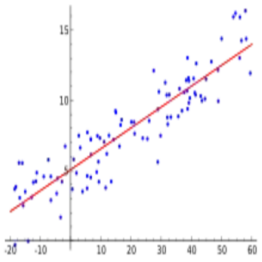
General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression is a method of finding the best straight line fitting to the given data, meaning finding the best linear relationship between the independent and dependent variables(target variable on which we will predict).

Technically linear regression is a machine learning algorithm that finds the best linear-fit relationship on any given data, between independent and dependent variables.

The diagram below explains the point mentioned above i.e the straight line that fits the given data ,values in y are dependent(target variable) and values in x are independent variables



Example :

Simple example would be salary and no years of experience ,they are linearly dependent(i.e salary increases with experience).

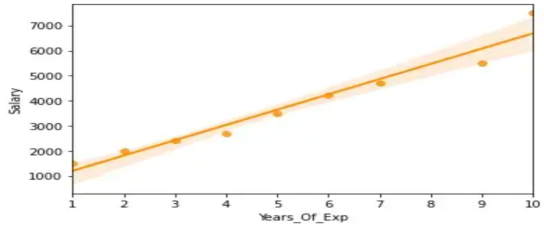
Terminologies in Linear Regression

1. Regression Line

Each value of x that determines value in y and straight line is represented using the standard equation **$Y = \beta_0 + \beta_1.X$ (i.e y=c + mx)**

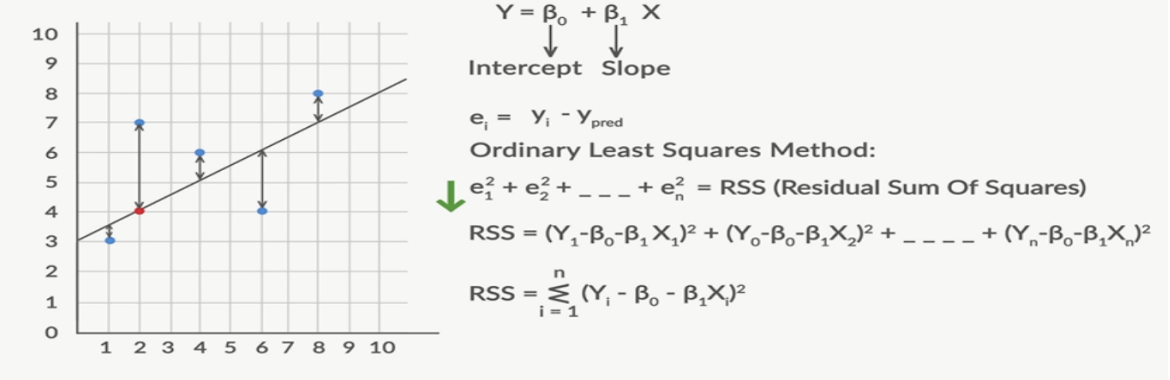
- In a 2D plane , 'β_1' is the slope of the line(change in Y divided by change in X).
- Y is the output(In this example salary) which is determined by input X(no of years). How much of an impact X(years of experience) has on Y (salary) is determined by "β_1".
- 'β_0' is the constant (value of Y when X is zero).

For the salary vs years of experience below diagram represents regression line



2. Best Fit Line

The best-fit line is found by minimising the expression of RSS (Residual Sum of Squares) which is equal to the sum of squares of the residual for each data point in the plot. Residuals for any data point is found by subtracting predicted value of dependent variable from actual value of dependent variable



Type of Linear Regression

Simple Linear Regression

The basic type of regression model is the simple linear regression which explains the relationship between a dependent variable and one independent variable using a straight line. The straight line is plotted on the scatter plot of these two points.

Multiple linear regression

It is a model or a technique to understand the relationship between one dependent variable and several independent variables (explanatory variables).

Equation is represented in a plane as **$Y = \beta_0 + \beta_1.X1 + \beta_2.X_2 .. + \beta_n.X_n$** ($\beta_1..\beta_n$ are the slopes or weights of independent variables)

Measuring Accuracy/Strength of Linear Regression Model

R^2 or Coefficient of Determination

Formula **$R^2 = 1-(RSS/TSS)$** , RSS & TSS are defined as follows

RSS (Residual Sum of Squares): Defined as the total sum of error across the whole sample. It is the measure of the difference between the expected and the actual output.

Formula RSS = $\sum^{n}_{i=1} (y_i - (\alpha+\beta x_i))^2$

TSS (Total sum of squares): It is the sum of errors of the data points from the mean of the response variable. Mathematically, TSS = $\sum^{n}_{i=1} (y_i - y_{mean})^2$

2. Explain the Anscombe's quartet in detail.

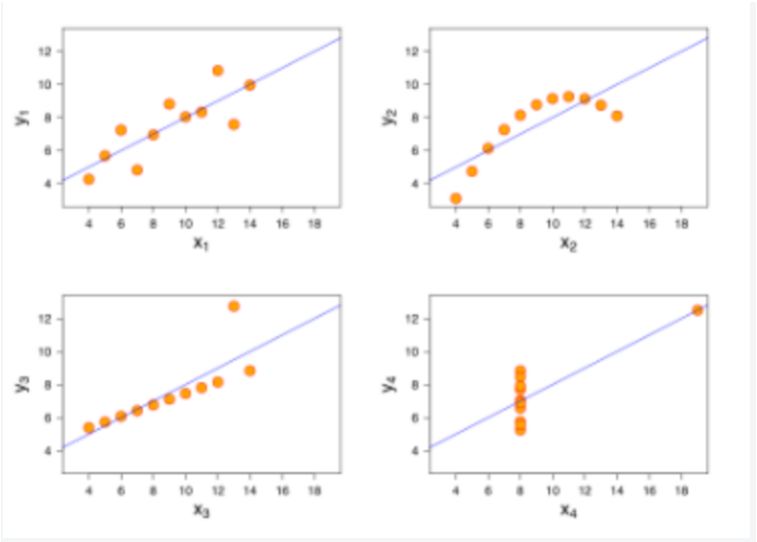
Anscombe's quartet comprises four data sets and each data-set consists of eleven (x,y) points that have nearly identical simple descriptive statistics(same mean, variance, standard deviation etc) but different graphical representation.

They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data when analysing it, and the effect of outliers,linear separability and other influential observations on statistical properties.

Dataset representing the Anscombe's quartet

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
				Summary Statistics							
N	11	11		11	11		11	11		11	11
mean	9.00	7.50		9.00	7.500909		9.00	7.50		9.00	7.50
SD	3.16	1.94		3.16	1.94		3.16	1.94		3.16	1.94
r	0.82			0.82			0.82			0.82	

From the above tabular column all the four datasets have the same mean ,variance and standard deviation. Even though the stats are same when plotted they look as follows



- Dataset 1: this fits the linear regression model well as the relationship is linear.
- Dataset 2: Linear regression model cannot be used on the data as the data is non-linear(quadratic).
- Dataset 3 & 4 : Depicts the linear regression model's sensitivity to outliers

We should never ever run a regression without having a visual look at our data as it is effective only if the data is linearly related.

Interesting quote from Francis Anscombe : "Intended to counter the impression among statisticians that "numerical calculations are exact, but graphs are rough"

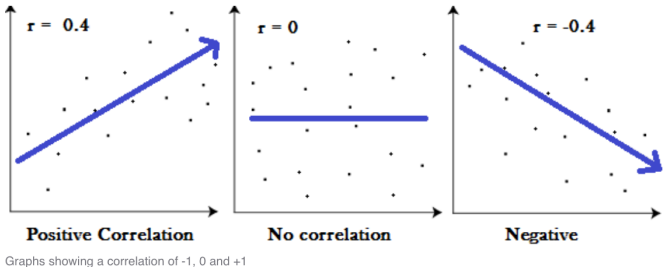
3. What is Pearson's R?

