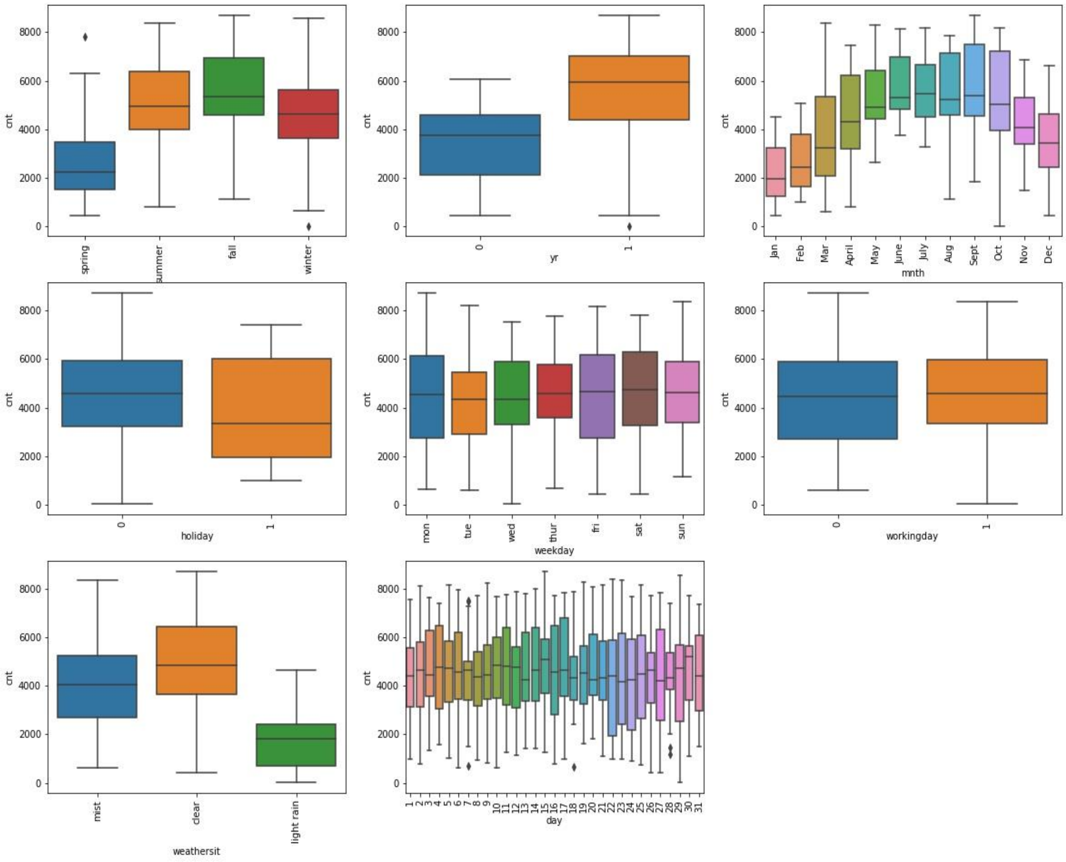
Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer abouttheir effect on the dependent variable? (3 marks)

The boxplot of categorical variables with dependent variable is as shown below.



Inferences:

season : Bike usage is highest during the fall season and lowest during spring. There is considerable difference in bike usage among different season and hence season could be a good predictor for the target variable.

yr: We see that the demand for shared bikes have increased and almost an increase of 50% from 2018 to 2019. Hence we see there is increasing demand over years.

mnth: We see the demand increasing across the year and peak in September and decrease thereafter. Hence mnth can be good predictor variable for target variable.

holiday: median bike demand is higher during non-holidays. This is opposite to what we expected. Hence the dataset could be biased or it may be getting affected by other factor.

weathersit: We see highest demand when weather is clear and least during light rain weather. Also no demand during high rain weather. Hence weather can be a good predictor for target variable.

day, weekday, workinday: For these variables , the difference in demand among respective categories is very small. Hence we can assume their effect independently on demand is not much.

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

Dummy variable is created for categorical variables in dataset and is one of the preprocessing steps in modelling. Categorical variables are not numeric and hence cannot be directly input to a model. By dummy variable creation, each categorical variable value is converted to 0 or 1. Each categorical variable creates as many new columns as it has categories with each new column named as (categorical variable name)_(category name). The value in new column is 1 if the sample belongs to the corresponding category and 0 otherwise.

For example, in dataset variable "season" has 4 categories as follows. After dummy variable creation, 4 new columns are added as follows: season_fall, season_summer, season_spring, season_winter.

We see that p categories leads to p new columns. This can be made more efficient. We can assign one category to be true when all other categories are false.

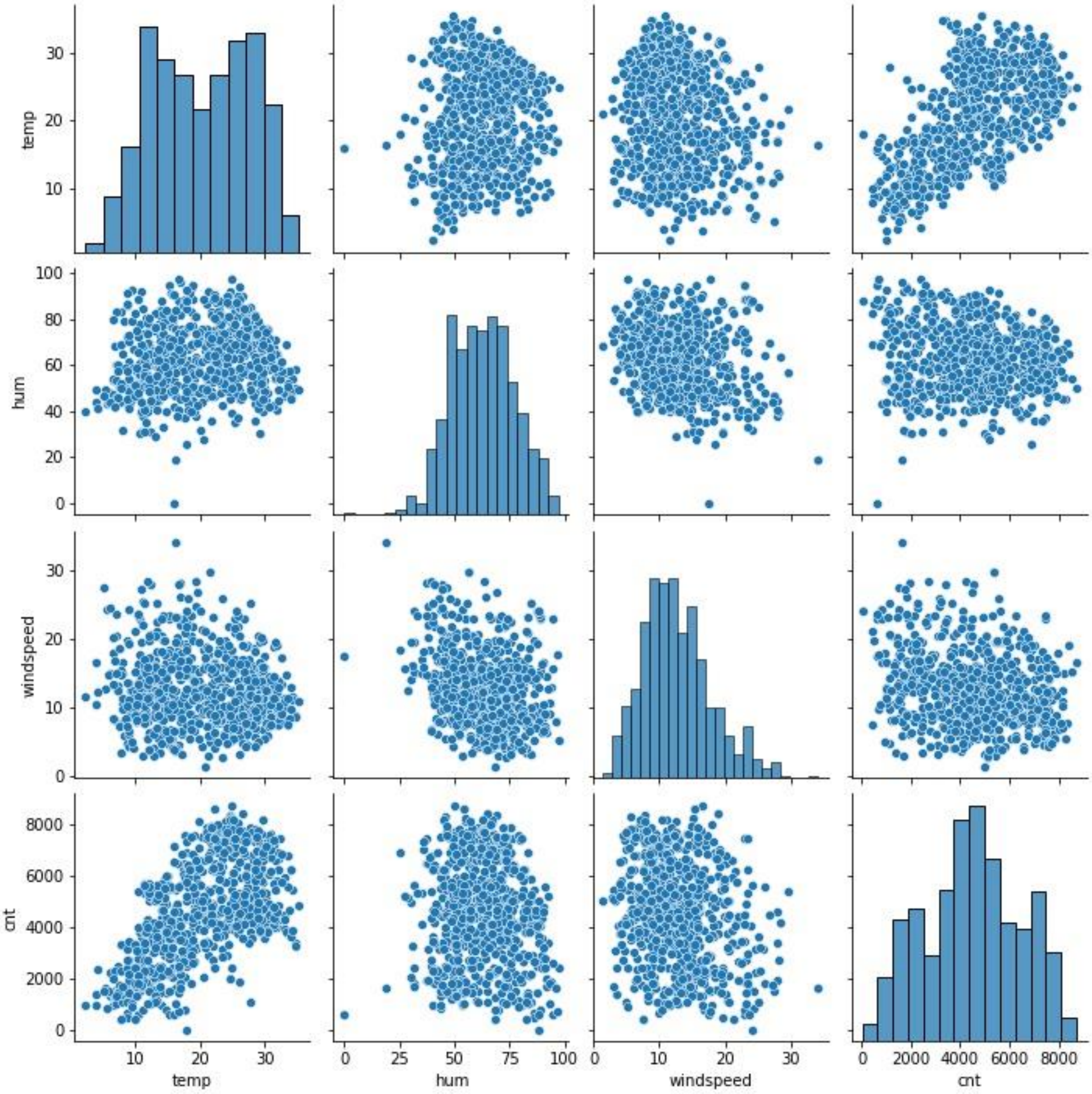
For example we can drop " season_fall " which will be recognised when remaining 3 new columns are all 0.

The drop_first=True thus helps to reduce number of the extra column created during dummy variable creation. So for a variable with "p" categories , only(p-1) new columns will be added. This reduces the correlations created between the dummy variables and make the model simpler.

In our dataset, we would have had 36 independent variables instead of 32 if we didn’t use drop_first=True as we creates dummy variables for 5 input variables. Hence the model became simpler by reducing correlation between variables.

