

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?

Answer:

I have plotted the categorical variables with the target variables on barplot and has inferred the following effects on target:

- Season: 3: fall has highest demand for rental bikes.
- I see that demand for next year has grown
- Demand is continuously growing each month till June.
- September has the highest demand. After September, demand is decreasing
- When there is a holiday, demand has decreased.
- Weekdays are not giving a clear picture about demand.
- The clear weathersit has the highest demand.
- Booking count has increased significantly in 2019 compared to 2018.
- Count reduced during holidays.
- Count was higher during May to October months.
- Count was higher in the fall(Autumn) season followed by Summer.
- Count was higher on Clear, Few clouds, Partly cloudy days followed by Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist days. No records were found for Heavy Rain + Ice Pellets + Thunderstorm + Mist, Snow + Fog weather.

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

Answer:

- **`drop_first=True`** is important to use, as it helps in reducing the extra column created during dummy variable creation.
- Hence, it reduces the correlations created among dummy variables.
- Suppose, in our Bike_sharing_Datset, we have a column for seasons that contains 4 variables: 'Summer', 'Winter', 'Fall', 'Spring'. So, a season is 'Summer' or 'Winter' or 'Fall'. If the season is neither of the 3, then it's ultimately 'Spring'.
- If we do not drop one of the dummy variables created from categorical variables, then it becomes redundant with the dataset as we have a constant variable(intercept) which creates multicollinearity issues.

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

Answer:

- Temperature('temp') and feeling temperature('atemp') have almost the same correlation with the target variable (count).
- The feature "temp" has the highest correlation. It is very well linearly related with target count ("cnt")

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

Answer:

I have validated the following assumptions of Linear Regression:

- Residuals are normally distributed with mean 0.
- Checked Multicollinearity using VIFs
- Checked Linearity.
- Ensured the overfitting by looking at the R-Squared value and Adjusted R-Squared Values.

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

Answer:

- Features 'weathersit3', 'temperature ' and 'season' are highly related with the target column, so these are the top contributing features in model building.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Answer:

In the most simple words, Linear Regression is the supervised Machine Learning model in which the model finds the best fit linear line between the independent and dependent variable i.e., it finds the linear relationship between the dependent(y) and independent variable(x).

Linear Regression is of two types: Simple and Multiple.

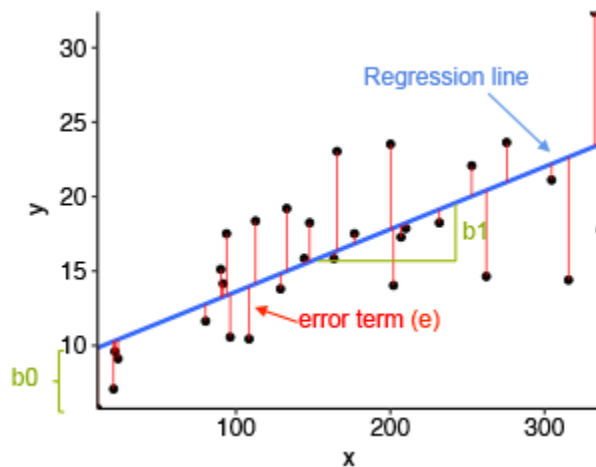
Simple Linear Regression is where only one independent variable is present and the model has to find the linear relationship of it with the dependent variable whereas, In Multiple Linear Regression there are more than one independent variable for the model to find the relationship.

Equation of Simple Linear Regression is $y = b_0 + b_1x$ where b_0 is the intercept, b_1 is coefficient or slope, x is the independent variable and y is the dependent variable.

Equation of Multiple Linear Regression, $y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \dots + b_nx_n$ where b_0 is the intercept, $b_1, b_2, b_3, \dots, b_n$ are coefficients or slopes of the independent variables $x_1, x_2, x_3, \dots, x_n$ and y is the dependent variable.

A Linear Regression model's main aim is to find the best fit linear line and the optimal values of intercept and coefficients such that the error is minimized. Error is the difference between the actual value and Predicted value and the goal is to reduce this difference.

Let's understand this with the help of a diagram.



In the above diagram,

- x is our independent variable which is plotted on the x -axis and y is the dependent variable which is plotted on the y -axis.
- Black dots are the data points i.e the actual values.
- b_0 is the intercept which is 10 and b_1 is the slope of the x variable.
- The blue line is the best fit line predicted by the model i.e., the predicted values lie on the blue line.
- The vertical distance between the data point and the regression line is known as error or residual. Each data point has one residual and the sum of all the differences is known as the Sum of Residuals/Errors.

2. Explain the Anscombe's quartet in detail.

Answer:

Anscombe's quartet is a group of four data sets that are nearly identical in simple descriptive statistics, but there are peculiarities that fool the regression model once you plot each data set. As you can see, the data sets have very different distributions so they look completely different from one another when you visualize the data on scatter plots.

