Acknowledgement: Your acknowledgement for collaboration and other sources goes here.

Name: Badri Narayanan

Linear Regression - Assignment 2

From your analysis of the categorical variables from the data set, what could you infer about their effect on the dependent variable?

The categorical variables present in the data set are - season, weathersit, year, month, holiday. After our modelling and analysis, we have the following categorical variables that make up the final equation we derive at the end.

$demand = -0.098 \times holiday + 0.49 \times temp - 0.14 \times windspeed - 0.07 \times spring + 0.04 \times summer + 0.0871 \times winter - 0.0742 \times cloudy - 0.2849 \times lightrain + 0.2308 \times year + 0.2050$
The categorical variables holiday, spring, cloudy, light-rain have a negative impact the demand of bikes whereas summer, winter and year have a positive influence on the dependent variable.

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

If we have n levels or categories in our data, we need only n-1 columns of categories to represent any of the category. For instance, for 2 categories 'yes' and 'no' where yes means 1 and no means 0, we can remove the column 'yes', we can still represent the category 'yes', i.e. when 'no' = 0.
Hence, during dummy variable creation we can drop the first category as it might save us some space when the data gets larger.

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

From the heat-map that we have visualised in our analysis, we can see that temp(observed temperature) and atemp(feeling temperature) have the highest correlation with the target variable followed by the feature variable '2019'.

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

- Residual analysis on training error. Plot a histogram between y_train and y_train_cnt (which is the predicted value) and we can observe that they have zero mean and follow a normal distribution. The error terms also have constant variance and are independent of each other.
- We can see a linear relationship between y_test and y_pred as shown in the end of the

notebook. This means that we can draw a line that passes through the data points and is able to explain the data.

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

Looking at the coefficient values, we can see the top three features are:
1. temperature which positively affects demand of bikes
2. the year '2019' which also positively affects demand of bikes
3. light-rain which seems to negatively affect demand of bikes

General subjective questions

Explain the linear regression algorithm in detail.

Linear regression is a statistical technique that can be used to solve regression problems - estimating relationship between a target variable and the independent feature variables. For example, in our assignment, we tried to establish a relationship between demand(target variable) and other feature variables such as weather, temperature, holiday, etc . The following are some the steps that are followed in general for a linear regression problem.

1. let's assume that the validity of linear regression holds true. Hence, the linear relationship between two variables in case of simple linear regression would be
Y = beta1*X + C
where beta1 and C are the coefficient of the independent variable and the constant that has to be learnt by the model. The coefficient is also the slope of the line that can be drawn through the points and C is the intercept.

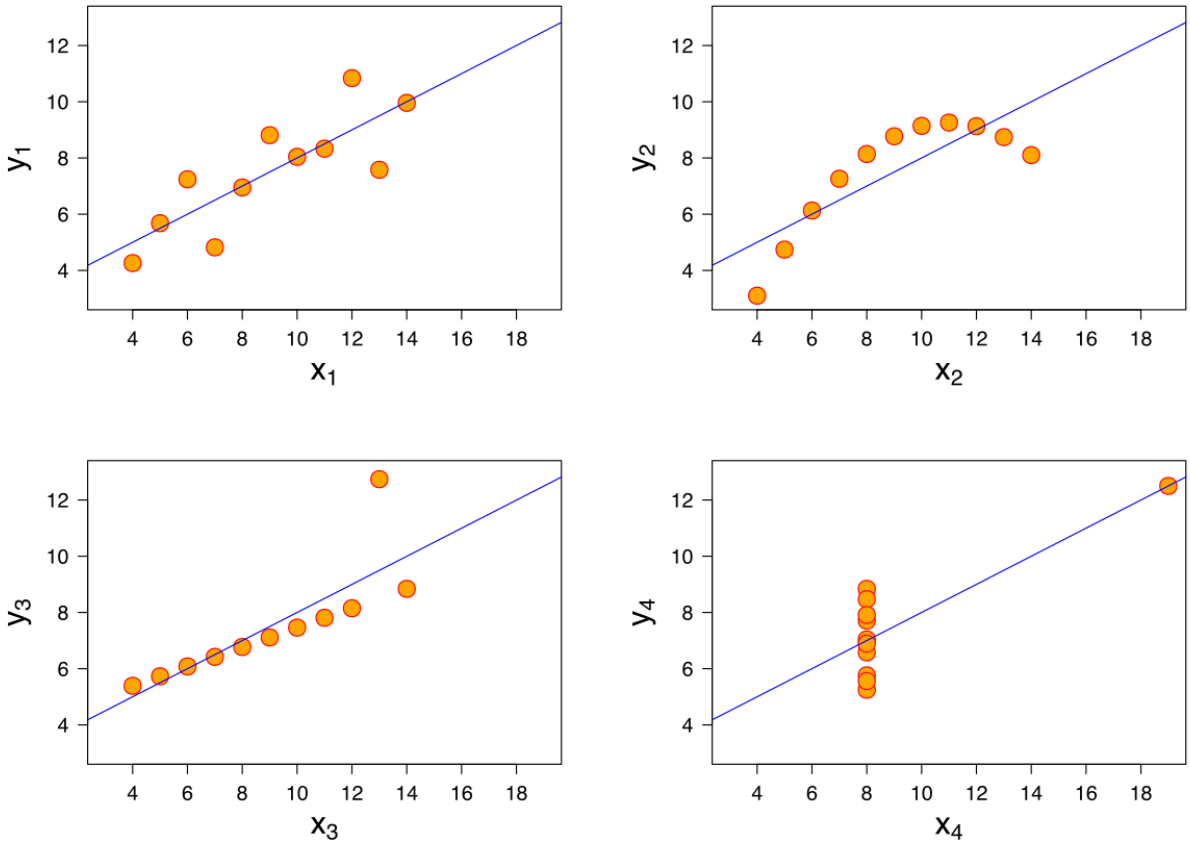
2. In order to fit a line, we need to see how good the measure of fit is. For this, we can use an algorithm like Ordinary Least Squares method, which tries to minimise the error between the predicted and observed values.

