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 variable?

- **Season** Fall is more optimal for Shared bike demand. Summer season is following Season fall with Median greater the 5000. Demand is low in spring. Fall is most optimal season to use Shared Bike.
- Demand of bike is high in **month** of sep and oct. We can say booking is high fall.
- Booking is low when **weather** is light snow and in clear weather demand is high. So we can say clear weather is optimal for ranted bike.
- Bike demand of shared bike is less in ***holidays***. Also demand on holiday is less in 2019 in comparison to 2018
- Booking is little high on **Weekdays**-Saturday and Friday otherwise almost same throughout the weekdays.
- As we can see bike demand is slightly high in **working day** as compared to non-working day. No significant change we can say.
- Demand for shared bike is increased in year 2019. Year could be a one feature which influence demand of bike because median for 2019 is quite high from 2018.

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

- Using get_dummies method we can create n dummies from variable having n levels.
- Actually we don't need n dummies we can represent n levels by n-1 level
- So use **drop_first=True** to drop first dummy variable as it is uniquely represent by one of the combination.
- If we keep it will create redundancy of one level.
- Dropping one level reduces Multicollinearity in the dataset, which is one of the Assumption of Multiple Linear Regression. E.g. : season -(1:spring, 2:summer, 3:fall, 4:winter) having 4 levels.

  Example- Season having 4 levels-Spring , Summer , Fall , winter. We can represent using 3 levels.

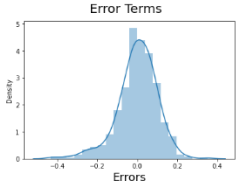
  i) 000 will correspond to Season1
  ii) 100 will correspond to Season2
  iii) 010 will correspond to Season3
  iv) 001 will correspond to Season4

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

Target variable 'cnt' has highest correlation with predictor variable temp

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

i. **Error Terms are normally distributed with mean zero (not x , y)** – this is validated using plot of residual(y_train-y_train_pred). If plot **is** normally distributed with mean 0 we can say assumption is respected.

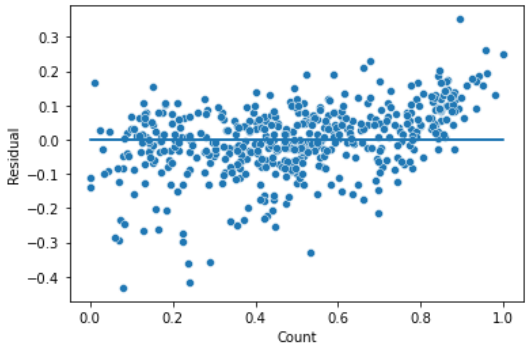


ii. **Multicollinearity check –** this is validated using VIF. If VIF of all the variables are below 5 then we can see there no multicollinearity existing between the predictor variables.

iii. **Error term are independent of each other**- This can be verified using durbin_watson test. If value is between 1.5 to 2.5,there is no corelation.

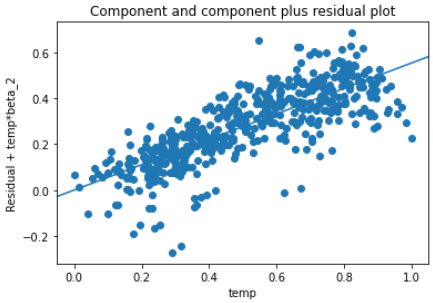
durbin_watson(lrm6.resid)

iv. **Error terms have constant variance(Homoscedasticity)-** this can be verified using residual vs fitted Plot.



v. **Linear relationship-** relations between **the independent and dependent variables** must be linear. This is done using below plot

sm.graphics.plot_ccpr(lrm7, 'temp')



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

- Temp – coefficient (0.5402) is high for temp variable. Positively corelate with target. If Temp increase de m and of shared bike increased by 0.5402 units
- Year - coefficient of year is 0.2548. Demand will increase as increase in yr.
- Weathersit-LightSnow this feature is negatively corelated (-0.2700) with target. If weather sit is Lightsno w then demand will decrease by 0.2700 units.