Anscombe's quartet tells us about the importance of visualizing data before applying various algorithms to build models. This suggests 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 (outliers, diversity of the data, linear separability of the data, etc.). Moreover, the linear regression can only be considered a fit for the data with linear relationships and is incapable of handling any other kind of data set.

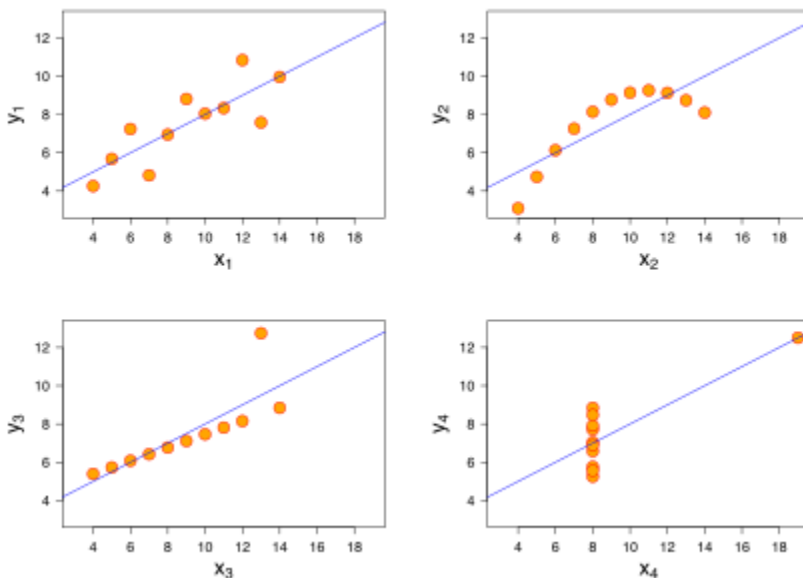
We can define these four plots as follows:

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

The statistical information for these four data sets are approximately similar. We can compute them as follows:

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	

All four sets are identical when examined using simple summary statistics, but vary considerably when graphed



We can describe the four data sets as:

- **Data Set 1:** fits the linear regression model pretty well.
- **Data Set 2:** cannot fit the linear regression model because the data is non-linear.
- **Data Set 3:** shows the outliers involved in the data set, which cannot be handled by the linear regression model.
- **Data Set 4:** shows the outliers involved in the data set, which also cannot be handled by the linear regression model.

3. What is Pearson's R?

Answer:


The Pearson correlation coefficient, also called Pearson's R, is a statistical calculation of the strength of two variables' relationships. In other words, it's a measurement of how dependent two variables are on one another.

The Pearson product-moment correlation coefficient depicts the extent that a change in one variable affects another variable. This relationship is measured by calculating the slope of the variables' linear regression.

The value of Pearson r can only take values ranging from +1 to -1 (both values inclusive). If the value of r is zero, there is no correlation between the variables.

If the value of r is greater than zero, there is a positive or direct correlation between the variables. Thus, a decrease in the first variable will result in a decrease in the second variable.

If the value of r is less than zero, there is a negative or inverse correlation. Thus, a decrease in the first variable will result in an increase in the second variable.



When plotted on a diagram, a positive correlation will see a line which slopes downwards from left to right and a negative correlation will see a line which slopes downwards from right to left.

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

Answer:

Scaling is a method to normalize the range of independent variables. It is performed to bring all the independent variables on the same scale in regression. If Scaling is not done, then the regression algorithm will consider greater values as higher and smaller values as lower values.

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.

Example Weight of a device = 500 g, and weight of another device is 5 kg. In this example a machine learning algorithm will consider 500 as greater value which is not the case. And it will make a wrong prediction.

Machine Learning algorithm works on numbers not units. So, before regression on a dataset it is a necessary step to perform.

Scaling can be performed in two ways: Normalization: It scales a variable in range 0 and 1.

Standardization: It transforms data to have a mean of 0 and standard deviation of 1

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

Answer:

If there is perfect correlation, then $VIF = \text{infinity}$. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get $R^2 = 1$, which leads to $1/(1-R^2)$ infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

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

Answer:

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

It is used for determining if two data sets come from populations with a common distribution. A Q-Q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. Whether the distribution is Gaussian, Uniform, Exponential or even Pareto distribution, it can be found out.

Few advantages:

- It can be used with sample sizes also
- 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.

It is used to check following scenarios:

If two data sets

- come from populations with a common distribution
- have common location and scale
- have similar distributional shapes
- have similar tail behavior

Interpretation:

A Q-Q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.

Below are the possible interpretations for two data sets.

- Similar distribution: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x -axis
- Y-values < X-values: If y-quantiles are lower than the x-quantiles.
- X-values < Y-values: If x-quantiles are lower than the y-quantiles.

Thank You,

Devi Mullapudi