3. Calculate the coefficients after OLS and use hypothesis testing to check the significance of these coefficient values along with checking for correlation between variables in case of multiple linear regression.

4. Check for the validity of linear regression assumptions by calculating residuals and checking for linear relationship between predicted values and test values.

Explain the Anscombe's quartet in detail.

Anscombe's quartet contains of four different datasets that have similar statistical properties but appear very different when they are plotted. Let's take a look a them as shown below in the graphs.



For all 4 of them, the slope of the regression line is 0.500 (to three decimal places) and the intercept is 3.00 (to two decimal places). This shows that just by knowing the quantitative values, we might not be able to conclude anything. Visualising these values gives us more insight into the relationship between the variables, whether or not it has a linear or non linear relationship.

1. In the first graph, there seems to be a linear relationship between x and y.
2. In the second right graph, there seems to be a non-linear relationship between x and y.
3. In the bottom left graph, there is a linear relationship except for one outlier value.
4. The bottom right graph has one point at the end and other points are highly correlated.

What is Pearson's R?

Pearson's R, or also known as correlation coefficient is a measure of how much of a linear correlation there is between two sets of data. It is given by the formula:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

The values of correlation coefficient ranges from -1.0 to +1.0. If the value is positive and high, this means that increase in the value of one variable is likely to increase the value of other variable as well. When the value is negative and high, this means that the increase in value of one variable is likely to result in decrease in value of other variable.
Calculating correlation coefficient allows to to determine relationship between variables and make use of them during feature selection process to drop any, if necessary.

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

Scaling is a technique which we can use to bring down variables which fall under multiple range of values to come under a given interval of values. This gives us data which falls under a certain scale and this is extremely useful for two major reasons:
1. If we don't have comparable scales, some of the coefficients might be very large or small compared to the others. This might prove to be a problem during model evaluation.
2. Scaling also helps with the optimisation step. During gradient descent, the gradients converge much faster than when the values are not scaled.

We have two major methods of scaling:
1. min-max scaling
2. standardisation

Normalised scaling or min-max scaling squishes the value between 0 to 1 whereas standardised scaling does not restrict the value to a particular range. The value of standardised value is obtained by subtracting the mean from the value and dividing it by the standard deviation.

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

If there is a VIF value of infinity, this means that there is a very high correlation. In case of almost perfect correlation, let's assume that $R\hat{2}$=1. Which means that 1/(1-$R\hat{2}$) = infinity. To solve this, we can drop some of the variables based on multicollinearity and build our model again to rectify the problem of high VIF.

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

Q-Q or quantile-quantile plots show us the plot of quantiles of a sample distribution against that of a theoretical distribution. Doing so helps us determine if the given dataset follows a particular distribution such as normal, uniform, exponential, etc.
In our case of linear regression, they can also be used to see if the residuals follow some sort of a normal distribution pattern or not and if the error terms are centred around 0. Below shown graph is one where both the sample and theoretical data have the same kind of distribution.