Subjective Questions

Answers

Santosh K U - 15/3/2022

Assignment based questions:

Question 1:

- season : counts is similar for summer, winter and fall, counts sort of dips for spring
- Yr : There is good increase in counts in 2019 compared to 2018.
- mnth :months also provide similar inference as season, there is dip in number of the months in spring.
- holiday : Clearly there is huge increase in numbers during holidays.
- weekday : count looks the smae on all weekdays.
- workingday : sort of give a picture that when in holidays and working days, the count is almost similar, strange!
- weathersit : clearly the count is less on a rainy/snow day

Assignment based questions:

Question 2:

- For categorical variables with n categories, it is important to encode these categories.

- As n-1 categories will implicitly represent all n categories of a variable, With encoding the categories can be optimised to n-1.

- By using drop_first=True, Encoding always ensures that first category for the variable will be dropped out of n categories.

- By reducing one column of each categories the correlations will also be reduced, this helps the model building.

Assignment based questions:

Question 3:

- ‘atemp and temp’ has the highest correlation with target variable ‘cnt’ with a value of 0.63

Assignment based questions:

Question 4:

- First, we plotted a pair pot to identify if the dependent variable has linear relation with the independent variables.
- With the error plots the variance was checked for homoscedasticity, since the error plot is normally distributed with zero mean and constant variance, the model is not impacted.
- Multicollinearity was checked by plotting heat maps consisting the correlation between all the variables. Variables with high correlation was removed. Later VIF values were also checked to tune the model better by removing variables with VIF(high correlation).
- If the errors are not normally distributed the model could be biased, this point was checked by plotting the errors. We found that the errors are normally distributed.

Assignment based questions:

Question 5:

- mnth_9 - with factor 0.39
- mnth_8 and mnth_6 - with factor 0.35
- mnth_5 and mnth_7 - with factor 0.32

General Questions :

Question 1:

- Linear regression is a supervised learning, where we build a regression model to predict an outcome based on independent variables.

- With this model, we find the relation between the dependent and independent variable, with this relation we can predict the result based on the independent variables.

- Based on the number of independent variable the linear regression model is classified as simple and multi linear regression. As the name suggests simple linear regression has one independent variable and multi linear regression has more than one independent variable.

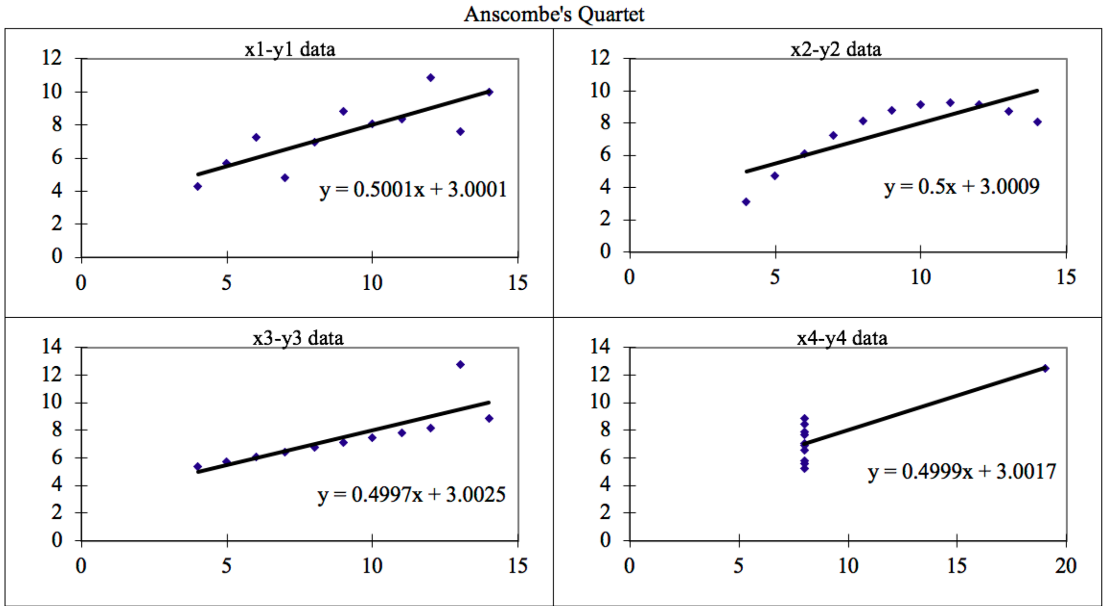
- Simple linear regression - y = mx + c , y - Dependent variable, x - Independent variable, m - slope/coefficient of x, c - constant

