LINEAR REGRESSION ASSIGNMENT

Assignment-based Subjective Questions

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

Final Equation:- cnt = 0.2537*Yr(Latest year) + 0.0630 *September + 0.0553*Sunday + 0.0482*workingday - 0.0402*summer - 0.0780*MistCloudy - 0.0872*winter - 0.2035*windspeed - 0.2947*spring + 0.5338 (intercept) $

Out of the list of 9 variables which are part of my final model(equation), windspeed is the only numerical variable and rest all are categorical types. Yr and workingday might be numerical flags but they don't have numerical significance. These along with the other categorical variables were treated separately by creating dummy columns. Month, weather and season categorical variables have shown strong influence on the the dependent/target variables. Hence it can clearly seen that overall the categorical variables were the key independent variables in my predictive model.

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

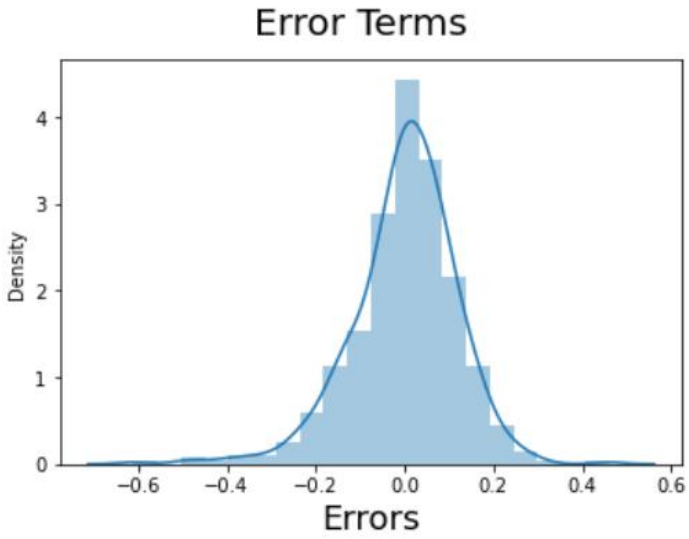
It helps us to have one column less while creating dummy columns for each of the categorical columns. N unique values will have N-1 dummy columns because of this. And when all the dummy columns have 0 as their values it means the Nth value is being portrayed.

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

As per the pair-plot, the atemp and temp columns appeared to have the highest correlation (Ignoring the casual and registered columns). And later on it was seen that the heatmap of correlation of the columns showed atemp having highest correlation 0.65, closely followed by temp column 0.64.

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

To validate the assumptions if the error terms are also normally distributed a histogram of the error terms was plotted.



The graph validated our assumptions.

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

In terms of +ve influence Yr (0.2537), September (0.0603) & Sunday (0.0553).

In terms of -ve influence spring (-0.2947), windspeed (-0.2035) & winter (0.0872).

In terms of absolute values Spring, Yr and windspeed are the top 3 contributing features explaining the demand of the shared bikes

General Subjective Questions

Q: Explain the linear regression model in detail:

Linear regression is one of the most predominantly used predictive analysis which shows a linear relationship between a dependent variable and one or more independent variable. The logic behind the linear regression model is to examine 2 things

1) Does a set of predictor variables do a good job in predicting an outcome (dependent) variable?
2) Which variables in particular are significant predictors of the outcome variable and in what way do they–indicated by the magnitude and sign of the beta estimates–impact the outcome variable?

The simplest form of the regression equation with one dependent and one independent variable is defined by the formula y = c + b*x, where y = estimated dependent variable score, c = constant, b = regression coefficient, and x = score on the independent variable.

Similarly in case of multiple linear regression model we have more than 1 independent variable and with one dependent/target variable which is defined by the formula y = c + b1*x1 + b2*x2 + b3* x3 ….. bn* xn , where y = estimated dependent variable score, c = constant and b1,b2,b3…bn are the regression coefficients, and x1,x2,x3 ….. xn are the scores of the independent variables.

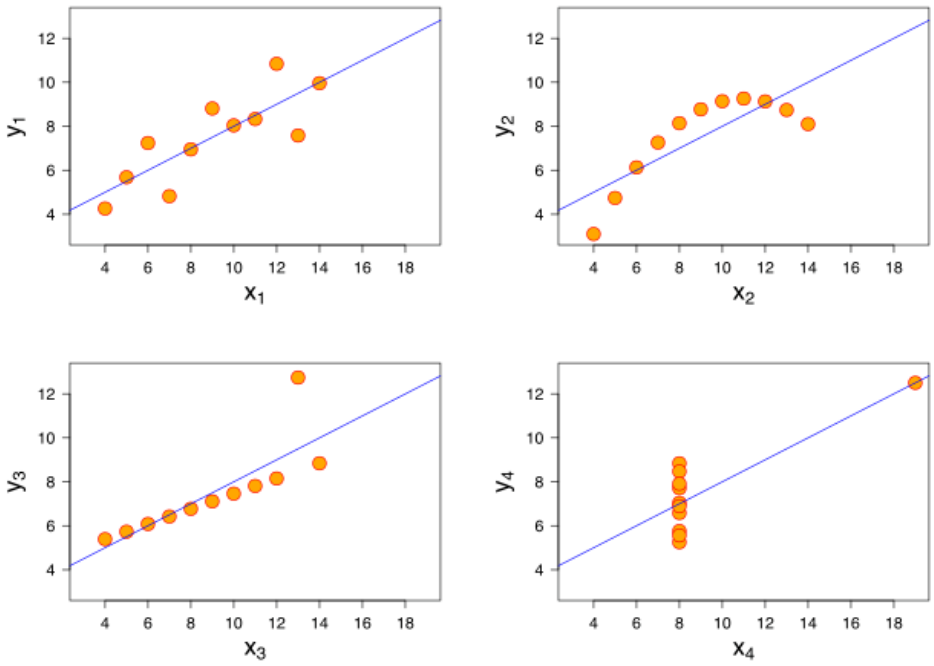
Linear regression analysis is based on a few assumptions about the data. There are:

- Homogeneity of variance, meaning the size of the error in the prediction, does not change significantly across different values of the independent variable.
- Independence of observations in the dataset, referring to no hidden relationships.
- Normality of data distribution for the dependent variable. You can check the same using the hist() function in R.

Q: Explain the Anscombe's quartet in detail:

Anscombe's quartet, constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting the graphs before analyzing and model building, and the effect of other observations on statistical properties is a group of four datasets that are nearly identical in simple descriptive statistics, which provides same statistical information that involves variance, and mean of all x, y points in all four datasets.

For Example, in the below graph all the four sets are identical statistically but vary considerably when graphed.



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. Anscombe's quartet helps us to understand the importance of data visualization and how easy it is to fool a regression algorithm.

Q: What is Pearson's R?

The Pearson's Correlation Coefficient is also referred to as Pearson's r, named after Karl Pearson, the Pearson product-moment correlation coefficient (PPMCC), or bi-variate correlation. It is a statistic that measures the linear correlation between two variables. Like all correlations, it also has a numerical value that lies between -1.0 and +1.0.
It is the covariance of the two variables divided by the product of their standard deviations. The form of the definition involves a "product moment", that is, the mean (the first moment about the origin) of the product of the mean-adjusted random variables; hence the modifier product-moment in the name.

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

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

The Pearson's correlation coefficient varies between -1 and +1 where:
r = 1 means the data is perfectly linear with a positive slope (i.e. both variables tend to change in the same direction)
r = -1 means the data is perfectly linear with a negative slope (i.e. both variables tend to change in different directions)
r = 0 means there is no linear association
0 < r <0. 5 means there is a weak association
0.5 < r <0. 8 means there is a moderate association
r > 0.8 means there is a strong association

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

Scaling is a technique to standardize the columns present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values. It improves in the interpretation of the weights of the feature variables.

Normalization

1) Known as min-max scaling or min-max normalization, it is the simplest method and consists of rescaling the range of features to scale the range in [0, 1].
2) The general formula is x'= (x- min(x))/(max(x)- min(x))
3) Helpful when the distribution of data does not follow a Gaussian distribution.
   This can be useful in algorithms that do not assume any distribution of the data like K-Nearest Neighbors and Neural Networks
4) It has bounding ranges.

Standardization

1) Feature standardization makes the values of each feature in the data have zero mean and unit variance. The general method of calculation is to determine the distribution mean and standard deviation for each feature and calculate the new data point
2) The general formula is
   x'= (x- μ)/ σ where μ is the mean of the feature values and σ is the standard deviation of the feature values.
3) Helpful in cases where the data follows a Gaussian distribution.
4) It does not have bounding ranges.

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

If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between the independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To resolve this we can drop one of the variables from the dataset which is causing this perfect multi collinearity.

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 as well).

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

The quantile-quantile plot is a graphical method for determining whether two samples of data came from the same population or not. A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. Quantile refers to a fraction (or percent) of points below a given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value. A 45 degree angle is plotted on the Q-Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.

If the two distributions being compared are similar, 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. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.

QQ plots are very useful to determine:

1) If two populations are of the same distribution
2) If residuals follow a normal distribution. Having a normal error term is an assumption in regression and we can verify if it's met using this.
3) Skewness of distributions
4) Compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.