Linear Regression Subjective Questions

Question 1: From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

ANSWER 1: Categorical variables are treated and visualise using bar and box plots and some of the inferences are –

- ✓ Spring season attract less bookings while fall attracts more and trends is same in years but count has increase in subsequent year(i.e. from 2018-2019).
- ✓ May to October sees more booking and same trend persist from 2018 to 2019.
- ✓ Clear weather sees more booking and again booking is more in 2019 compared to previous year, i.e 2018.
- ✓ Weekday like Thu, Fir and weekend have more number of bookings as compared to the first three days of week.

- ✓ On Holidays there is less booking this might be because people spend time with family friends or go on vacation
- ✓ Booking on either working or no working is same but number of bookings increases from 2018 to 2019.
- ✓ 2019 shows more booking compared to the previous year.

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

ANSWER 2: drop_first = True is used in get dummies in pandas which Convert categorical variable into dummy/indicator variables.

drop_first = bool (by default its False) using True can get n-1 dummies out of n categorical levels

and this vital to use as this reduce the extra column created during dummy variable creation. And eventually help in understanding correlation better between dummy variable

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

ANSWER 3: Pairplot shows that 'temp' independent variable has strong correlation with target variable

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

ANSWER 4: Validation of assumption in regression model is done based on following facts-

- ✓ Checked the linear relationship between independent and predicted (dependent) variables
- ✓ Error terms are normally distributed and have mean zero, this is done by plotting error terms histograms
- ✓ Error terms or residuals are independent of each other, this is done checking whether there is any pattern observe in residuals

- ✓ Residuals(Error terms) should have constant variance (homoscedasticity) i.e. in other words the variance should not follow any particular pattern (or should not decrease or increase). This is validated by plotting predicted vs actual numbers
- ✓ The independent variables should have insignificant multicollinearity, this is validated by analysing VIF and Correlation coefficients.

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

ANSWER 5:

- ✓ Temp
- ✓ season winter
- ✓ mnth_sep

General Subjective Questions

Question 1: Explain the linear regression algorithm in detail.

ANSWER 1: Linear regression is a very basic and fundamental algorithm of machine learning where we predict the behaviour of based on some independent variables. Linear suggests dependent and independent variables are linearly correlated.

As an example, lets say an automobile manufacturer from developing country wish to make an entry in third world country outlashing current market players from developed countries. Hence automobile manufacturer wants to understand the factors affecting the pricing of cars in third world country, since those may be very different from the their own market. The company wants to know: Which variables are important in predicting the price of a car and How well those variables correlated the price of a car.

The manufacturer need to model price of cars as a function of independent variables Which can be use by management to understand dynamics of market.

In these cases, independent variables the value will be linearly upward that means whenever X is increasing Y is also increasing or vice versa that means they have a correlation or there will be a linear downward relationship.

Linear regression is used to predict a quantitative response Y from the predictor variable X.

Mathematically, $Y = bx + a$

Where,

b = Slope of the line.

a = y-intercept of the line.

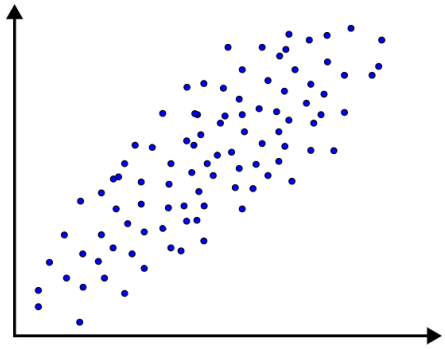
x = Independent variable from dataset

y = Dependent variable from dataset

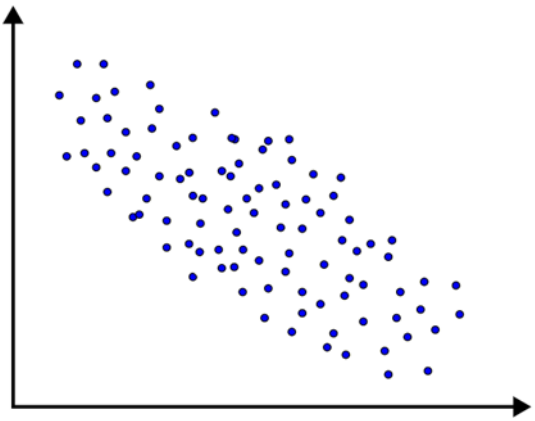
$$b\,(slope) = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$$