It is a statistic that measures the linear correlation or in simple terms relationship between two variables.It has a numerical value that lies between -1.0 and +1.0.

However, it cannot capture nonlinear relationships between two variables and cannot differentiate between dependent and independent variables.

Interpretation of pearson coefficient

- 1 indicates a strong positive relationship(value between 0 to 1).
- -1 indicates a strong negative relationship (value between -1 to 0).
- A result of zero indicates no relationship at all.



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

Pearson's correlation coe√fficient is the covariance of the two variables divided by the product of their standard deviations.

Formula to calculate pearson R coefficient

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

Example : Let's take a simple example of computing pearson coefficient between age and weight of 6 samples

SNO	AGE(X)	WEIGHT(Y)	XY	X^2	Y^2
1	21	70	1470	441	4900
2	25	60	1500	625	3600
3	31	55	1705	961	3025
4	38	80	3040	1444	6400
5	40	78	3120	1600	6084
6	47	66	3102	2209	4356
∑	202	409	13937	7280	28365

By using above formula

$$R = \frac{6*(13937) - (202*409)}{\text{sqrt}((6*7280) - 202*202) * (6*28365) - (408*408)} = 1004/2892.45 = 0.347 \sim 0.35$$

Value of pearson coefficient for the above dataset is 0.35

The positive value of Pearson's coefficient implies that if we change either of these variables, there will be a positive effect on the other.

Assumptions when using pearson coefficient

- Relationship between variables should be linear
- Both the variables should be normally distributed
- Each variable should be continuous i.e interval or ratio(ex age ,weight,income)
- Every assumption of independent variable should have corresponding observation of dependant variable
- Homoscedascity -Error term (that is, the "noise" in the relationship between the independent variables and the dependent variable) is the same across all values of the independent variables

4. What is scaling? Why is scaling performed? What is the difference between normalised scaling and standardised scaling?

Feature Scaling is a technique to standardise the independent features that have highly varying magnitudes or values or units present in the data in a fixed range. It is done as part of preprocessing . It is important to have everything on the same scale for the model to be easily interpretable

Feature Scaling needs to be done for the following reasons

- If the datasets have features with varying magnitudes then features with high magnitude will have low coefficients and lower magnitude will have higher coefficients. So there is a possibility that we will interpret as lesser coefficients mean lesser significance
- Another big advantage with feature scaling is that the algorithm will converge fast and more performant as the data is shrinked to a range.

Normalization/Min-Max Scaling:
It brings all of the data in the range of 0 and 1 and computed as follows

$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

Standardization Scaling:
Standardization brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).

$$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$$

Example
Let say take simple dataset as following

CAR	AGE_OF_CAR(Years)	DISTANCE TRAVELLED
FORD IKON	10	10000
SWIFT	12	20000
HONDA CITY	8	15000
TRAVERA	15	30000

In the above dataset Age of car is ranging from 8 to 15 years, whereas Distance Travelled is from 10000km to 50000km. When we compare both the ranges, they are at very long distances from each other. As mentioned above, interpretation of coefficients could be

misleading and this pushes us to feature scaling for better interpretation and faster convergence.

After applying normalization and standardization data looks as below

CAR	AGE_OF_CAR(Years)	DISTANCE TRAVELLED	DISTANCE_TRAVELLED_NORMALIZED	DISTANCE_TRAVELLED_STANDARDIZED
FORD IKON	10	10000	0.0	-1.183216
SWIFT	12	20000	0.50	0.169031
HONDA CITY	8	15000	0.25	-0.507093
TRAVERA	15	30000	1	1.521278

