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)

Below is the final equation of line received after performing multiple linear regression:

cnt = 0.2401yr - 0.0985holiday + 0.4623temp - 0.2857lightSnowRain - 0.0754MistCloudy - 0.1258Spring + 0.0587Winter + 0.1724

The categorical columns have below slope coefficients:

+0.2401yr

-0.0985holiday

-0.2857lightSnowRain

-0.0754MistCloudy

-0.1258Spring

+0.0587Winter

This signifies direct relationship between these categorical variables and the target variable 'cnt'. Their slopes showing positive or negative dependence, along with the degree of dependence.

Keeping other predictor variables constant, the value of bike sharing will increase/decrease based on the sign and value of slopes of each of below predictor variables.

Specific inferences are:

+0.2401yr = There has been a 24.01 % increase in bike use per year.

-0.0985holiday = Holiday affects the bike use by 9.85 %.

-0.2857lightSnowRain = When the weather is Light snow or rain, bike use is decreased by 28.57 %.

-0.0754MistCloudy = When the weather is mist/cloudy, bike share use decreases by 7.54 %.

-0.1258Spring = During Spring season, bike use decreases by 12.58 %.

+0.0587Winter = Winter sees an increase in bike sharing by 5.87 %.

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

Dropping of one of the columns among newly created set of Dummy variables column is done to reduce complexity while not compromising the result.

Example If we have a Categorical column with 3 different values ‘Yes’, ‘No’ and ‘May be’; if we assign 1 for Yes, 0 for No, then a record with 00 (0 in Yes and 0 in No), infers a scenario of ‘Maybe’ (coz it conveys this record is neither Yes nor No).

Hence, we are conveying the same meaning and reducing the complexity of model building by using drop_first=True during dummy variable creation.

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

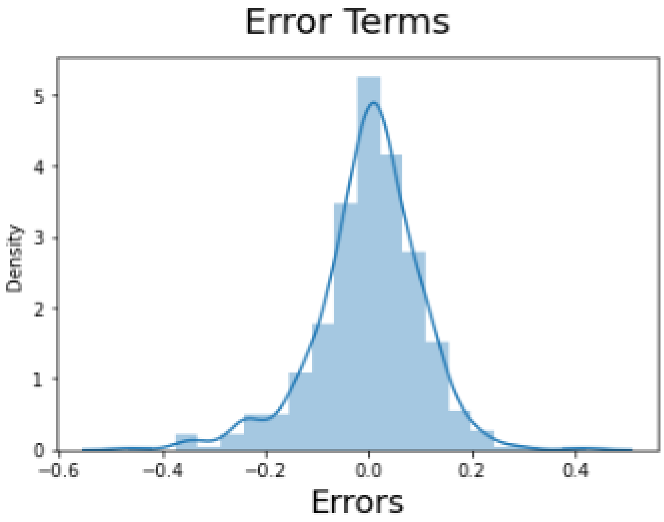
with the target variable? (1 mark)

‘Registered’ users have the highest correlation.

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

training set? (3 marks)

Error terms are normally distributed, as can be seen with graph we received below:



There is a linear relationship between dependent and target variable. There is constant variance of residuals at all levels of x.

We could see by plot that the target and the independent variable are linearly dependent.

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

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

Looking at the final equation we received:

cnt = 0.2401yr - 0.0985holiday + 0.4623temp - 0.2857lightSnowRain - 0.0754MistCloudy - 0.1258Spring + 0.0587Winter + 0.1724

We can say that temp (positive relation by slope of 0.4623), year(positive relation by slope of 0.2401) and LightSnowRain(weather being Light snow or rainy with negative slope of 0.2857) are the top 3 features, in order, contributing significantly towards explaining the demand of the shared bikes

General Subjective Questions

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

Linear regression is a form of machine learning method, where we are training a model to predict the behavior of our data-based on some variables. The predictor variables are linearly correlated with the target variable.

The linear regression equation as:

$$y = a + bx$$

Where a and b given by the formulas:

$$b(slobe) = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$$

$$a(intercept) = \frac{n\sum y - b(\sum x)}{n}$$

Here, x and y are two variables on the regression line.

b = Slope of the line

a = y-intercept of the line

x = Independent variable from dataset

y = Dependent variable from dataset

For a multiple linear regression, where more than two dependent variables are involved in a linear relationship, we have below formula:

Y=β0+β1X1+β2X2+...+βpXp+ϵ

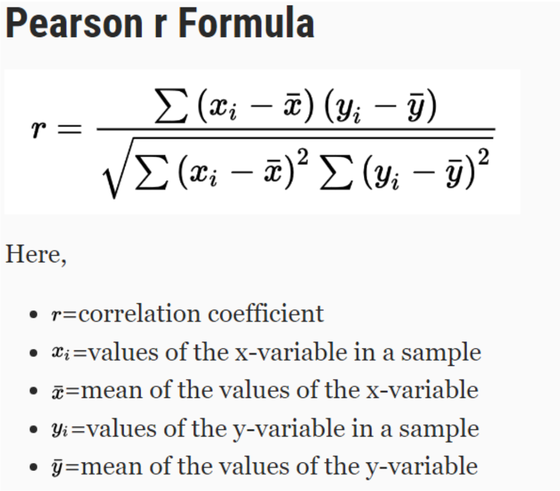
2. Explain the Anscombe's quartet in detail. (3 marks)

Anscombe's Quartet is a group of four data sets which are nearly identical in simple statistical analyis, but there are some peculiarities in the dataset that misleads the regression model built. They have very different distributions and appear differently when plotted on scatter plots.

It illustrates the importance of plotting a graph before going to build model. , all the important features in the dataset must be visualised before implementing any machine learning algorithm on them which will help to make a good fit model.

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

Pearson's R is the product-moment correlation coefficient (PPMCC), which is a measure of linear correlation between two sets of data. The coefficient lies between -1 to +1. If the value is 0, there is no relationship.



Pearson r Formula

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

Here,

- 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

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

Scaling is a method of bringing all the variables being analyzed in a model to a common scale, thereby not compromising there separate interpretations. Example if we have data in our table with values 0 and 1, and another column with values 1000 to 9000, we would like to scale down the larger values.

This is done because collected data set may contain features highly varying in magnitudes, units and range. If scaling is not done, then algorithm only takes magnitude in account and not units hence may lead to incorrect modeling.

Normalization scaling will scale all the feature between 0 to 1 and standardization scaling will scale the feature with mean 0 and 1 std deviation.

Normalization is generally used when our dataset has outliers and standardization when we are caring about the distribution.

Standardization: X_new = (X - mean)/Std

Normalization: X_new = (X - X_min)/(X_max - X_min)

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

(3 marks)

This happens when there is a perfect correlation. In case of perfect correlation, we get R2 = 1 , which lead to 1/(1-R2) infinity. Hence, 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).

In order to solve this problem, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

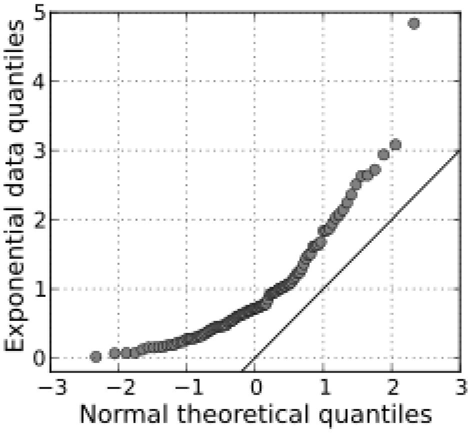
6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

(3 marks)

It is a plot of two quantiles against each other. Q-Q plot is used to check if the random variable X follows any certain distribution. It is done by taking percentile of the provided random variable X and comparing with other Normally distributed random variable.

It presents data such as- if the two distributions being compared are similar, then the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x.

Below is a sample Q-Q plot with the 45 degree reference line:



5
4
3
2
1
0
−3
−2
−1
0
1
2
3
Exponential data quantiles
Normal theoretical quantiles