- Multi linear regression - y = m1 x1 + m2 x2 + ……+ mn xn + c + e, y - Dependent variable, x1,x2…xn - Independent variables, m1,m2…mn - slope/coefficients of x, c - constant, e - error considered since a sample if considered for analysis.

General Questions :

Question 2 :

- Anscombe's quadtet consists of four datasets which have similar statistical features but totally different distributions, when each dataset plotted with the output variable, the data has different distributions, please find the details below:



General Questions :

Question 2 :

- x1-y1 : The disctribution is linear.
- X2-y2 : There is a non-linear relation between the variables.
- X3-y3 : There is a linear relation with between the variables, but there is a outlier, the regression line has to change when compare with x1-y1.
- X4-y4 : There is no linear relation between the varibles, but since there is a outlier at higher side, it establishes significance with the linear regression, which is totally different when looked in the plot.
- With this an awareness is provided on how linear regression impacted with outliers and special observations.

General Questions :

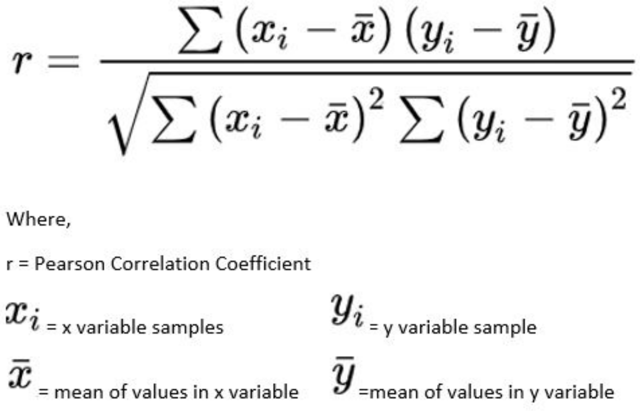
Question 3 :

- Pearson’s R, this is the measure of linear correlation between the variables. The values ranges between -1 to 1.

- If the value is 1, it means that the variables have positive linear correlation.

- If the value is -1, it means that the variables have negative linear correlation.

- If the value is 0, it means that the variables have no linear correlation.



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

Where,

r = Pearson Correlation Coefficient

x_i = x variable samples y_i = y variable sample

$\bar{x}$ = mean of values in x variable $\bar{y}$ =mean of values in y variable

Pearson’s R formula

General Questions :

Question 4 :

- Scaling is method to bring all the variables to have a common range.
  This is an important step for in linear regression, without scaling the the coefficients of the independent variables will either be very high/very low.

- With coefficients varying too much it is difficult to get inference from the model.

- Normalisation - Normalisation is technique to bring the data within the range 0 and 1
  MinMaxScaling : x = (x-min(x))/(max(x) - min(x))

- Standardisation - Standardisation replaces the values with Z score, this brings all the data points into standard normal distribution.
  x = x-mean(x)/sd(x)

General Questions :

Question 5 :

- VIF provides estimate of the variance involved in estimation of coefficients of the independent variables in linear regression.

- The formula of VIF is
  VIF = 1 / (1 - R^2)

- VIF will be infinity when R^2 = 1, this means RSS is zero.

- RSS is sum of square of residuals, if RSS is zero it means the predicted and actual value are the same. Hence the residuals are zero.

- This happens when there is a perfect correlation between the independent variable and dependent variable.

General Questions :

Question 6 :

- Q-Q plot are used when train and test data are from different datasets. It helps us understand if the datasets have a common distribution.

- Q-Q plots are plots of each quantile of the datasets against each other, if the plot fits on/parallel it indicates that the datasets are from same distributions.

- With Q-Q plot we can get inferences on data distribution, location, scale, distributional shapes and tail behaviour.