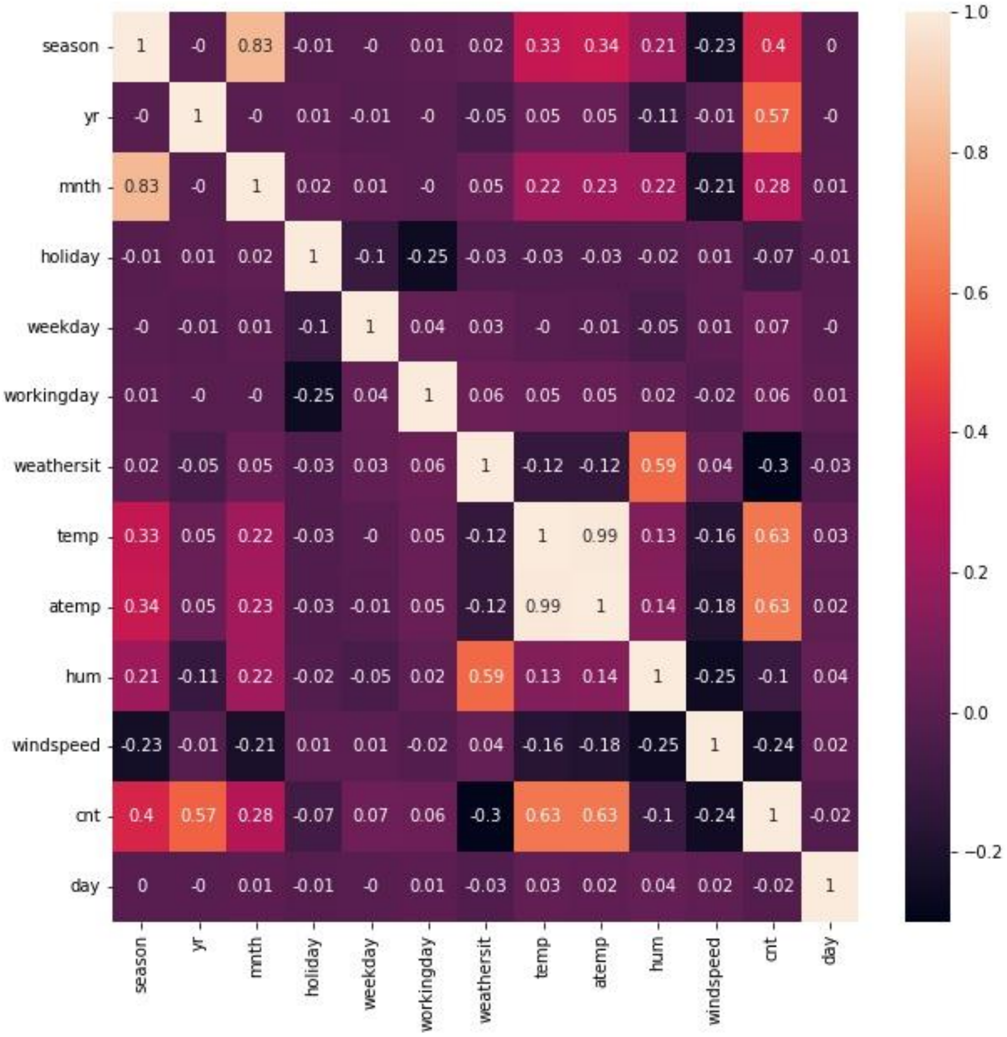
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

The pair plot of the numerical variables with target is as shown.



We can see that "temp" variable has highest correlation with target among these numeric bariables..

Let’s validate by plotting heatmap.



So we see temp has highest correlation of 0.63 with target variable.

Note : Variables "temp" and "atemp" has almost most perfect correlation of 0.99.Hence we have dropped "atemp" in our analysis.

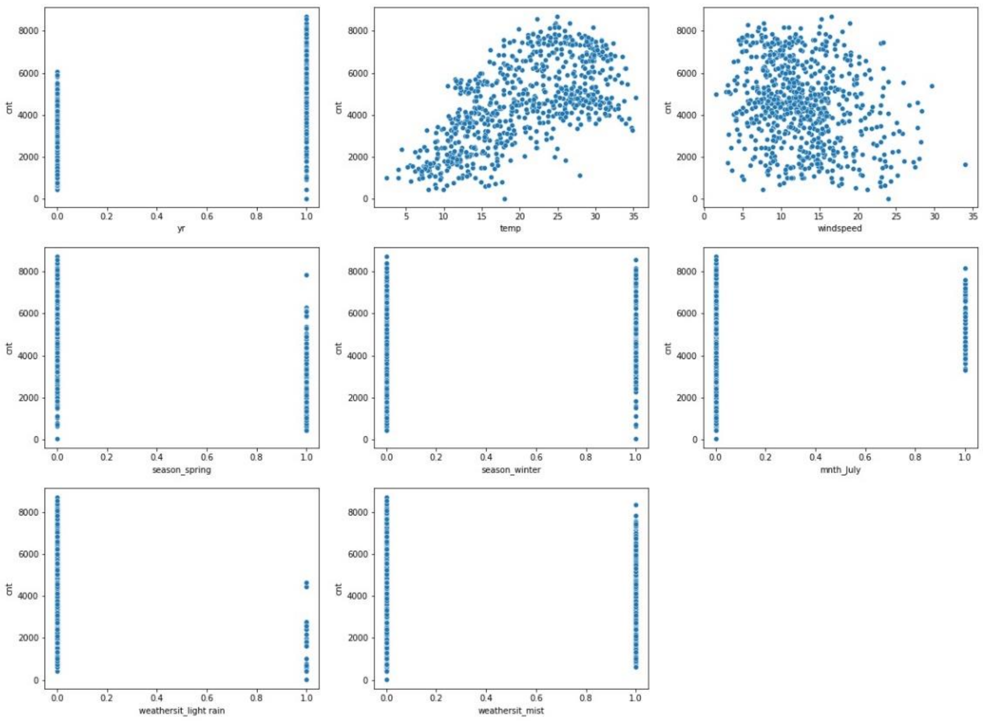
4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Linear Regression makes the following assumptions

i) There is linear relationship between target variable and independent features.

The independent features are: yr, temp, windspeed, season_spring, season_winter, mnth_July, weathersit_light rain and weathersit_mist.

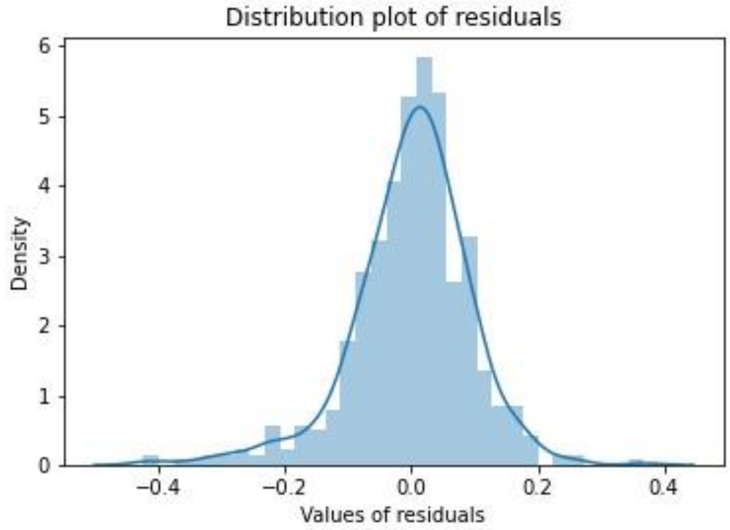
The scatter plot of target variable with independent features are as shown.



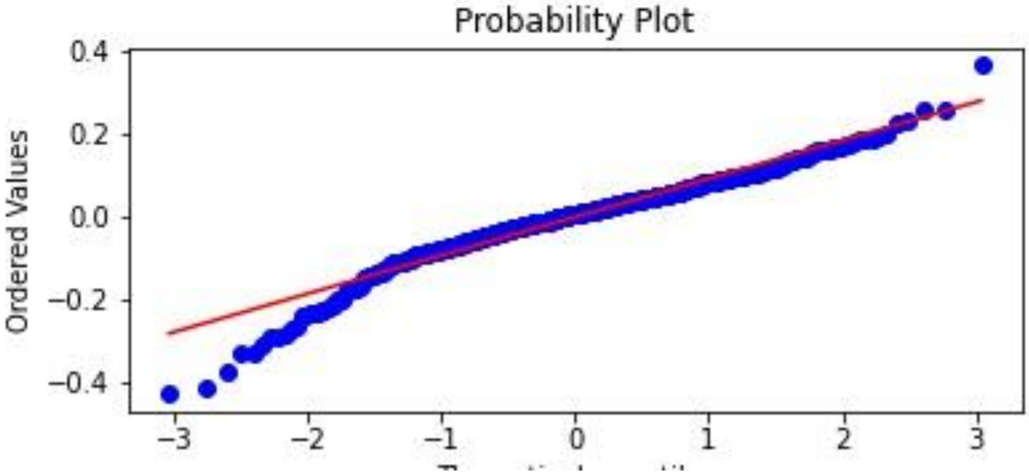
From the scatter plots we see that variables like "temp" has a linear relationship with the target variable. Thus it is validated using scatter plot of target with independent features.

ii) Residuals ie error terms are normally distributed with mean 0.

We can use the distribution plot of error terms for this.



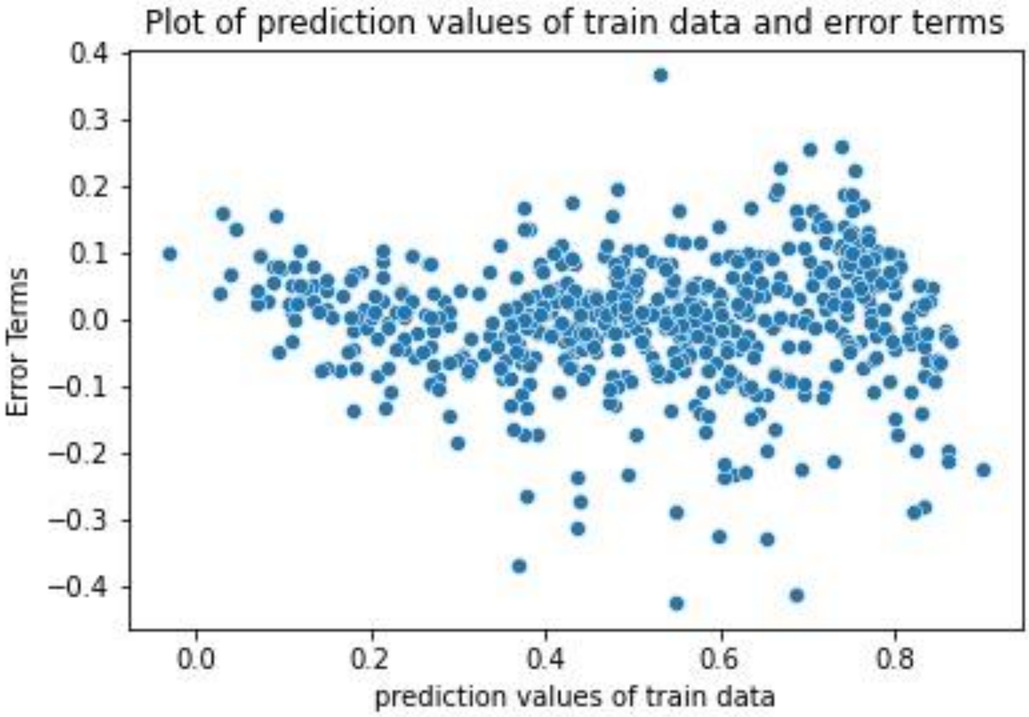
From the plot we can see that error terms are normally distributed with mean 0. For validating this, we can also use Q-Q plot between residual distribution and normal distribution.



The Q-Q plot follows a straight line. This means that error term distribution is similar to the default normal distribution. Hence the error terms are normally distributed.

iii) Homoscedasticity : assumes that error terms have constant variance.

Scatter plot of error terms with our model prediction is as shown.



From the plot we can say that residuals are having constant variance.

iv) Independence of error terms

From the scatter plot above we can see that there is no clear pattern between residuals and are completely random.

Hence the assumption that residuals are independent is validated.

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

According to final model, the top 3 features contributing significantly towards explaining the demand of the shared bikes are as follows: year, temperature and weathersit_light.

These 3 have highest value of coefficients in model.

temperature: has positive correlation of 0.49. Means if temperature increase by one unit, bike demand increase by 0.43.

year: has positive correlation of 0.23. Means year on year, bike demand increase by 23%. So we can see bike demand increase yearly.

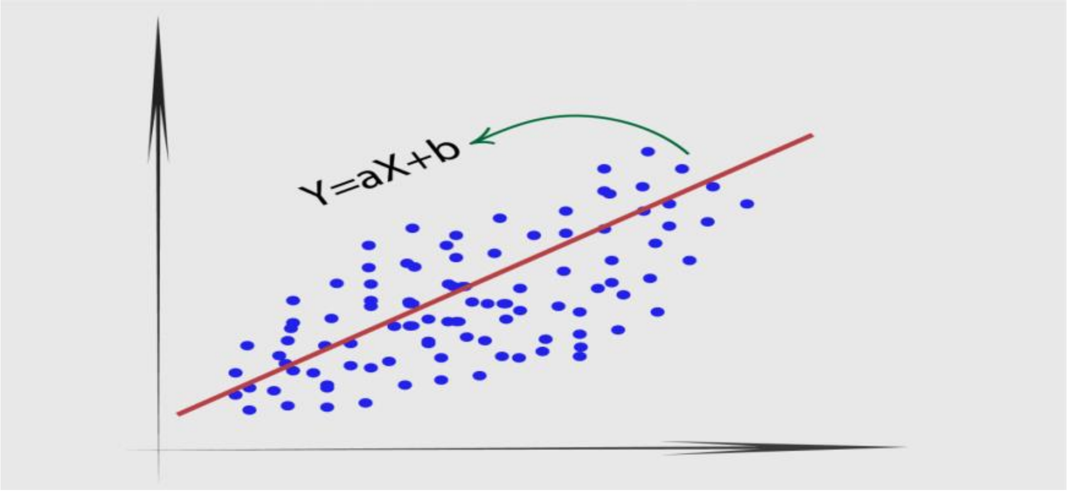
light rain : has negative correlation of 0.28. Means if light rain increase by one unit, bike demand decrease by 0.28. This is in sync with our common knowledge that bike usage is lower during rain.

General Questions

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

Linear regression is a supervised machine learning algorithm. By supervised it means that the dataset is labelled. It is used when there seems to be a linear relationship between dependent variable and independent. It is a Regression model, that is the output variable should be continuous.

Linear Regression algorithm tries to fit a straight line on train data as show below.



The equation of a simple linear regression is as follows

Y = aX + B

where Y is dependent variable

X is independent variables

A is slope of the line

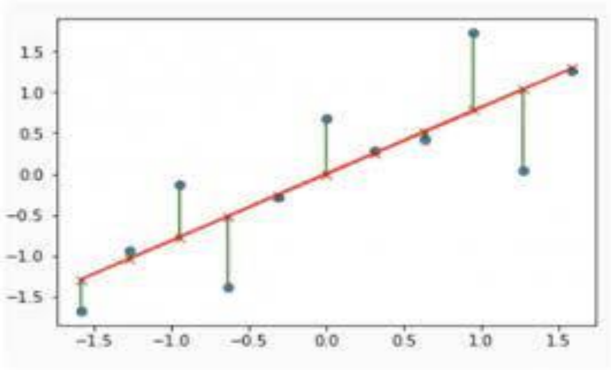
B is constant denoting y intercept of the line

The goal of linear regression algorithm is to find best possible values of coefficients which give the best fit line for the data.

For this Ordinary Least square (OLS) method can be used. The best fit line is found by minimising the RSS (Residual sum of squares) which is sum of squares of residual for each point in dataset. Residual of a point is difference between actual value and predicted value of dependent variable at that point.

RSS measure the variance in dataset that is not explained by the regression model.

A plot of residuals is shown below



Red line : regression line, blue round : actual data point, green line : residual

Error terms that is residuals is given by

$E_i = y_{actual} - y_{pred}$

Where y_{pred} : is predicted value at that point

y_{actual} : actual value at that point

Then we find the residual square sum (RSS) by adding squares of all residuals.

RSS = E_1(square) + E_2(square) + E_3(square) +…

So RSS is given by

RSS = $\Sigma_{i=1 \text{ to } n} (Y_i - B_0 - B_i)^2$

where y is dependent variable

x is independent variables

B_i is coefficient of the corresponding variables x

B_0 is constant denoting y intercept of the line

We will have to find best fit line of the form which is value of coefficients that give minimum RSS.

RSS is the cost function here. To minimise RSS usually an optimisation algorithm like gradient descent approach. In gradient descent, we start with random coefficients and then iteratively move to better coefficients values such that cost function is minimised. As the minimum RSS value, we get the best value of coefficients.

Assumptions of Linear Regression

i) There exist linear relation between dependent and independent variables.

ii) Residuals which is error terms are normally distributed with mean 0.

iii) Homoscedasticity : assumes that error terms have constant variance

iv) Independence of error terms

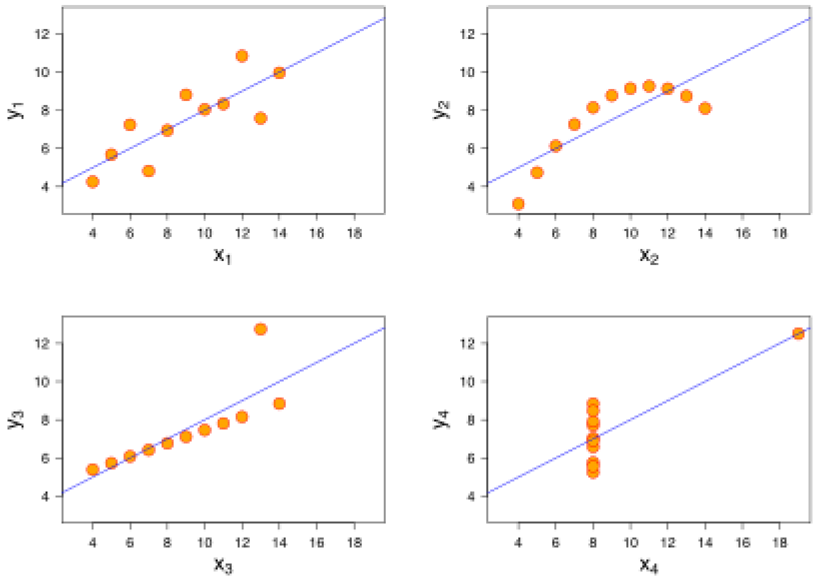
Linear Regression has lower performance compared to more complex models like decision tree and neural networks. But it is still used in industry due to its high interpretability. The coefficient of a variable shows the magnitude of change in output for unit change in that variable.

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

Statistics plays an important role in data analysis. But often statistic properties alone are not enough in describing a real life dataset.The Anscombe’s quartet shows us why visualizing a dataset is as important as its summary statistics for data analysis.

The Anscombe’s quartet consists of four datasets are similar by statistical properties but different in their actual distributions which can be seen from plotting their graphs.

The four datasets are as shown.



The four datasets have similar statistical properties: mean, variance, sample mean, sample variance, correlation, R2 and linear regression line but their distributions are different as seen from their visualisations.

Dataset 1 (top left): is a linear relation with y being a Gaussian distribution

Dataset 2 (top right): is a nonlinear relation and needs are more general regression. R2 value will be more appropriate than Pearson correlation coefficient.

Dataset 3 (bottom left): is a linear relation but has one outlier of large value that which influence the correlation. Hence needs a different regression line.

Dataset 4 (bottom right) : has just one point of large value that leads to high correlation value even though other points don't show a relation between x and y.

Hence the Anscombe's quartet shows the inadequacy of summary statistics alone in understanding a real life dataset. Hence we should visualize the data set along with summary statistics before analysing data.

3. What is Pearson's R? (3 marks)

It is the Pearson's Correlation coefficient. It is a popular ways of measuring linear correlation between 2 variables in a data.

It is given by the formula

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

Where,

r = Pearson Correlation Coefficient

x_i = x variable samples $\quad$ y_i = y variable sample

$\bar{x}$ = mean of values in x variable $\quad$ $\bar{y}$ =mean of values in y variable

It can be seen that Pearson coefficient is the normalised covariance of 2 variables. This is ratio of covariance between variables and product of their standard deviations.

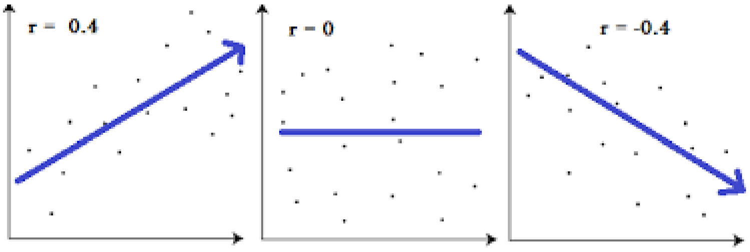
It has range -1 to 1.

1 : means perfect linear correlation

0: no linear correlation

-1: means perfect linear correlation in opposite direction.

Some examples plots of different R values are:



Positive Correlation No correlation Negative

One important point to note is that it only shows the linear relation between variations and ignores any other relation between them. The squared value of Pearson coefficient is used in evaluating linear regression models and is called R2 (r square value) of model.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Scaling is one of the pre-processing steps before modelling. In scaling, the variables in data are scaled to uniform scale so that the interpretation of coefficients of a model is meaningful.

For example, in bike sharing dataset, temp variable has range between 2 to 35 while some binary categorical variables like holiday have values 0 and 1. If not scaled, their coefficients won’t reflect magnitude of change in output for unit change in that variable.

How scaling helps and why it is performed

i) It makes the coefficients of a model more interpretable. Coefficients clearly reflect magnitude of change in output for unit change in that variable

ii) Faster training of model. Scaling to a smaller range leads to faster convergence of gradient descent optimization algorithm and hence reduce training time.

iii) Provide a degree of regularisation

But scaling does not affect model performance. All the summary statictics like R2 value and p value and F-statistics are independent of scaling.

