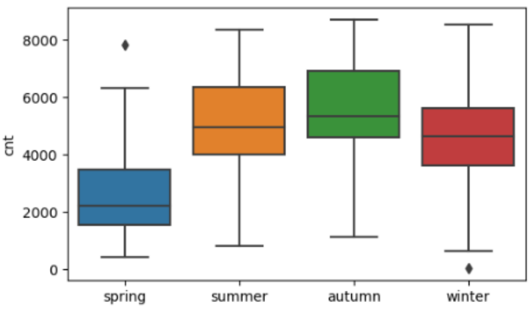
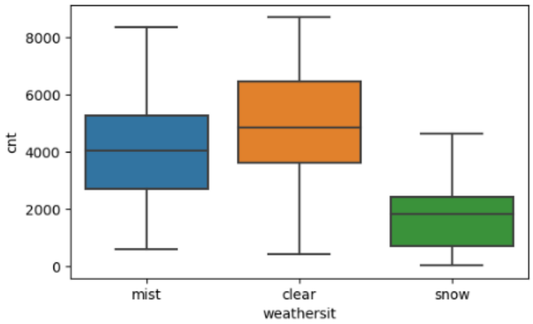
Assignment-based Subjective Questions

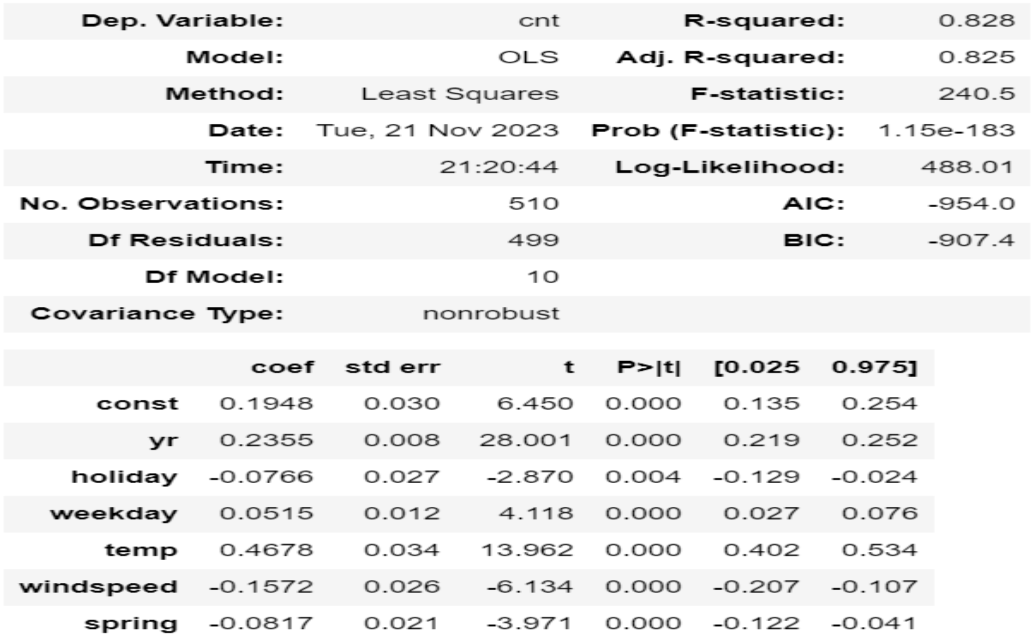
1. From your analysis of the categorical variables from the dataset, what could you infer about

their effect on the dependent variable? (3 marks)

Categorical variables used in the Bike sharing dataset is Seasons. While we have categorised into spring autumn summer and winter. Its noted that summer and autumn have a positive effect on the dependant variables whereas spring has a negative impact. Similarly, **presence of snow and mist has a negative** impact on overall demand.