$$a\,(intercept) = \frac{n\sum y - b(\sum x)}{n}$$

Positive correlation



Negative Correlation



When training the model the best fit line is obtained by finding best a and b coefficient. Once we find the best a and b values, we can use model for prediction i.e. it will predict the value of y for the input value of x. **The best values of a and b are obtained by defining Cost Function(residual i.e. error difference between predicted and actual value) and minimising the same.** Cost function(J) of Linear Regression is the Root Mean Squared Error (RMSE) between predicted y value (pred) and true y value (y).

Linear regression is of the following two types

- ✓ Simple Linear Regression
- ✓ Multiple Linear Regression

There four main assumption in linear regression

1. **Linear relationship**: There is a linear relationship between the independent variable, x, and the dependent variable, y.

2. **Independence**: The residuals are independent. i.e. there is no correlation between consecutive residuals in time series data.

3. **Homoscedasticity**: The residuals have constant variance at every level of x.

4. **Normality**: The residuals of the model are normally distributed.

Question 2: Explain the Anscombe's quartet in detail.

ANSWER 2: Anscombe's quartet talks about different sets which have nearly identical standard deviation and mean based but when visualized graphically it looks completely different distribution.

This concept was devised by a statistician Francis John “Frank” Anscombe who noticed that 4 sets of 11 data-points have similar descriptive statistics viz. mean, standard deviation, and correlation between x and y.

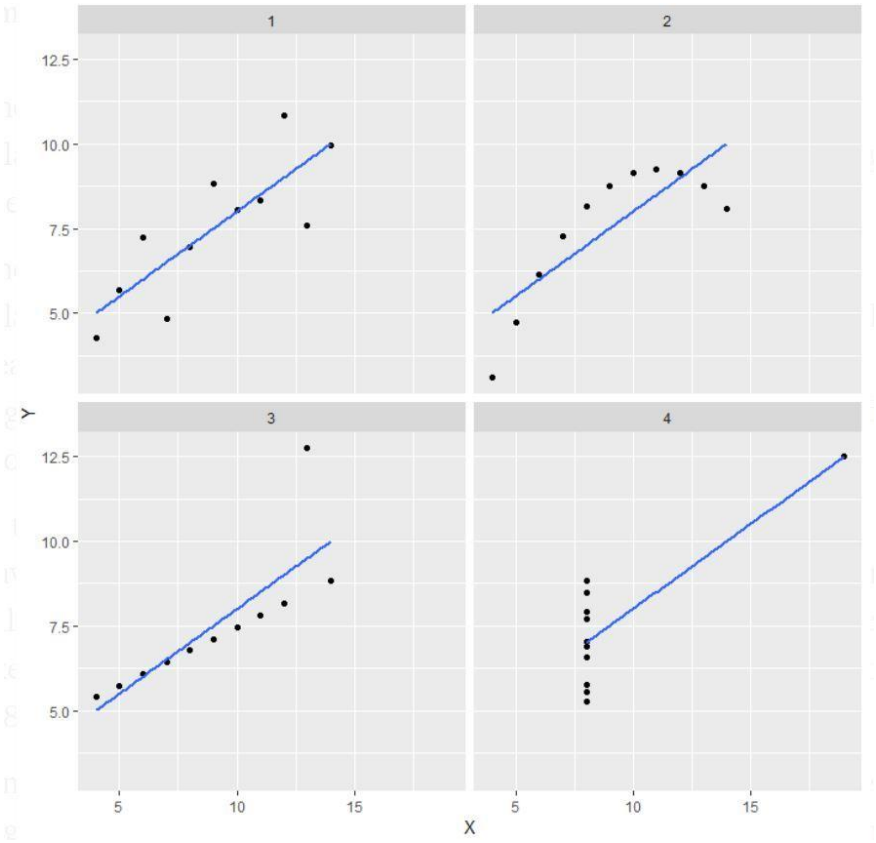
I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

Standard deviation and Mean

Summary

Set	mean(X)	sd(X)	mean(Y)	sd(Y)	cor(X,Y)
1	9	3.32	7.5	2.03	0.816
2	9	3.32	7.5	2.03	0.816
3	9	3.32	7.5	2.03	0.816
4	9	3.32	7.5	2.03	0.817

Correlation



The above four datasets show:

Dataset 1: This fits the linear regression model.

Dataset 2: This does not fit linear regression model on the data as the data is non-linear.

Dataset 3: this shows that linear relation fits well but outliers involved in the dataset is not be handled by linear regression model.

Dataset 4: shows the outliers handled by model and produces high correlation

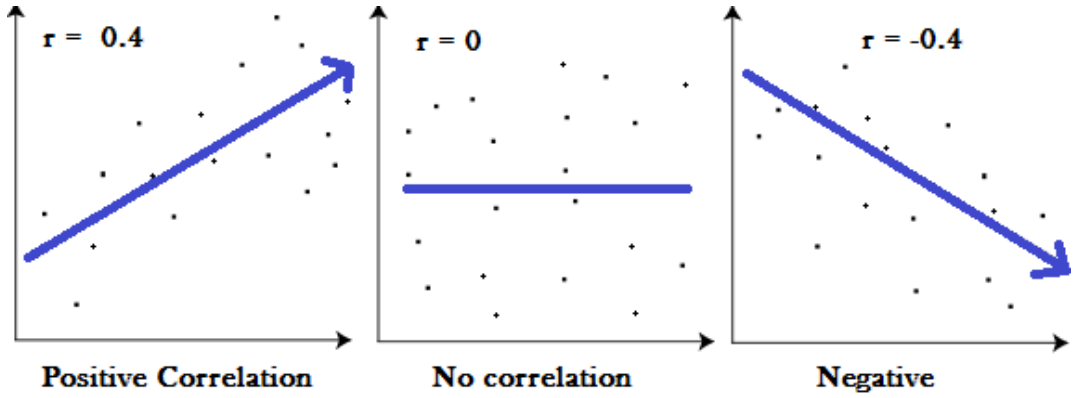
Above shows that it is imperative that visualise the data graphically before starting to analyse it directly, and descriptive statistics is not just enough

Question 3: What is Pearson’s R?

ANSWER 3:

Pearson’s Correlation Coefficient is named after Karl Pearson. In Statistics, the Pearson's Correlation Coefficient is also called to as Pearson's r. This is used to find out correlation between two variables and its value lies in between -1.0 and +1.0.

- ✓ 1 indicates a strong positive relationship between variables
- ✓ -1 indicates a strong negative relationship between variables
- ✓ A result of zero indicates no relationship at all between variables



Pearson's correlation coefficient is given by

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

Where:

Cov(X,Y) is covariance

σ_x standard deviation of X

σ_y standard deviation of Y

But there few problems associated with the Pearson's Coefficient : Pearson coefficient cannot capture nonlinear relationships between two variables and cannot differentiate between dependent and independent variables.

Real life example:

This R value can be used to predict the if there is a relationship between how genetically modified brinjal has any correlation with its productions.

Question 4: What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

ANSWER 4:

Many a times data set under analysis have highly varying in nature (viz. magnitudes, units and range). Before performing analysis if this data is not scaled then algorithm considered magnitude of variable but not unit which leads to incorrect modelling. In order to cater this, scaling is done to bring all the variables to the same level of magnitude. If not scale, the feature with a higher value range starts dominating when calculating distances. This normalization or standardization helps in speeding up the calculations in an algorithm.

Normalization is rescaling the values into a range of [0,1] while Standardization is rescaling data to have a mean of 0 and a standard deviation of 1 (unit variance).