It I removed any of these feature R^2 drops significantly.

General Subjective Questions

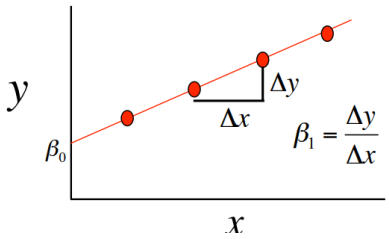
Q1 Explain the linear regression algorithm in detail.

- It's a Technique used for the modelling and analysis of numerical data
- Linear regression is a simple statistical regression technique used in predictive analysis that shows relatio nshi between continuous variables.
- Linear regression is called linear regression because it shows a linear relationship between the independ ent variable (x-axis) and the dependent variable (y-axis).
- When there is one independent variable then we call it as simple linear regression. If there is more the o ne independent variables such linear regression is called multiple liner regression.
- In layman terms to find best fit straight line having slop which describe the relationship within the varia bles. A regression line can be a Positive Linear Relationship or a Negative Linear Relationship.

To calculate best fit line below straight line equation is used-

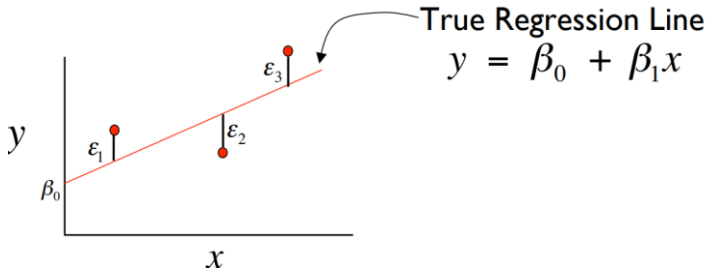
Simple linear regression equation -
Y=B0+B1*X *+E*

$$y = \beta_0 + \beta_1 x$$



Multiple linear regression equation:-
Y=B0+B1*X1 *+B2*X2+.....Bn*Xn+E*

where **B0**=intercept , **B1**=Coefficient of X's/Slope , **X** = Independent variable, **Y**= Output Var. E=random error

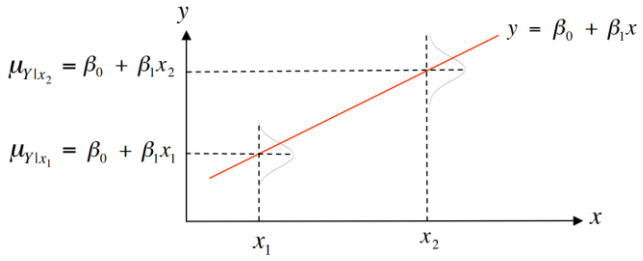


Error is nothing but difference between actual and predicted value

The goal of the linear regression algorithm is to get the best possible values for B0 and B1 to find the best fit line. To get best fit line we use cost Function to get value of beta's. Cost Function for linear regression is Root mean Squared error. The least error means the error between predicted values and actual values should be minimized.

There are four assumptions associated with a linear regression model:

Linearity: The relationship between independent variables dependent variable is linear.
Homoscedasticity: The variance of residuals should be equal.
Independence: error terms are independent of each other.
Normality: error terms are normally distributed. Which is explained by below graph.



The performance of the regression model can be evaluated by using various metrics like Fstat,Pvalue, R-squared etc.

R2=1-RSS/TSS

R^2 statistical measure that explain how close the data points are with fitted line. R-squared is always between 0 and 100%

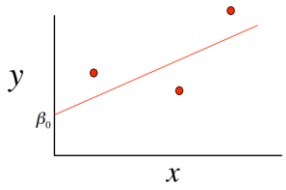
RSS-The residual sum of squares (RSS) measures the level of variance in the error term, or residuals, of a regressio n model. The smaller the residual sum of squares, the better your model fits your data

TSS-Total sum of square which is the sum of the squares of the prediction from the linear regression minus the m ean for that variable

In general, a model fits the data well if the differences between the observed values and the model's predicted val ues are small and unbiased.