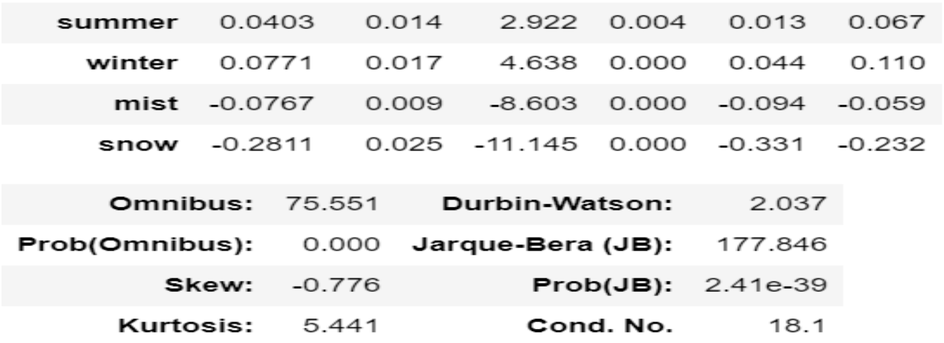






Dep. Variable:	cnt	**R-squared:**	0.828
Model:	OLS	**Adj. R-squared:**	0.825
Method:	Least Squares	**F-statistic:**	240.5
Date:	Tue, 21 Nov 2023	**Prob (F-statistic):**	1.15e-183
Time:	21:20:44	**Log-Likelihood:**	488.01
No. Observations:	510	**AIC:**	-954.0
Df Residuals:	499	**BIC:**	-907.4
Df Model:	10		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	0.1948	0.030	6.450	0.000	0.135	0.254
yr	0.2355	0.008	28.001	0.000	0.219	0.252
holiday	-0.0766	0.027	-2.870	0.004	-0.129	-0.024
weekday	0.0515	0.012	4.118	0.000	0.027	0.076
temp	0.4678	0.034	13.962	0.000	0.402	0.534
windspeed	-0.1572	0.026	-6.134	0.000	-0.207	-0.107
spring	-0.0817	0.021	-3.971	0.000	-0.122	-0.041



summer	0.0403	0.014	2.922	0.004	0.013	0.067
winter	0.0771	0.017	4.638	0.000	0.044	0.110
mist	-0.0767	0.009	-8.603	0.000	-0.094	-0.059
snow	-0.2811	0.025	-11.145	0.000	-0.331	-0.232

Omnibus:	75.551	**Durbin-Watson:**	2.037
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	177.846
Skew:	-0.776	**Prob(JB):**	2.41e-39
Kurtosis:	5.441	**Cond. No.**	18.1

Demand(y) = 0.19 + Year x 0.23 + holiday x (-0.07) + weekdayx0.05 + temp x 0.47 +windspeed x (-0.16) + spring x (-0.08) + summer x 0.04 + winter x 0.07 + mist x (-0.07) + snow x (-0.28)

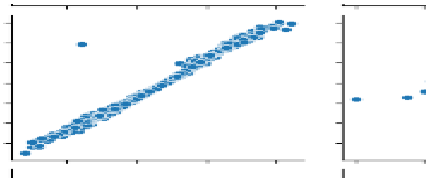
2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)

Drop first = True drops one of the first column used as dummy variable in categorical data set. By doing this helps in reducing the extra column. Hence it reduces the correlations created among dummy variables.

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

with the target variable? (1 mark)

Temperature

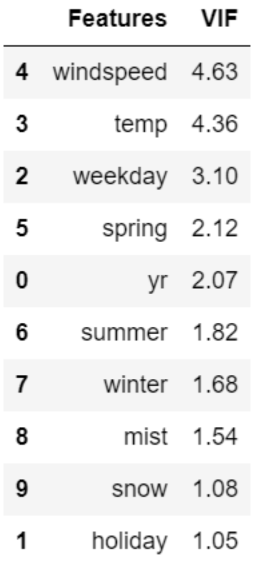


4. How did you validate the assumptions of Linear Regression after building the model on the

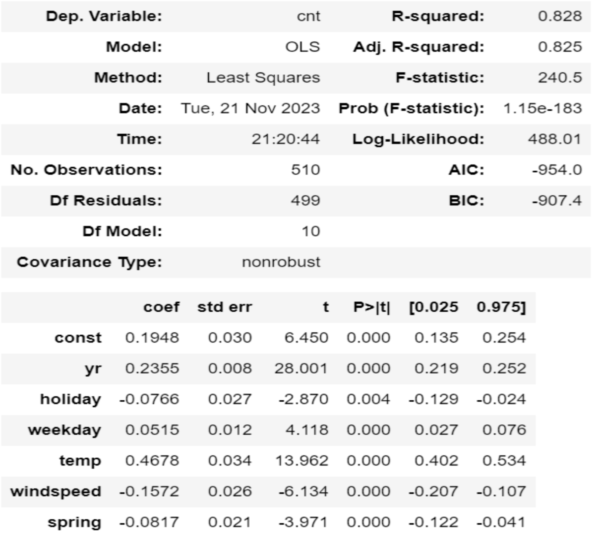
training set? (3 marks)

During model building, add variables one by one and fit into the model.