Normalization/Min-Max Scaling:

It brings all of the data in the range of 0 and 1

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

Standardization Scaling:

Standardization alters the values by their Z score. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).

One more important thing to notice is that *scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.*

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

ANSWER 5:

VIF is defined by

VIF = 1 /(1-R^2)

It can be infer from above formulae that if VIF = infinity, then there is perfect correlation. Int his case R2 will be 1 which indicates variation

of a dependent variable is explained properly by the independent variable(s) in a regression model.

Then, we need to remove one of the variables from the dataset which is causing this perfect multicollinearity. In other words, only one variable can be expressed as a linear combination of another variable.

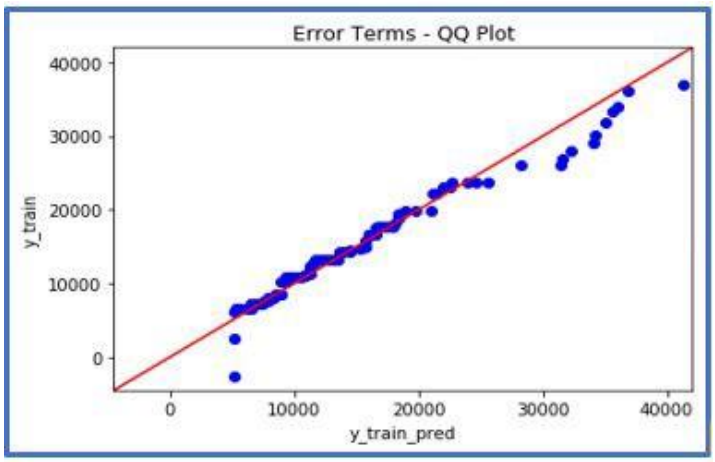
Question 6: What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

ANSWER 6:

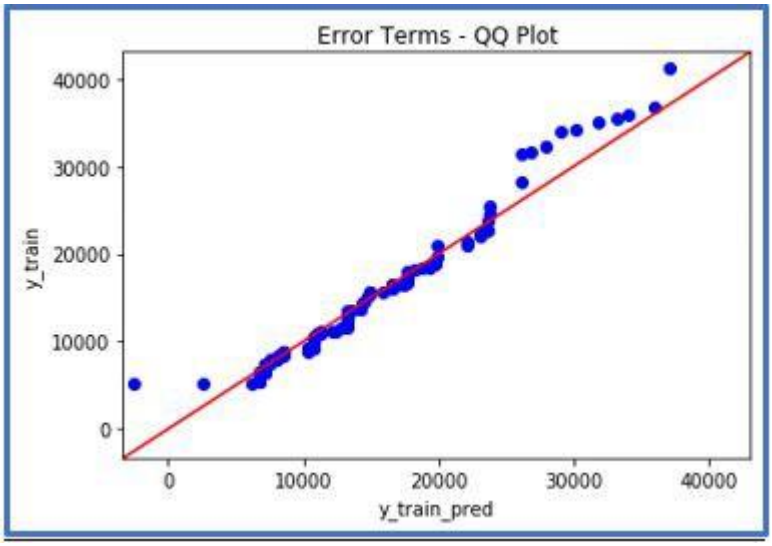
The Q-Q plot, or quantile-quantile plot, is a useful tool to understand if a set of data plausibly came from some theoretical distribution such as a Normal or exponential. It also helps to determine if two data sets come from populations with a common distribution.

a) **Similar distribution**: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x -axis

b) **Y-values < X-values:** If y-quantiles are lower than the x-quantiles.



c) X-values < Y-values: If x-quantiles are lower than the y-quantiles.



d) Different distribution: If all point of quantiles lies away from the straight line at an angle of 45 degree from x -axis

The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions.

Importance of Q-Q plot:

In linear regression model this helps, when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.

But in case if two data sets differs, then Q-Q plots useful to understand or get insight of differences which might be difficult in certain scenarios using analytical methods like chi square test or Kolmogorov-Smirnov 2-sample tests