Point estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ of and are obtained by the principle of least squares

$$f(\beta_0,\beta_1) = \sum_{i=1}^{n}[y_i - (\beta_0 + \beta_1 x_i)]^2$$



Predicted, or fitted, values are values of y predicted by the least squares regression line obtained by plugging in x 1,x2,…,xn into the estimated regression line. Technically, ordinary least squares (OLS) regression minimizes the su m of the squared residuals.
P va

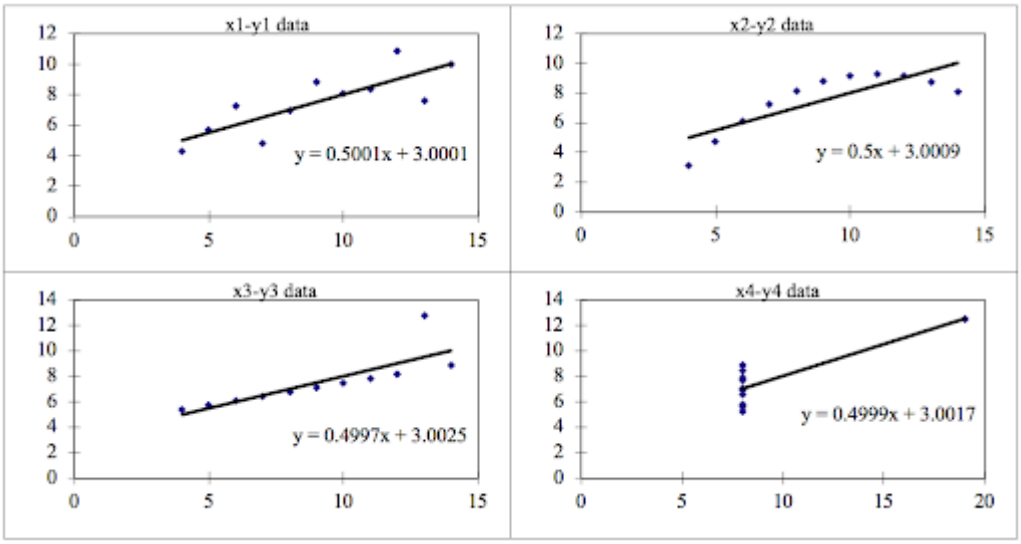
Q2 .Explain the Anscombe's quartet in detail

- Anscombe's Quartet is basically a modal example to demonstrate the importance of data visualization before applying various algorithms out there to build models out of them which suggests that the data f eatures must be plotted in order to see the distribution of the samples that can help to identify the vario us anomalies present in the data like outliers, diversity of the data, linear separability of the data, etc.
- Anscombe's Quartet was developed by statistician Francis Anscombe.
- It comprises four datasets each containing eleven (x,y) pairs.
- 4 datasets share same statistical information that involves **variance**, and **mean** of all x,y points in all four datasets
- But things change completely, when they are graphed. Each graph tells a different story irrespective of t heir similar **summary statistics.**

Below is the data set and graph

Anscombe's Data								
Observation	x1	y1	x2	y2	x3	y3	x4	y4
1	10	8.04	10	9.14	10	7.46	8	6.58
2	8	6.95	8	8.14	8	6.77	8	5.76
3	13	7.58	13	8.74	13	12.74	8	7.71
4	9	8.81	9	8.77	9	7.11	8	8.84
5	11	8.33	11	9.26	11	7.81	8	8.47
6	14	9.96	14	8.1	14	8.84	8	7.04
7	6	7.24	6	6.13	6	6.08	8	5.25
8	4	4.26	4	3.1	4	5.39	19	12.5
9	12	10.84	12	9.13	12	8.15	8	5.56
10	7	4.82	7	7.26	7	6.42	8	7.91
11	5	5.68	5	4.74	5	5.73	8	6.89

	I		II		III		IV	
	x	y	x	y	x	y	x	y
	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
SUM	99,00	82,51	99,00	82,51	99,00	82,50	99,00	82,51
AVG	9,00	7,50	9,00	7,50	9,00	7,50	9,00	7,50
STDEV	3,32	2,03	3,32	2,03	3,32	2,03	3,32	2,03



We can see the statistics summary of all the dataset is same also when we see plot linear line are same but each dataset behaving differently.