1. Rsquared value: During model fitting, observe the Rsquared value. Any slight increase in Rsquared value indicates that the variable is having an impact in the model.
2. Adjusted Rsquared value – adding more features will add a minor penalty to the model
3. The value of Prob (F-statistic) should be low.
4. Pvalues – Pvalues should be low or below 0.05. Any variable with high p values should be removed from the model
5. VIF – Variance Inflation factor. Ideal value is below 5. During variable selection, the variables are dropped one by one with higher VIF. Each time you drop the variable VIF values should be checked again.
6. RFE – Recursive Feature elimination. This will give you the actual ranking based on the factors which are used for model building.

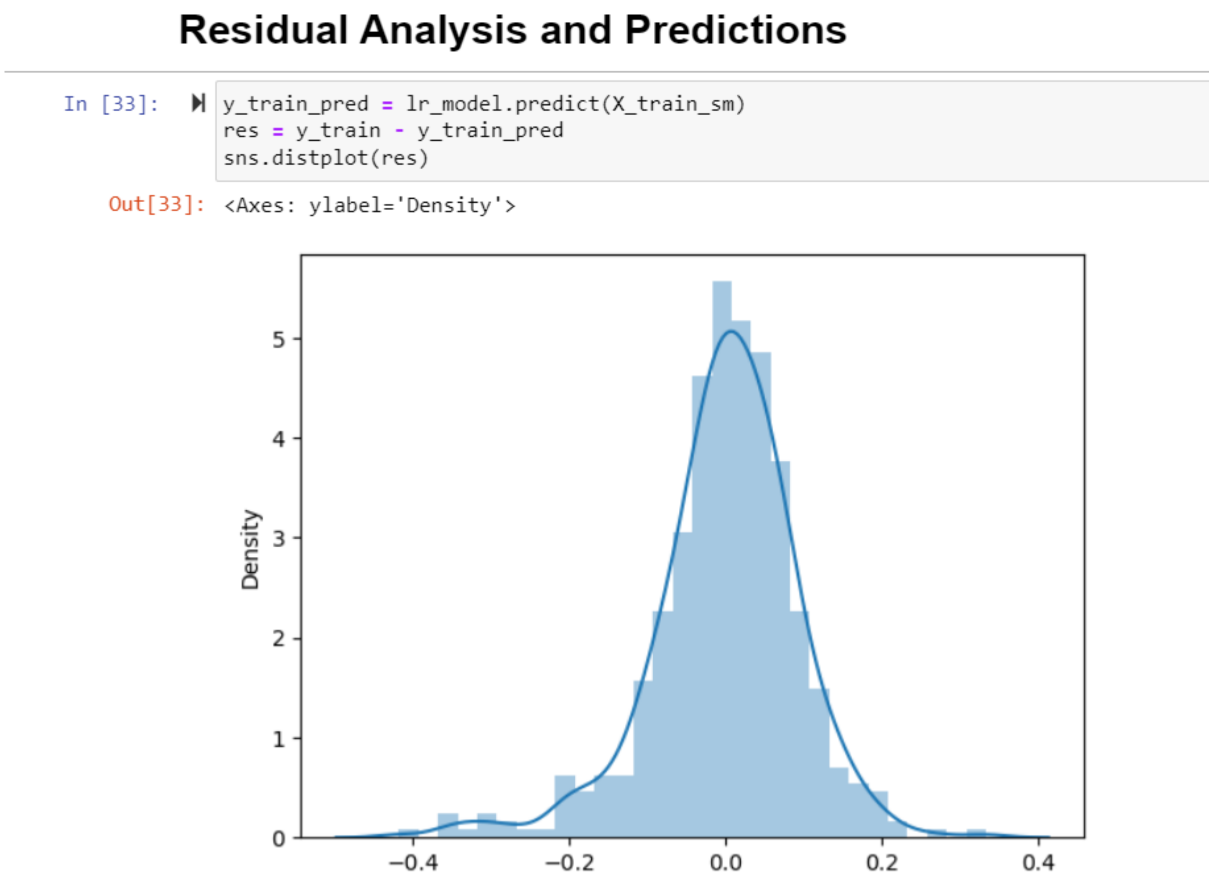


	Features	VIF
4	windspeed	4.63
3	temp	4.36
2	weekday	3.10
5	spring	2.12
0	yr	2.07
6	summer	1.82
7	winter	1.68
8	mist	1.54
9	snow	1.08
1	holiday	1.05



```
      Dep. Variable:                cnt       R-squared:      0.828
              Model:                OLS  Adj. R-squared:      0.825
             Method:      Least Squares      F-statistic:      240.5
               Date:   Tue, 21 Nov 2023 Prob (F-statistic): 1.15e-183
               Time:           21:20:44   Log-Likelihood:     488.01
  No. Observations:                510              AIC:      -954.0
       Df Residuals:                499              BIC:      -907.4
           Df Model:                 10
    Covariance Type:          nonrobust

               coef   std err        t   P>|t|  [0.025   0.975]
     const   0.1948     0.030    6.450   0.000   0.135    0.254
        yr   0.2355     0.008   28.001   0.000   0.219    0.252
   holiday  -0.0766     0.027   -2.870   0.004  -0.129   -0.024
   weekday   0.0515     0.012    4.118   0.000   0.027    0.076
      temp   0.4678     0.034   13.962   0.000   0.402    0.534
 windspeed  -0.1572     0.026   -6.134   0.000  -0.207   -0.107
    spring  -0.0817     0.021   -3.971   0.000  -0.122   -0.041
```



Residual Analysis and Predictions

```
In [33]:  y_train_pred = lr_model.predict(X_train_sm)
          res = y_train - y_train_pred
          sns.distplot(res)

Out[33]: <Axes: ylabel='Density'>
```

5. Based on the final model, which are the top 3 features contributing significantly towards

explaining the demand of the shared bikes? (2 marks)

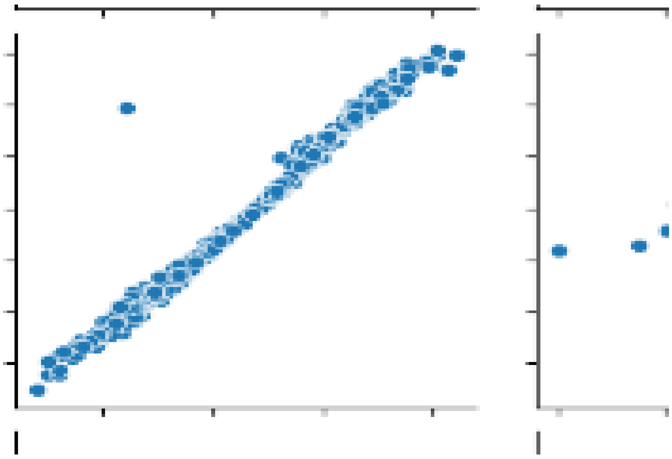
Demand(y) = 0.19 + Year x 0.23 + holiday x (-0.07) + weekdayx0.05 + temp x 0.47 +windspeed x (-0.16) + spring x (-0.08) + summer x 0.04 + winter x 0.07 + mist x (-0.07) + snow x (-0.28)

1. Temperature/ Feeling Temperature [positive effect]
2. Year or gaining popularity [positive effect]
3. Snow/Mist [Negative effect]

General Subjective Questions

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

Linear regression is a type of supervised(labelled) machine-learning algorithm that learns from the labelled datasets and maps the data points to the most linear functions. which can be used for prediction on new datasets. The algorithm computes the linear relationship between a dependent variable and one or more independent features. When the number of the independent feature, is 1 then it is known as Univariate Linear regression, and in the case of more than one feature, it is known as multivariate linear regression. The goal of the algorithm is to find the best linear equation that can predict the value of the dependent variable based on the independent variables. The equation provides a straight line that represents the relationship between the dependent and independent variables. The slope of the line indicates how much the dependent variable changes for a unit change in the independent variable.



In regression, set of records are present with X and Y values and these values are used to learn a function so if you want to predict Y from an unknown X this learned function can be used. In regression we have to find the value of Y, So, a function is required that predicts continuous Y in the case of regression given X as independent features.

Here Y is called a dependent or target variable and X is called an independent variable also known as the predictor of Y. There are many types of functions or modules that can be used for regression. A linear function is the simplest type of function. Here, X may be a single feature or multiple features representing the problem.

Below are the properties

Linearity: The independent and dependent variables have a linear relationship with one another or changes in the dependent variable follow those in the independent variable in a linear fashion.

Independence: The observations in the dataset are independent of each other. This means that the value of the dependent variable for one observation does not depend on the value of the dependent variable for another observation.

