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)

   Answer:

   Analysis of categorical variables from the dataset is:

   • **season:** Highest booking happening in season3 with a median of over 5000 booking. This was followed by season2, season4 and season1 in order.

   • **mnth:** Demand is continuously growing each month till June. September month has highest demand. After September, demand is decreasing.

   • **weathersit**: Demand is high if the weather is clear. Less rent when there is light rain/snow.

   • **holiday:** There is no demand, if it is a holiday.

   • **weekday:** Since this does not shows any trend, this is of no use.

   • **working day:** Median is very close, does not have much impact.

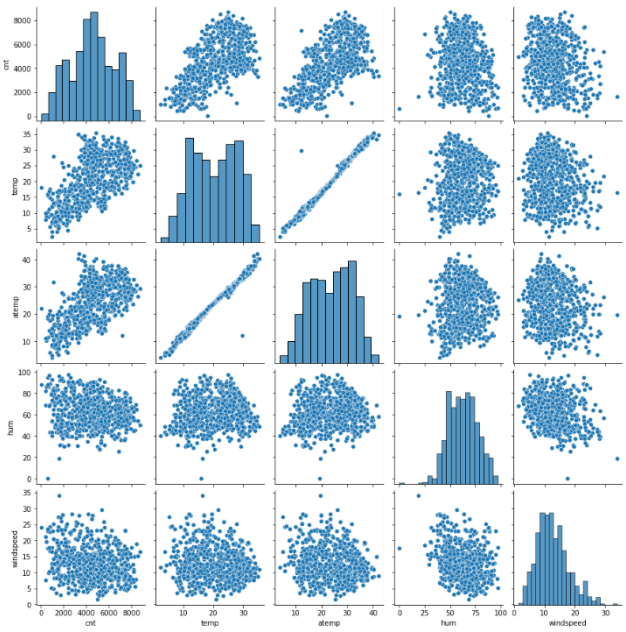
2. Why is it important to use **drop_first=True** during dummy variable creation? (2 marks)

   Answer:

   It helps in avoiding the extra column created during dummy
   Variable creation. Hence, It reduces multi-collinearity in the dataset.

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

   Answer:

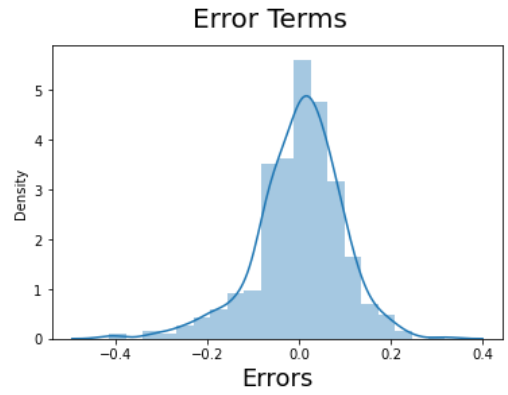


   “temp” and “atemp” are highly correlated to each other on target variable “cnt”.

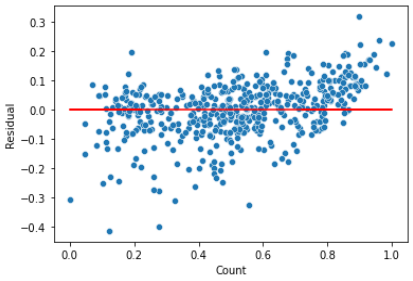
4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Answer:

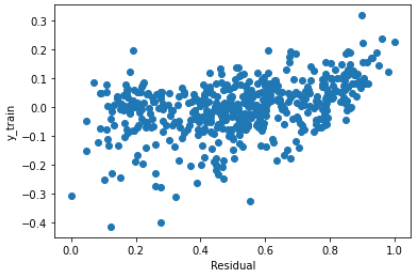
1. Error terms are normally distributed with mean zero.



2. Homoscedasticity



3. Linearity check



4. No multi-collinearity between predictor variables.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Answer:

Features “yr”, “temp” and season “windspeed” are highly related with target column, so these are top 3 contributing features in model building.

General Subjective Questions

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

Answer:

Linear regression may be defined as the statistical model that analyzes the linear relationship between a dependent variable with given set of independent variables. Linear relationship between variables means that when the value of one or more independent variables will change (increase or decrease), the value of dependent variable will also change accordingly (increase or decrease).

Mathematically the relationship can be represented with the help of following equation −

$Y = mX + b$

Here, Y is the dependent variable we are trying to predict

X is the dependent variable we are using to make predictions.

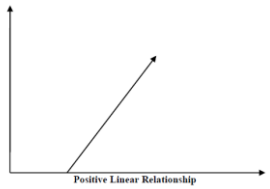
m is the slop of the regression line which represents the effect X has on Y

b is a constant, known as the Y-intercept. If X = 0,Y would be equal to b.

Furthermore, the linear relationship can be positive or negative in nature as explained below –

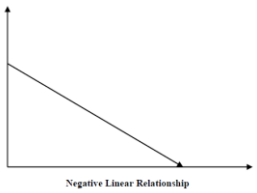
Positive Linear Relationship

A linear relationship will be called positive if both independent and dependent variable increases. It can be understood with the help of following graph −



Negative Linear relationship

A linear relationship will be called positive if independent increases and dependent variable decreases. It can be understood with the help of following graph −



Types of Linear Regression

Linear regression is of the following two types −

- Simple Linear Regression
- Multiple Linear Regression

Simple Linear Regression (SLR)

It is the most basic version of linear regression which predicts a response using a single feature. The assumption in SLR is that the two variables are linearly related.

Multiple Linear Regression (MLR)

It is the extension of simple linear regression that predicts a response using two or more features.

Assumptions

The following are some assumptions about dataset that is made by Linear Regression model −

Multi-collinearity − Linear regression model assumes that there is very little or no multi-collinearity in the data. Basically, multi-collinearity occurs when the independent variables or features have dependency in them.

Auto-correlation − Another assumption Linear regression model assumes is that there is very little or no auto-correlation in the data. Basically, auto-correlation occurs when there is dependency between residual errors.

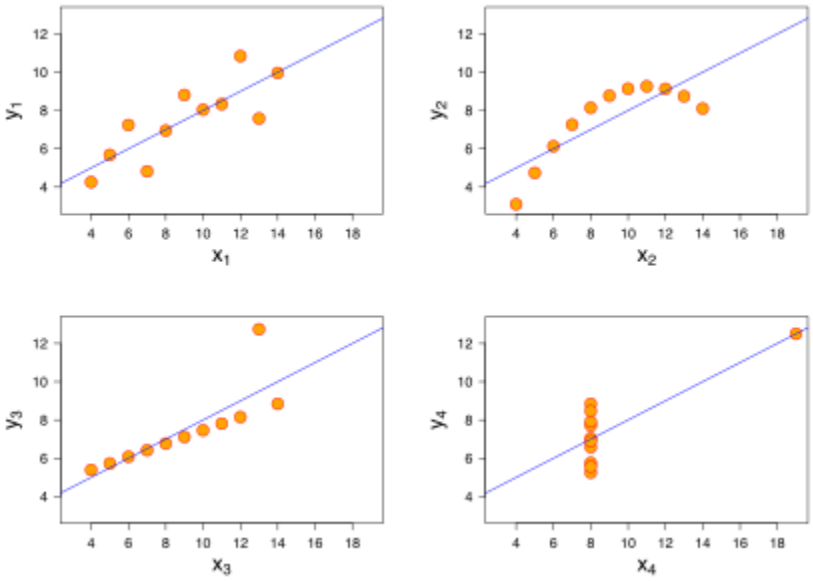
Relationship between variables − Linear regression model assumes that the relationship between response and feature variables must be linear.

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

Answer:

Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed. The essential thing to note about these datasets is that they share the same descriptive statistics. But things change completely, and I must emphasize completely, when they are graphed.

Each graph tells a different story irrespective of their similar summary statistics.



• Dataset I appears to have clean and well-fitting linear models.

• Dataset II is not distributed normally.

• In Dataset III the distribution is linear, but the calculated regression is thrown off by an outlier.

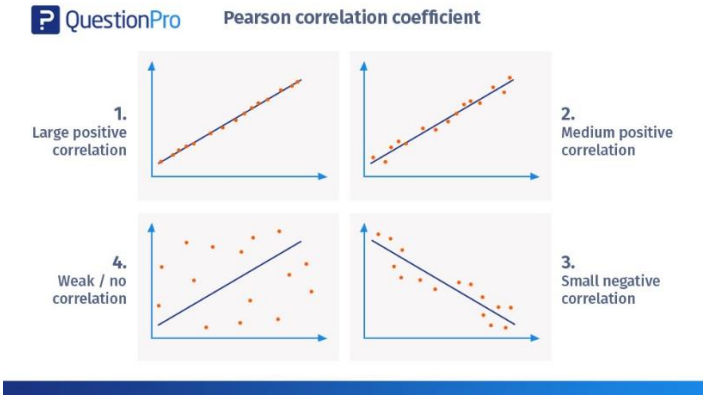
• Dataset IV shows that one outlier is enough to produce a high correlation coefficient.

This quartet emphasizes the importance of visualization in Data Analysis.

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

Answer:

Correlation coefficients are used to measure how strong a relationship is between two variables. There are several types of correlation coefficient, but the most popular is Pearson's. Pearson's correlation (also called Pearson's R) is a correlation coefficient commonly used in linear regression. If you're starting out in statistics, you'll probably learn about Pearson's R first. In fact, when anyone refers to the correlation coefficient, they are usually talking about Pearson's.



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

Answer:

Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units. It is extremely important to rescale the variables so that they have a comparable scale. If we don't have comparable scales, then some of the coefficients as obtained by fitting the regression model might be very large or very small as compared to the other coefficients.

Normalized scaling means to scale a variable to have values between 0 and 1, while standardized scaling refers to transform data to have a mean of zero and a standard deviation of 1

What?

It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

Why?

Most of the times, collected data set contains features 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 have to 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:

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

Strandardized Scaling:

• 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).

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

Answer:

The Variance Inflation Factor (VIF) is a measure of collinearity among predictor variables within a multiple regression. It is calculated by taking the ratio of the variance of all a given model's betas divide by the variance of a single beta if it were fit alone. If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables. 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).

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

(3 marks)

Answer:

A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight. 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.

Few advantages:

a) It can be used with sample sizes also.

b) Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the
presence of outliers can all be detected from this plot.

If two data sets —

- Come from populations with a common distribution.
- Have common location and scale.
- Have similar distributional shapes.
- Have similar tail behavior.

Below are the possible regressions for two data sets:

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.

Submitted By,
K.KRITHIGA.