Assignment-based Subjective Questions

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

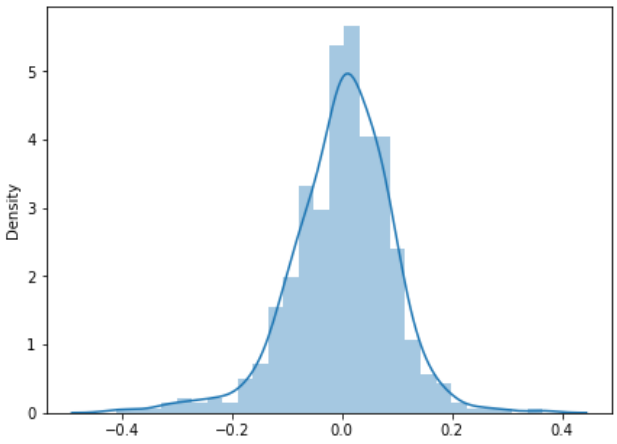
It depends on the model. If you don't drop the first column then your dummy variables will be correlated (redundant as Dimitre shows in the post below). This may affect some models adversely and the effect is stronger when the cardinality is smaller. For example, iterative models may have trouble converging and lists of variable importances may be distorted.
For example, if you have a variable gender, you don't need both a male and female dummy. Just one will be fine. If male=1 then the person is a male and if male=0 then the person is female. However if you have a category with hundreds of values, it's not suggested dropping the first column.
For n categories, if you know the result of n-1 of them then you can easily predict the result of nth category. So it is safe to drop in case of multiple categories also.

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

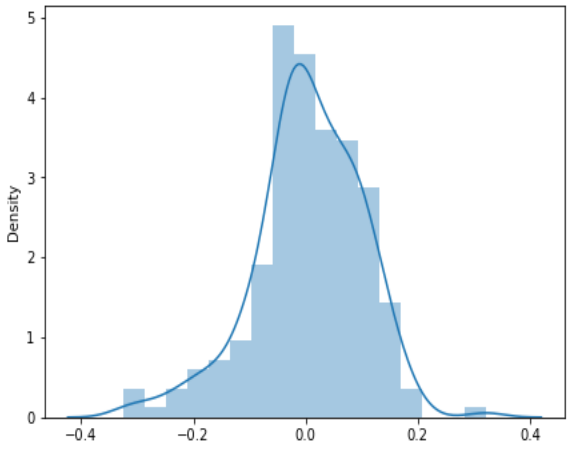
Atemp and temp variables has the highest correlation (0.63)

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

- There is linear relationship between X(features) and Y(dependant variable):
  The temp and windspeed has linear relationship with count.
- Error terms are normally distributed:

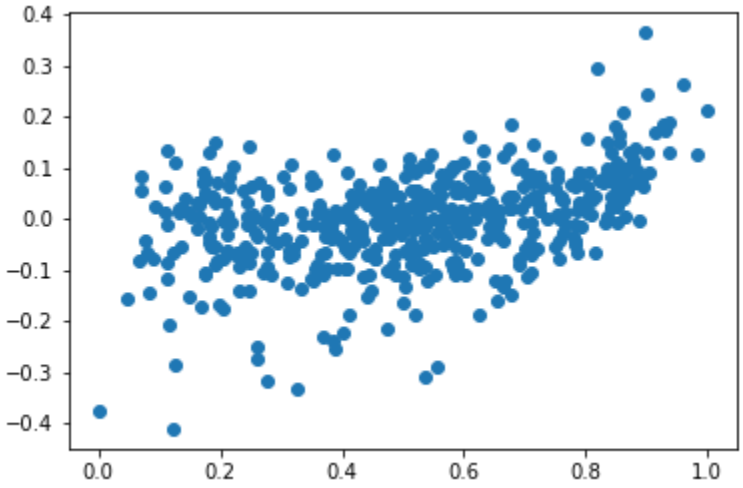


Residual of X_train data



Residual of X_test data

- checking whether error terms have constant variance(homoscedasticity):



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

1) Wet bulb temperature of simply temperature = feeling temperature in Celsius
2) Windspeed = wind speed
3) Lightrain = Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds

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

- Presence of rain affects the most in a negative sense to the total rental bikes being used.
- The year column: since these bike-sharing systems are slowly gaining popularity, the demand for these bikes is increasing every year.
- September has strong positive correlation, as The USA enjoys a beautiful climate in September with plenty of warm sunshine and clear blue skies for most of the month.