The four datasets can be described as:

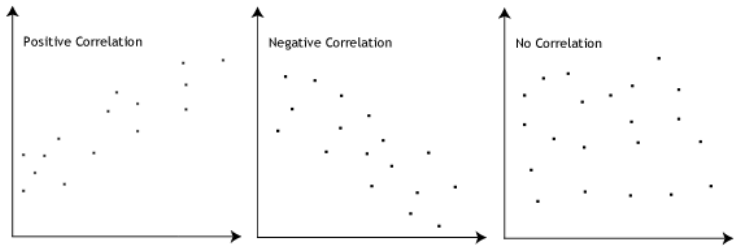
a) Dataset 1: fits the linear regression model pretty well.
b) Dataset 2: Data is non linear so could not fit linear regression model on the data quite well
c) Dataset 3: it shows the outliers involved in the dataset which cannot be handled by linear regression model.
d) Dataset 4: shows the outliers involved in the dataset which cannot be handled by linear regression model.

Q3 What is Pearson's R?

- Pearson correlation coefficient is defined as measurement of the strength of the relationship between two variables and their association with each other.
- In other word Pearson's R calculates the effect of change in one variable when the other variable changes. It basically seeks to draw a line through the data of two variables to show their relationship.
- This linear relationship can be positive or negative.
- Mathematically, it is denoted as the covariance of the two independent variables divided by the product of their standard deviations.

The Pearson's correlation coefficient varies between -1 and +1 where:

- r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)
- r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)
- r = 0 means there is no linear association
- r > 0 < 5 means there is a weak association
- r > 5 < 8 means there is a moderate association
- r > 8 means there is a strong association



Pearson r Formula

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

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

What is scaling-

- It is a data Pre-Processing step which is applied on continuous variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.
- Scaling means to change the range of values but without changing the shape of distribution. Range is often set to 0 to 1
- scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

Need of scaling –

- If continuous variables in dataset having different magnitude, unit and range. If we build the model without scaling the data it will consider only magnitude not units which result in incorrect modelling.
- Means to say it magnitude is high coefficient will be high which leads to incorrect modelling.

- To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

Two methods of scaling-

Standardization

- Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).
- Or we can say transforms the mean of the data to be 0 & its variance to be 1. As the data value tends towards infinity, variance of the data tends towards 1.

Standardization x= x-mean(x)/sd(x)

Normalization

- It brings all of the data in the range of 0 and 1. so that the data falls under a narrow range which helps the model to learn.

Normalization x=x-Xmean/Xmax-Xmin

Q5 You might have observed that sometimes the value of VIF is infinite. Why does this happen?

- VIF measure the corelation /Multicollinearity between the independent variables/predictor variables. If value is high means having high corelation between the variables and it is difficult to predict the contribution of predictors to a model
- Formula:-

VIFi= 1/1- Ri^2

- Based on above formula when the value of R=1 then VIF would be infinite. Means there is correlation in variables.
- If there is perfect correlation, then VIF = infinity.
- In this scenario we need to drop one of the corelated variable.

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

- Q-Q plot is a scatter plot created by plotting 2 different quantiles against each other.
- It used to graphically analyze and compare two probability distributions by plotting their quantiles against each other. If both the distributions are same then points will perfectly lie on straight line y-x.
- It is used to find if the points lie approximately on the line. If they don't, it means, our residuals aren't Gaussian (Normal) and thus, our errors are also not Gaussian.
- The first quantile is that of the variable you are testing the hypothesis for and the second one is the actual distribution you are testing it against
- If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight
- Importance of Q-Q plot: Below are the points:

i) The sample sizes do not need to be equal.

ii) Many distributional aspects can be simultaneously tested. For example, shifts in location, shifts in scale, changes in symmetry, and the presence of outliers.

iii) The q-q plot can provide more insight into the nature of the difference than analytical methods.

-