2 types of scaling is performed

i) Normalised scaling

Also called MinMax scaling. All variables are scaled to range 0 an 1.The formula is

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Where x_{scaled} : final scaled value

x : input value

x_{min} : minimum value of x in the dataset

x_{max} : maximum value of x in the dataset

It is used in application like image processing like change image intensity to range 0 to 1.There will be some loss in information especially outliers since the data is compressed to range 0 and 1.

ii) Standardized scaling

All variables are scaled to a normal distribution with mean zero and standard deviation 1. Use formula:

$$x_{\text{stand}} = \frac{x - \text{mean}(x)}{\text{standard deviation }(x)}$$

Used in clustering algorithm, SVM etc.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?(3 marks)

VIF is variance inflation factor. It represents the amount of multicolinearity in a set of multiple regression variables.

It is given by formula,

$\text{VIF}_i = 1 / (1 – R_i\text{(squared)})$

R_i(squared) here is R2 value of a model with all variables except corresponding variable as input to predict that variable.

Multicolienarity is phenomenon of having related predictor variables in a dataset. This leads to difficulty in interpretation of coefficients of model as coefficient of one variable could be affected by those of other related variables. Hence VIF that quantify multicolinarity is used during modelling.

VIF along with p value of variables is used to decide which variable to drop during each step of modelling. VIF range above 10 is not acceptable while between 5 and 10 can't be ignored and less than 5 is acceptable.

A VIF of “x” means that the corresponding variables coefficient is inflated by “x” times due to multicolinearity.

Sometimes a variable shows a VIF value of infinity. According to formula that means it has a R2 of 1 which perfect correlation.

That means that variable has perfect correlation with another independent variable or can be expressed as linear combination of some independent variables in dataset. Hence we will have to drop that variable or its related variable to reduce multicolinearity.

For example in our dataset, "temp" and "atemp" has correlation of 0.99. Hence if we use both of them in our model, they have VIF of very high value affecting coefficients of the model. So we have dropped one of them before modelling.

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

It is Quantile-Quantile plot and sometimes also called probability plot. It is used to compare two probability distributions by plotting quantiles of both distributions against each other. It is generally used to compare a dataset distribution to theoretical distributions.

It is a non-parameteric approach in statistics where it is not assumed that the distributions come from parametric family of distributions.

Q-Q plot is popularly used to assess the linear relation between distributions. In Linear Regression, it can be specifically used to validate the assumption of linear regression that the residuals (error terms) are normally distributed in residual analysis.

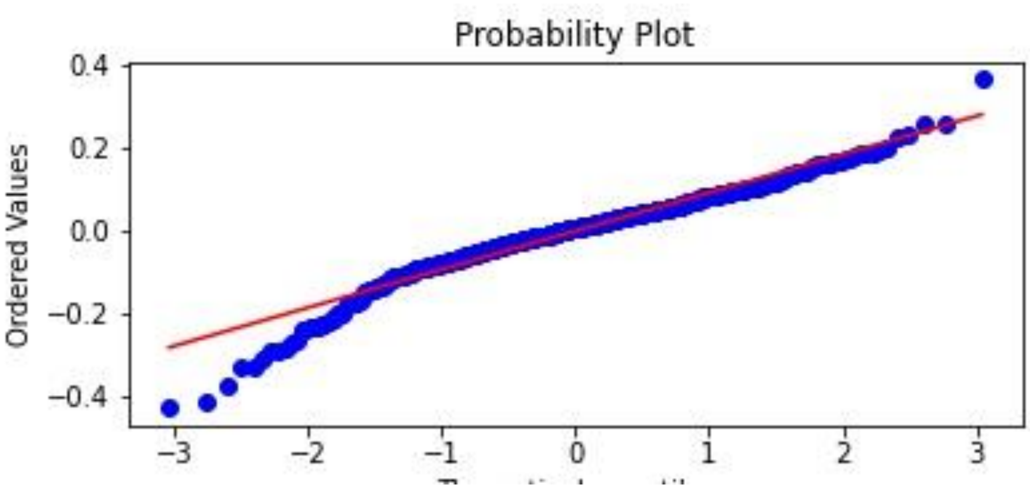
i) If Q-Q plot lies on a straight line, then it means there is linear relation.

ii) If Q-Q plot lies on an identity line y = x, then it means that they are similar distributions.

Q-Q plot in validation of linear regression assumption.

The probability plot is plotted between residuals and the normal distribution(theoretical distributions).If the plot follows a straight line, then it means that the residuals are normally distributed and thus validate our linear regression assumption.

The q-q plot between residuals and normal distribution for our data is as shown.



Since the plot follows a straight line, it validates our linear regression assumption.