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:

"Season, month , weather" are highly significant for more bikes being used. A clear or cloudy weather have great impact.

"yr": Observed that there is a growth in the number of bikes being rented from year 2018 to 2019. This means company is getting publicized more.

"Holiday": Bikes are more widely used on holidays. This implies that, customer are more interested to do physical activity on holidays.

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

Answer: *'*

drop_first=true' helps in reducing the extra column created during dummy vairable creation.Hence it reduces the correlations created amonth dummy variables.

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

Answer:

'registered' and then 'casual' are highly correlated with target variable -'cnt'. But, these two are holding subset data of target variable. Hence, we should look for other variables.

'temp','atemp' – temparature and feeling temparature are highly correlated with target variable.

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

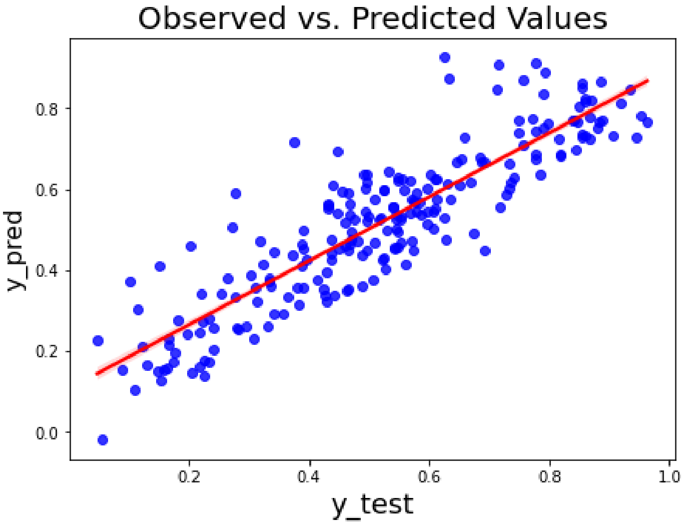
Answer:

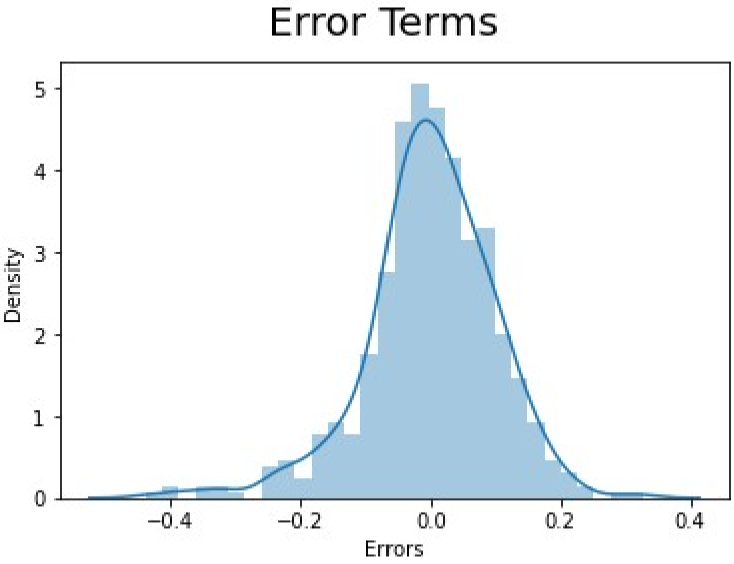
Based on below two factors I have validated my model

1. 'observed vs predicted' are similar

2. *Error terms are normally distributed*

Below two pictures captured from my jupyter notebook





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:

“Year”: More bikes are rented on 2nd year (2019)

“weather”: Clear or less cloudy climate leads to more bikes being rented

“temp”: normal tempature (not too low, not too high)

General Subjective Questions

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

Answer:

Linear Regression is a linear approach for modelling the relationship between a scalar response(target) and one or more explanatory variables(dependent and independent variables). Linear Regression fits a linear model with coefficients

$\Sigma_i = (Y_i - (\beta_0 + \beta_1 X))^2$

For Multiple Linear Regression it wil expand as Y=$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_k X_k$

Y = vector of observed values , also called as target / response

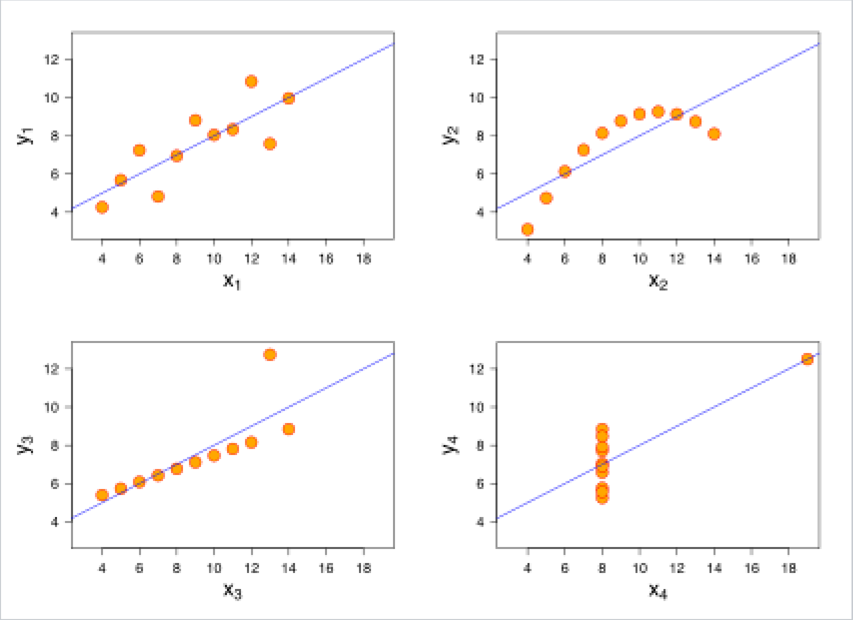
X = matrix of row vectors also known as explanatory vairables / input variables / predictor variables

β_0 *– intercept term*

β_1 , $\beta_2 .....\beta_k$ *– paritial derivates of the dependent variables with respect to the various indepdent variables.*

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

Answer: *(source: wikipedia)*



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

Anscombe's quartet *compireses of four data sets that have nearly identical simple descriptive statics, yet have very different distributions and appear very different when graphed. They were constructed in 1973 by the statistican Francis Anscombe to demonstrate both the importance of graphing data when analyzing it, and the effect of outliers and other influentials observations on statitical properties.*

1. First scater plot : simple linear relationship, corresponding to two variables correlated where y could be modelled as guassian with mean linearly dependent on x

2. Second scatter plot: while the relationship between the tow variables is obvious, it is not linear and the pearson correlation coefficient is not relevant

3. Third graph: modeeled relationship is linear , but should have different regression line. The calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816

4. Fourth graph: shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables

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

Answer:

The Pearson correlation coefficient (r) is ***the most common way of measuring a linear correlation****. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables. When one variable changes, the other variable changes in the same direction*

Pearson correlation coefficient (r) value	***Strength***	***Direction***
Greater than .5	*Strong*	*Positive*
Between .3 and .5	*Moderate*	*Positive*
Between 0 and .3	*Weak*	*Positive*
0	*None*	*None*
Between 0 and –.3	*Weak*	*Negative*
Between –.3 and –.5	*Moderate*	*Negative*
Less than –.5	*Strong*	*Negative*

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

Answer:

Scaling*: 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 is scaling performed*: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.*

Normalized scaling vs standardized scaling:

- *Normalization brings all of the data in the range of 0 and 1.*
- *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:

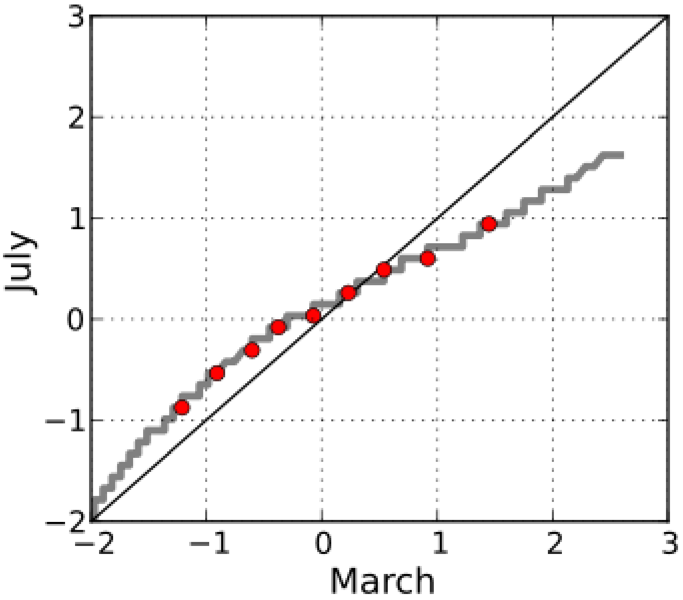
If there is a perfect correlation, then VIF = infinite

A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity.

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 plot of the quantiles of two distributions against each other, or a plot based on estimates of the quantiles. The pattern of points in the plot is used to compare the two distributions.*


July
March