Code used to generate above scaling

```
import pandas as pd
from sklearn.preprocessing import MinMaxScaler,StandardScaler
scaler = MinMaxScaler() #Normalization
sscaler= StandardScaler() #Standardization

data = [['FORD IKON','10', 10000], ['SWIFT','12', 20000], ['HONDA
CITY','8', 15000],['TRAVER','15', 30000]]
df = pd.DataFrame(data, columns=['CAR','AGE_OF_CAR(Years)','DISTANCE
TRAVELLED'])


df['DISTANCE_TRAVELLED_NORMALIZED'] = scaler.fit_transform(df[['DISTANCE
TRAVELLED']])
df['DISTANCE_TRAVELLED_STANDARDIZED'] =
sscaler.fit_transform(df[['DISTANCE TRAVELLED']])
df
```

The advantage of Standardisation over the other is that it doesn't compress the data between a particular range as in Min-Max scaling. This is useful, especially if there are outliers.

This doesn't mean that Min-Max scaling is not useful and it is widely used in image processing where pixels are within 0-255 and also in neural networks.

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

$VIF = 1/(1-R^2)$
When some of the independent variables are perfectly correlated (positively or negatively) with each other then we get $R^2=1$ causing VIF to reach infinity

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

This indicates the presence of multicollinearity, it doesn't affect the prediction of the model but affects the interpretation of the model which can affect the business model.

Example let say we have predictors X1 & X2 and when they are same following can be expressed by linear combination of others leading to infinite VIF

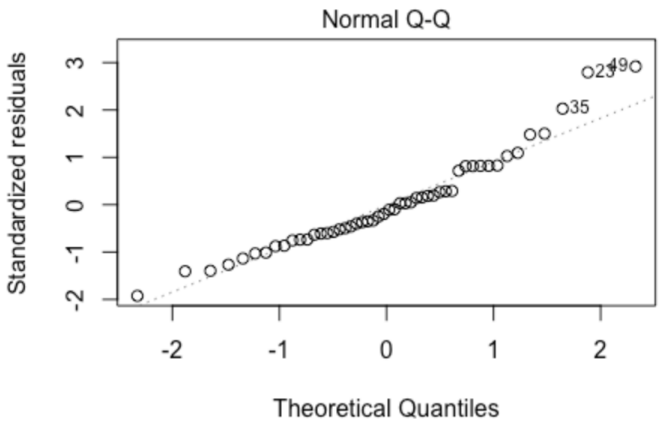
- 10X1
- 8X1+2X2
- 7X1+3X2

So to avoid this typical recommendation is to drop variables that VIF > 5

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

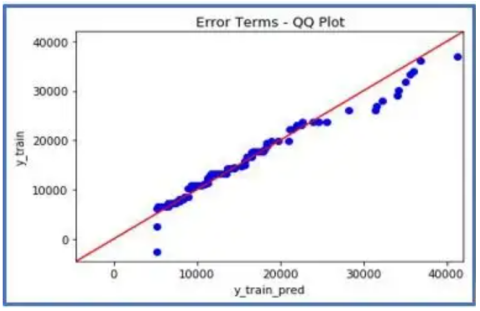
Q-Q plots are also known as Quantile-Quantile plots. quantiles of a sample distribution are plotted against quantiles of a theoretical distribution.This helps us determine if a dataset follows any particular type of probability distribution like normal, uniform, exponential,i.e if two sets of data come from the same distribution.

If the datasets we are comparing are of the same type of distribution type, we would get a roughly straight line on the data points. Here is an example of normal distribution.

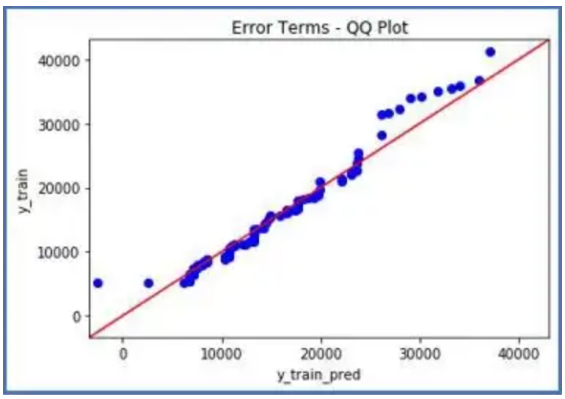


In linear regression QQ plots are used in

1. Verifying the Normal distributions on residuals , equal variance among groups
2. Skewness on the dataset
3. Identifying the outliers



The above diagram illustrates point 1 that is on plotting training vs predicted dataset it follows a normal distributions (most of the points lie on a straight line)



The above QQ plot indicates that there are some outliers on the right most side.

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?

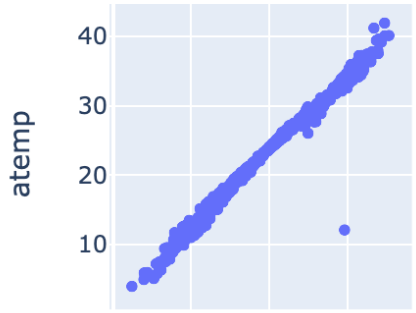
- Average usage/demand increases when the weather situation is clear and it decreases whenever there is light snow or rain or mist.
- Demand is slightly higher during working day
- Demand is higher during fall followed by summer and drastically reduces during fall season

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

The drop_first parameter specifies whether or not to drop the first category of the categorical variable that is being encoded.
If we have K categories K-1 dummy variables is enough to represent all the categories
By default, this is set to drop_first = False. This will cause get_dummies to create one dummy variable for every level of the input categorical variable.If set to true this will drop first category.

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

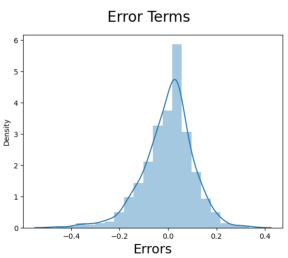
Temp and atemp are highly correlated based on the pair plot and correlation is 0.99 from heat map



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

Model has been validated on the following assumptions on residual data

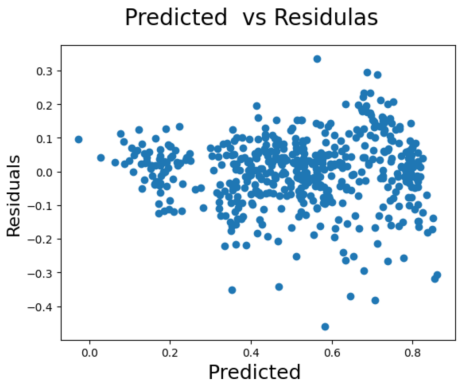
- Residuals are normally distributed



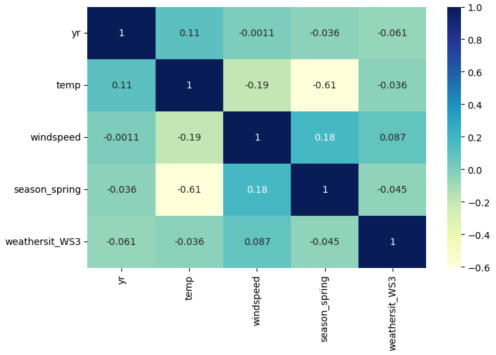
- Mean is zero or very close to zero

  Mean value is 2.185615522987563e-16

- Homoscedasticity (Error terms have equal variances or no specific patterns)



- Collinearity between features used for training are less



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

Below is the weightage on final 5 features used in the model

	coef
const	0.2772
yr	0.2371
temp	0.3857
windspeed	-0.1502
season_spring	-0.1476
weathersit_WS3	-0.2433

Top 3 features are

1. Temperature - Positive effect on target variable with others constant.
2. Yr - Positive correlation and also per EDA demand increased in 2019 compared to 2018.
3. Season Spring - Negative correlation and also EDA suggested demand is low during spring seasons.