The equation look like this

$$Totalbikes = 0.4746 \times atemp + 0.10546 \times September + 0.0692 \times summer + 0.0947 \times winter + 0.2347 \times Year - 0.0920 \times holiday - 0.13536 \times windspeed - 0.0553 \times spring - 0.0816 \times cloudy - 0.2792 \times lightrain + 0.0497 \times august + 0.1964 \times const$$

General Subjective Questions

1.Explain the linear regression algorithm in detail.

Linear regression may be defined as the model that analyzes the linear relationship between a dependent variable with a given set of independent variables. Linear relationship between variables means that when the value of one or more independent variables will change (increase or decrease), the value of dependent variable will also change accordingly (increase or decrease).

Mathematically the relationship can be represented with the help of following equation -

Y = mX + C

Here, Y is the dependent variable we are trying to predict

X is the dependent variable we are using to make predictions.

m is the slope of the regression line which represents the effect X has on Y

b is a constant, known as the Y-intercept. If X = 0,Y would be equal to C.

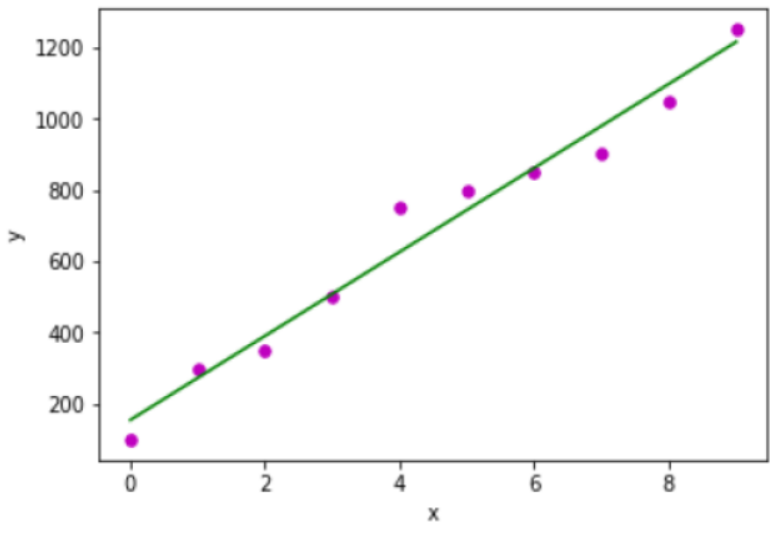
Linear regression is of the following two types −

- Simple Linear Regression
- Multiple Linear Regression

1) Simple linear regression

It is the most basic version of linear regression which predicts a response using a single feature. The assumption in SLR is that the two variables are linearly related.

Looks like this



2) multiple linear regression

It is the extension of simple linear regression that predicts a response using two or more features. Mathematically we can explain it as follows −

Consider a dataset having n observations, p features i.e. independent variables and y as one response i.e. dependent variable the regression line for p features can be calculated as follows −

$$h(x_i) = b_0 + b_1 x_{i1} + b_2 x_{i2} + \ldots + b_p x_{ip}$$

Here, h(xi) is the predicted response value and b0,b1,b2…,bp are the regression coefficients.

Multiple Linear Regression models always includes the errors in the data known as residual error which changes the calculation as follows −

$$h(x_i) = b_0 + b_1 x_{i1} + b_2 x_{i2} + \ldots + b_p x_{ip} + e_i$$

We can also write the above equation as follows −

$$y_i = h(x_i) + e_i \; or \; e_i = y_i - h(x_i)$$

Assumptions:

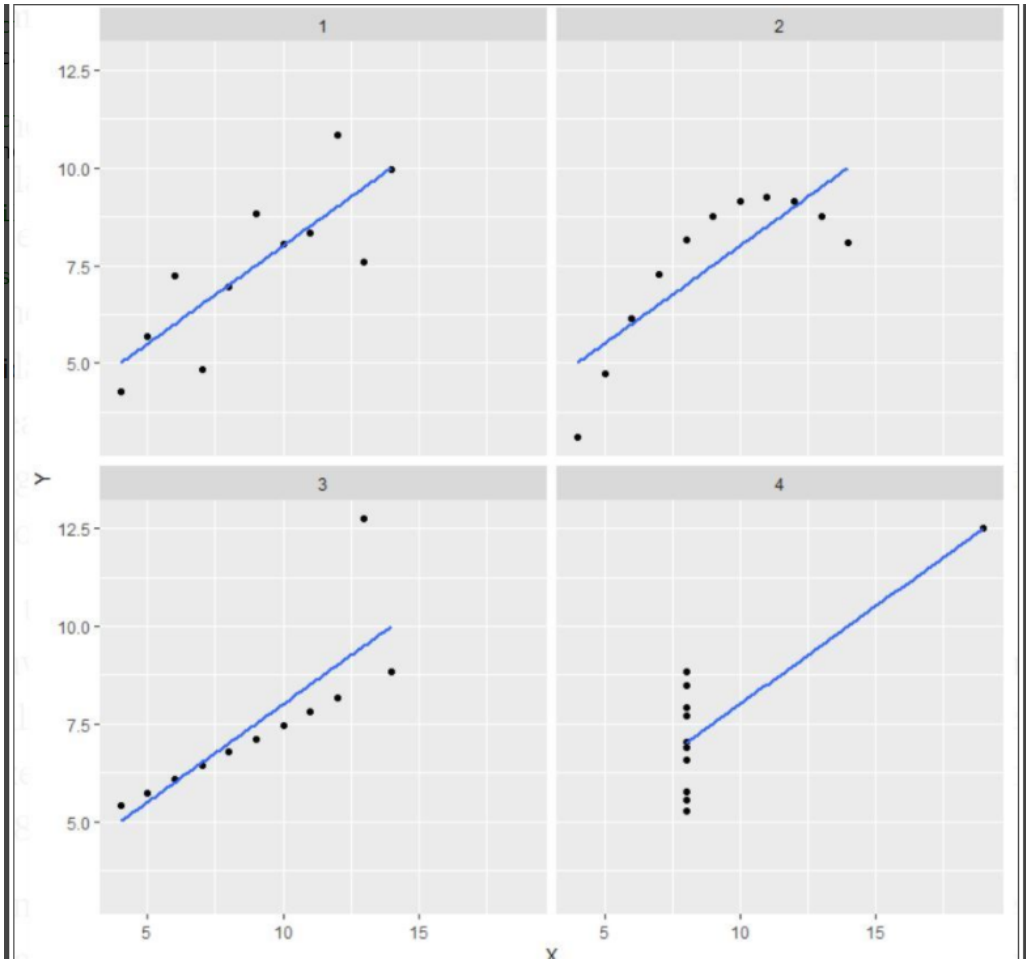
- Multicollinearity − Linear regression model assumes that there is very little or no multicollinearity in the data. Basically, multicollinearity occurs when the independent variables or features have dependency in them.
- Auto-correlation − Another assumption Linear regression model assumes is that there is very little or no auto-correlation in the data. Basically, auto-correlation occurs when there is dependency between residual errors.
- Relationship between variables − Linear regression model assumes that the relationship between response and feature variables must be linear.

2.Explain the Anscombe's quartet in detail.

According to the definition given in Wikipedia, Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers on statistical properties.

Once Francis John “Frank” Anscombe who was a statistician of great repute found 4 sets of 11 data-points in his dream and requested the council as his last wish to plot those points.
After that, the council analyzed them using only descriptive statistics and found the mean, standard deviation, and correlation between x and y.

The results were like this



Explanation:

- In the first one(top left) if you look at the scatter plot you will see that there seems to be a linear relationship between x and y.
- In the second one(top right) if you look at this figure you can conclude that there is a non-linear relationship between x and y.

- In the third one(bottom left) you can say when there is a perfect linear relationship for all the data points except one which seems to be an outlier which is indicated to be far away from that line.
- Finally, the fourth one(bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient.

Applications:

The quartet is still often used to illustrate the importance of looking at a set of data graphically before starting to analyze according to a particular type of relationship, and the inadequacy of basic statistical properties for describing realistic datasets.

3.What is Pearson's R?

Correlation is a bi-variate analysis that measures the strength of association between two variables and the direction of the relationship. In terms of the strength of relationship, the value of the correlation coefficient varies between +1 and -1. A value of ± 1 indicates a perfect degree of association between the two variables. As the correlation coefficient value goes towards 0, the relationship between the two variables will be weaker. The direction of the relationship is indicated by the sign of the coefficient; a + sign indicates a positive relationship and a - sign indicates a negative relationship.
Pearson's R:
As the title suggests, we'll only cover Pearson correlation coefficient. Pearson correlation coefficient is a measure of the strength of a linear association between two variables — denoted by r. You'll come across Pearson r correlation.
Assumptions:

- For the Pearson r correlation, both variables should be normally distributed. i.e the normal distribution describes how the values of a variable are distributed. This is sometimes called the 'Bell Curve' or the 'Gaussian Curve'.
- There should be no significant outliers
- The two variables have a linear relationship.
- The observations are paired observations. That is, for every observation of the independent variable, there must be a corresponding observation of the dependent variable.
- Homoscedasticity.

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

As its name says, scales the features of our data so that they all have a similar range.
It is performed as the algorithms find it easy to calculate the data if all the features are in similar range, also by interpretation point of view,if not done then the coefficient of larger scaled features will be very small and smaller range will be very large, so that may mislead us believing the 2nd one as more effect on our output variable.

Two major methods are employed to scale the variables: standardization and MinMax scaling. Standardization brings all the data into a standard normal distribution with mean 0 and standard deviation 1. MinMax scaling, on the other hand, brings all the data in the range of 0-1. The formulae used in the background for each of these methods are as given below:

- Standardisation: $x = \frac{x - mean(x)}{sd(x)}$
- MinMax Scaling: $x = \frac{x - min(x)}{max(x) - min(x)}$

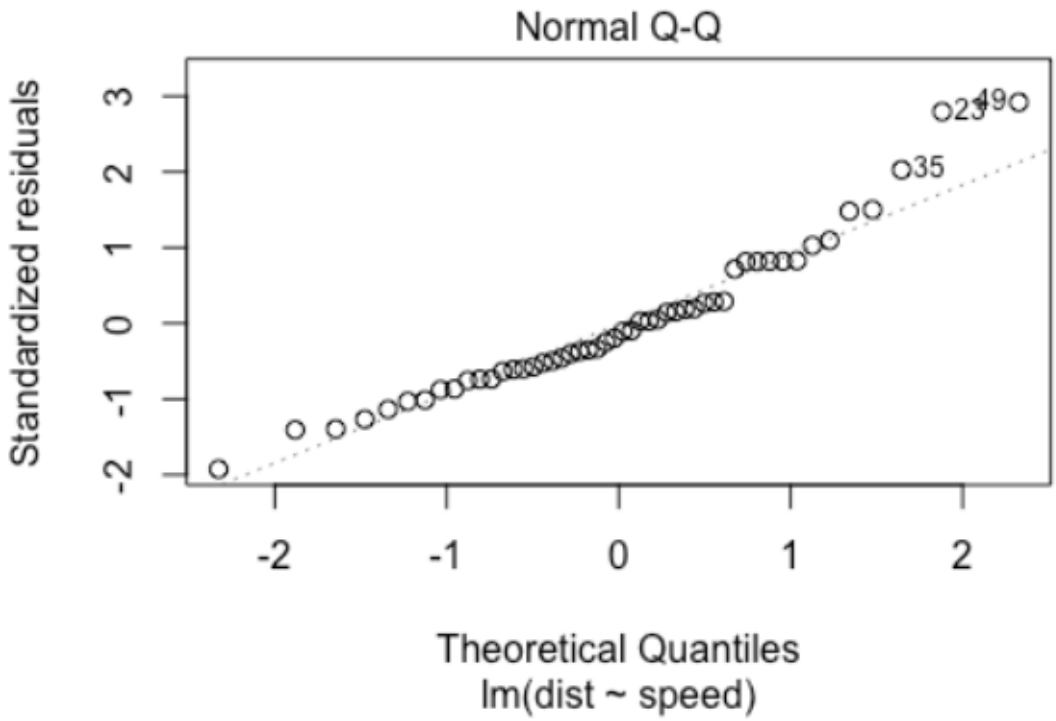
5.You might have observed that sometimes the value of VIF is infinite. Why does this happen?

In a single line, it represents the perfect correlation.
The case of perfect correlation, we get R2 =1, which leads to 1/(1-R2) infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.
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).

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

Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q Q plots is to find out if two sets of data come from the same distribution. 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.
A Q–Q plot is used to 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.



A Q Q plot showing the 45 degree reference line: