Linear Regression Assignment Subjective Questions

Abbas Bagwala

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: The dependent variable 'cnt' was influenced by several variables. It was mostly dependent upon the temperature ('temp'), whether or not there was light rain or light snow ('Light_Rain/Snow'), the year ('year'), and whether of not it was spring ('spring'). The dependent variable 'cnt' had a positive correlation with temperature and the year, and a negative correlation with light rain or snow and spring.
Moreover, it was affected, albeit to a lesser extent, by the following variables: whether or not it was a working day ('workingday'), the speed of the wind ('windspeed'), whether or not it was July, March, September ('july', 'mar', 'sep'), whether is was Monday ('mon'), whether it was misty ('misty'), and finally whether it was winter ('winter').
The effects can be summed up in the following formula:

cnt = 0.202 + (0.230 x year) + (0.044 x workingday) + (0.457 x temp) + (-0.097 x windspeed) + (-0.059 x july) + (0.046 x mar) + (0.065 x sep) + (0.053 x mon) + (-0.292 x Light_Rain/Snow) + (-0.080 x Misty) + (-0.126 x spring) + (0.043 x winter)

The variable that had the most sizeable impact on the dependent variable is temperature ('temp': 0.457), followed by light rain/snow ("Light_Rain/Snow": -0.292), year ("year": 0.230), and finally spring ("spring":0.126).

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

Answer: During the creation of dummy variables, one value is always expressed by all values being zero. For instance, if we have four categories called apples, oranges, guavas, and mangoes, then, giving the value '0' to the first three can represent that the fruit in consideration is a mango. Thus, we can drop it without any loss of data.
As in all coding exercises, brevity should be preferred. It saves space and time. Using drop_first=True drops the first variable from 'n' number of dummy variables, making the number of variables 'n-1'.

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

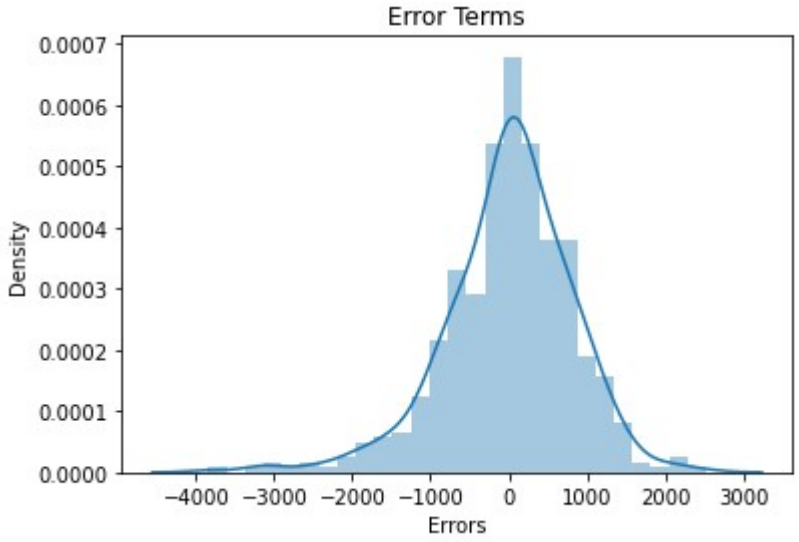
Answer: Among all the variables, the variables ‘temp’ and ‘atemp’ have the highest correlation with the target variable. And because we later drop ‘atemp’ due to VIF, we figure out that the variable ‘temp’ is actually the most highly correlated numerical variable with the target variable ‘cnt’.

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

Answer: The following are commonly held to be the five key assumptions of linear regression:

- A linear relationship
- Normality of the residuals
- No or little multicollinearity
- No autocorrelation
- Homoscedasticity

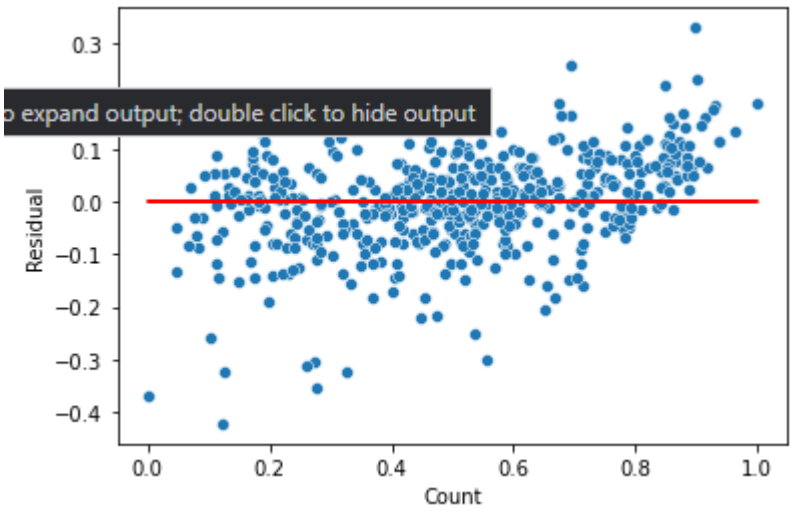
A linear relationship was established because **a straight line’s equation** clearly results in the target variable being explained. The **normality of the residuals** was done in the section under residual analysis, where a distplot of errors terms was plotted with its density (no. of times it occurred). This plot was clearly normal. See below:



We see that there **was none or little multicollinearity** by deciding upon a model with the optimal VIF scores.

We see that there was no autocorrelation between the variables by looking at the Durbin-Watson number, which in our case was 2.036, which indicates virtually **no autocorrelation between variables**.

Finally, the homoscedasticity was found out by doing a scatterplot between the target variable and the residuals, and running a straight line between them. See below:



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: Based on the model, the top three features contributing significantly towards explaining the demand of the shared bikes are temperature ('temp': 0.457), light rain/snow ("Light_Rain/Snow": -0.292), and year ("year": 0.230).

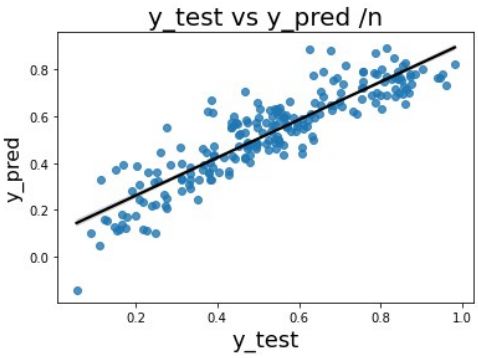
General Subjective Questions

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

Answer: A linear regression algorithm, which is a kind of supervised learning algorithm, is an algorithm that is used to predict the value of a dependent variable ('y') based on either one or several ideally independent variables from a dataframe ('X'). Mathematically, we may express it as:

$\mathbf{y = a + bx_1 + cx_2 + dx_3 \ldots}$ where (a, b, c, d are coefficients and x1, x2, x3… are ind. variables)

If there is only one independent variable (thus, one variable x in the dataframe X), then we call the linear regression a simple linear regression. If there are more than one variable, then the linear regression performed is termed multiple or multivariate linear regression.



The central aim of running a linear regression algorithm is to find out which 'line' best fits the given data. It is termed a linear regression algorithm because the data is supposed to be explained by a straight line. For datapoints that do not fit in a straight line, a linear regression algorithm should be (and cannot effectively be) used.

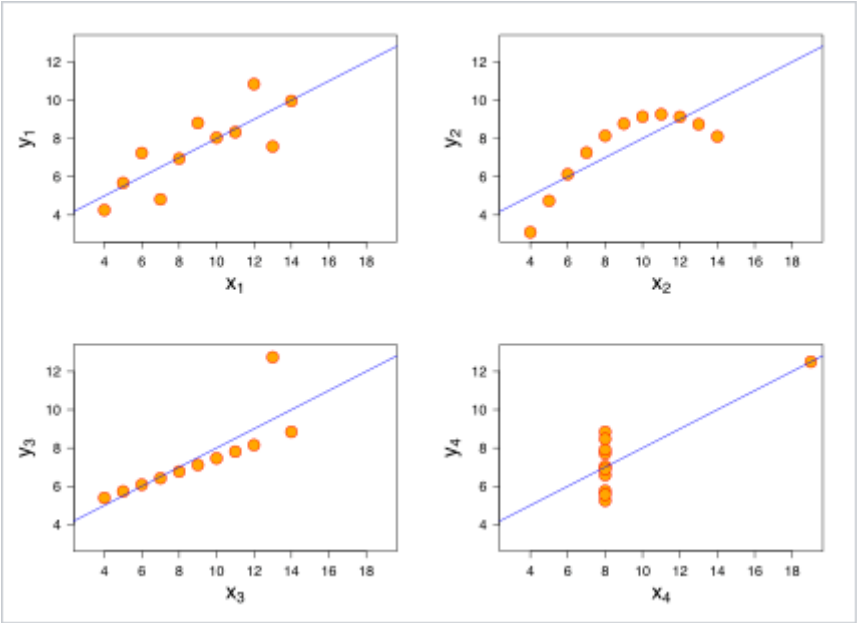
2. Explain the Anscombe's quartet in detail. (3 marks)

Answer: Anscombe's quartet was developed by an English statistician called Francis Anscombe to demonstrate why graphing data while analyzing it is an important and crucial step. The quartet is a graphical representation of four nearly identical descriptive statistics. The statistical data is shown in the table below.

Anscombe's quartet

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

All the four datasets in the data above have the same mean, sample variance, correlation, coefficient of determination, and other similar properties. If one were to only look at their statistics, one would be tempted to conclude that the data-sets are similar. However, the graphical representation of the four datasets show that this is not the case. This is how Anscombe's quartet shows that the graphical representation of datasets for data analysis is crucial. See below.



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

Answer: Pearson's correlation coefficient (Pearson's R) is a test in statistics that is used to measures the statistical relationship between two continuous variables. Pearson's correlation coefficient measures the association between two variables by using the method of covariance. It's calculated using the following formula:

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

r = correlation coefficient

x_i = values of the x-variable in a sample

$\bar{x}$ = mean of the values of the x-variable

y_i = values of the y-variable in a sample

$\bar{y}$ = mean of the values of the y-variable

If the value of r is near +1 or -1, then it is said to be a perfect correlation, between ± 0.50 and ± 1 is a high degree of correlation, between ± 0.30 and ± 0.49 a moderate correlation, and ±0.29 means a small correlation. A value of 0 indicates no correlation.

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

Answer: Scaling, also known as feature scaling, in an important data preprocessing step in machine learning. In scaling, the values of all variables/features are transformed to be put into one range so that values of all variables can be compared at once. Scaling is performed because some algorithms that compute distance between variables are biased towards large values. Another reason is that scaling makes it easier to understand the features/variables by merely looking at their values because their values can now be easily compared.

There are two ways in which scaling can be done: normalization and standardization. In normalization, minimum and maximum value of features are used, while in standardization, mean and standard deviation are used for scaling. Normalization is typically used when features are of different scales, while standardization is used when we want to ensure that the mean is zero and the standard deviation is one unit. Normalization is bounded between either [0,1] or [1,-1], while standardization is not bounded by a range. Normalization is really affected by outliers, while standardization is less affected.

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

Answer: The formula for calculating the VIF of a certain variable is given by

$$VIF_i = \frac{1}{1 - R_i^2}$$

R_i^2 represents the unadjusted coefficient of determination for refressing the i^{th} independent variable. When R_i^2 is equal to zero, the VIF will equal to 1, which means that the independent variable is not correlated to the remaining variables. However, if there is prefect correlation, then R_i^2 will be equal to 1, thus making VIF equal to 1/0, or infinity. In short, VIF is infinite when the variable under consideration is completely correlated to (and thus determined by) the remaining independent variables.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

Answer: The Q-Q plot is one of many methods used to determine the distribution of continuous variables graphically. The quantile-quantile, or Q-Q, plot is a scatter plot that is created by plotting two different quantiles against each other. The first one is of the variable we are testing the hypothesis for and the second one is the actual distribution that the hypothesis is being tested against.

A Q-Q plot is used to compare how distributions are shaped and to provide a graphical view of properties such as location, scale, and skewness. Given below is an example of a few Q-Q plots: