/standard deviation

Credit : Source

Q5- 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 your Model Building

Q6- What is Q-Q plot? Explain the use of and importance of a Q-Q plot in linear regression.

Q-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

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

Credit : Source