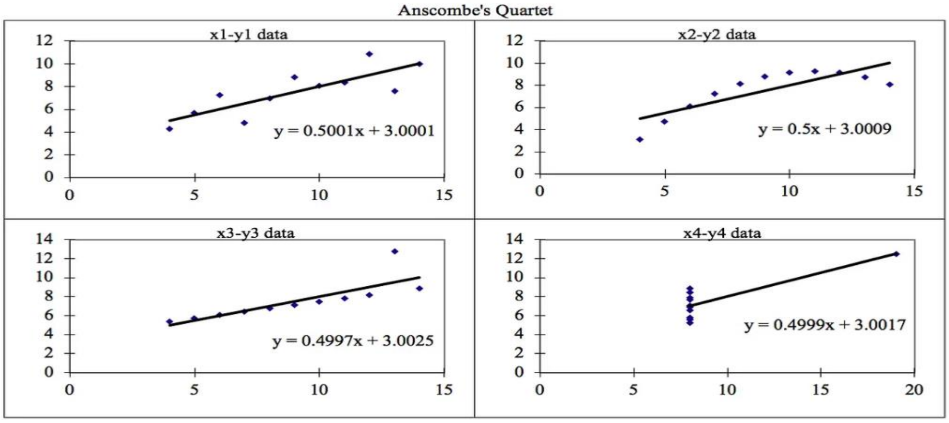
Homoscedasticity: Across all levels of the independent variable, the variance of the errors is constant. This indicates that the amount of the independent variable(s) has no impact on the variance of the errors.

Normality: The errors in the model are normally distributed.

No multicollinearity: There is no high correlation between the independent variables. This indicates that there is little or no correlation between the independent variables.

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

Its always important to plot the dataset graph for visualisation. The data looks nearly identical but when plotted it looks entirely different. A statistician Anscombe divided dataset into four types with nearly having the same statistical summary and while plotting the graph it looked entirely different.



All these graphs has identical statistical summary where as the data points are not always a linear form. This tells us about the importance of visualising the data before applying various algorithms out there to build models out of them which suggests that the data features must be plotted to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data. Linear Regression can be only be considered a fit for the data with linear relationships and is incapable of handling any other kind of datasets.

3. What is Pearson’s R? (3 marks)

Pearson’s correlation coefficient is the measure of the statistical relationship, or association, between two continuous variables. It is known as the best method of measuring the association between variables because it is based on the method of covariance. It gives information about the magnitude of the association, correlation and the direction of the relationship.

- Coefficient values can range from +1 to -1, where +1 indicates a perfect positive relationship, -1 indicates a perfect negative relationship, and a 0 indicates no relationship exists..
- Pure number: if one variable’s unit of measurement is in inches and the second variable is in quintals, even then, Pearson’s correlation coefficient value does not change.
- Symmetric: Correlation of the coefficient between two variables is symmetric. This means between X and Y or Y and X, the coefficient value of will remain the same.

Degree of correlation:

Perfect: If the value is near ± 1, then it said to be a perfect correlation: as one variable increases, the other variable tends to also increase (if positive) or decrease (if negative).

High degree: If the coefficient value lies between ± 0.50 and ± 1, then it is said to be a strong correlation.

Moderate degree: If the value lies between ± 0.30 and ± 0.49, then it is said to be a medium correlation.

Low degree: When the value lies below + .29, then it is said to be a small correlation.

No correlation: When the value is zero.

r = Covariance / (product of standard dev of (x,y))

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

Scaling is a Pre-Processing step which is applied to independent variables to normalize the data within a particular range. The dataset collected may have highly varying in magnitudes, units and range. If scaling is not done, then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we must do scaling to bring all the variables to the same level of magnitude. It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

Normalization/Min-Max Scaling: brings all of the data in the range of 0 and 1.

X = [X – min(X)] / [(max(X) – min(X))]

Standardization Scaling: 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).

X = [X – mean(X)] / sd(X)

One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers.

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

(3 marks)

This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity. Otherwise this can adversely affect the regression results. Thus, the variance inflation factor can estimate how much the variance of a regression coefficient is inflated due to multicollinearity.

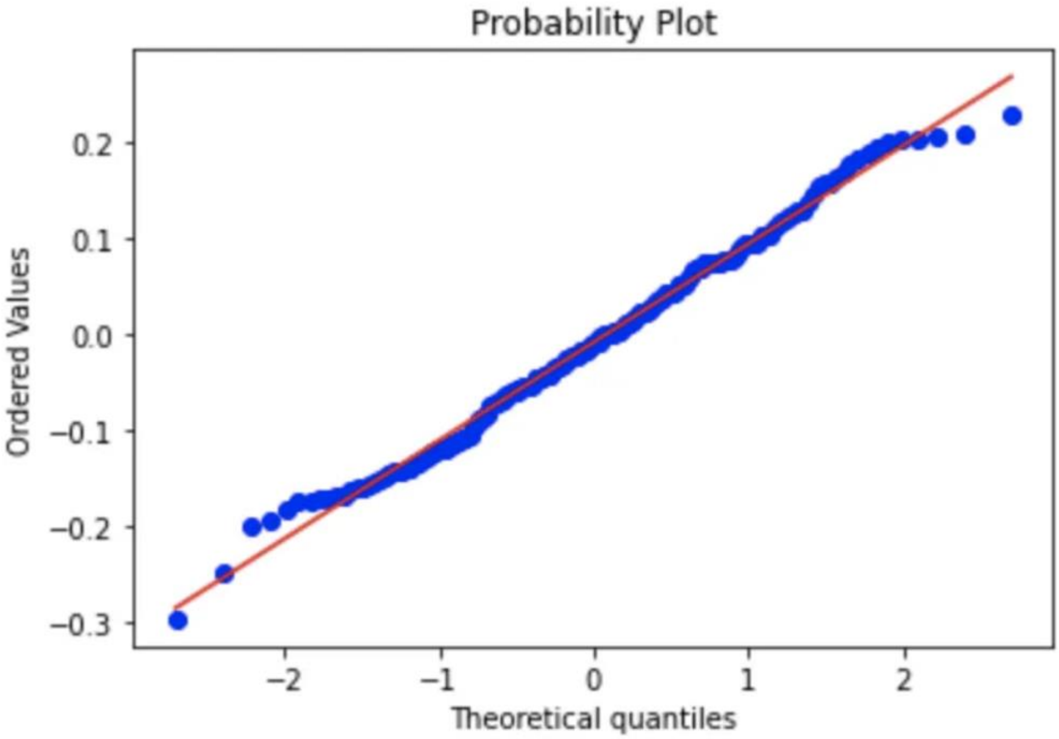
Multicollinearity creates a problem in the multiple regression model because the inputs are all influencing each other. Therefore, they are not actually independent, and it is difficult to test how much the combination of the independent variables affects the dependent variable, or outcome, within the regression model. While multicollinearity does not reduce a model's overall predictive power, it can produce estimates of the regression coefficients that are not statistically significant. In a sense, it can be thought of as a kind of double counting in the model.

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

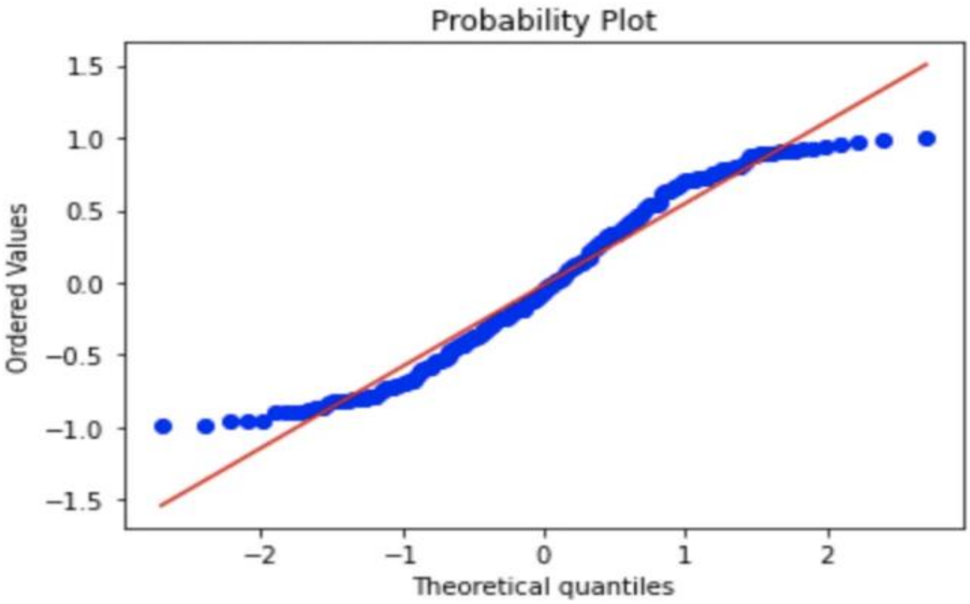
(3 marks)

Quantile-Quantile (Q-Q) plot is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.

This helps in a scenario of linear regression 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.



The fig above shows the underlying data sets are from a normal distribution



The fig above shows the underlying data sets are not from a normal distribution