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:
   - In Fall season bike rental is more compared to other seasons, and spring season has the least rental.
   - Bike rental in year 2019 is much better than year 2018.
   - There is no much difference in working day and non-working day.
   - When weather is clear then rental is more compared to other weather types.
   - Jan and Feb month has less rental.

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

   Answer: If we set drop_first=True then get_dummies method will return k-1 columns of categorical variable. Where k is total number of categorical values in a single variable.

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

   Answer: registered variable has the highest correlationwith the 'cnt'(target) variable

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

   Answer: I have checked by:
   - Plotting a distplot for y train pred which is derived by subtracting y_train_pred from y_train. And the distplot was showing a normal distribution.
   - All the variables p-value are below 5% and VIF is below 5.

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: Top 3 features contributing variables in decreasing order are:
   - temp
   - yr
   - winter

General Subjective Questions

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

Answer:

Linear regression is used for predictive analysis. With help of Linear regression, we get to know the correlation between continuous variables. In Linear regression X-axis denotes independent variable(s) and Y-axis denotes dependent variable.

To calculate best line fit for simple linear regression we use below equation:

$$y = \beta_0 + \beta_1.X$$

Where:

β_0 -> denotes intercept of line

β_1 -> denotes slope of line or linear regression' coefficient

y -> denotes dependent variable

X -> denotes independent variable

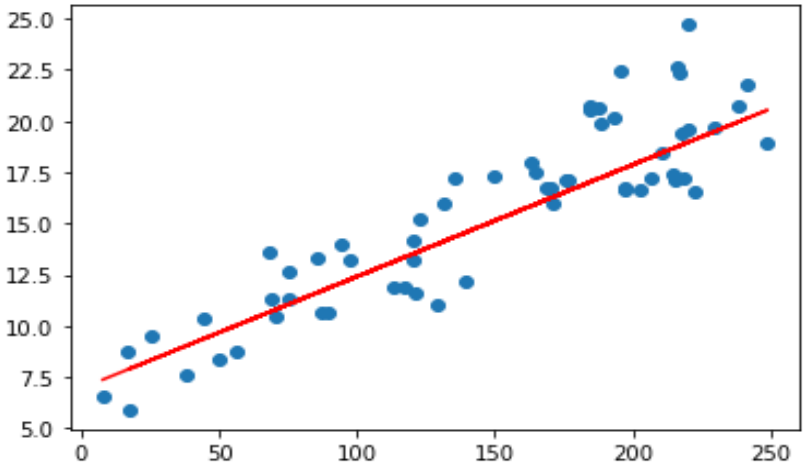


Fig 1: Positive correlation

If dependent variable's(y-axis) values increase and so of independent variables(X-axis) than it is known as positive correlation.

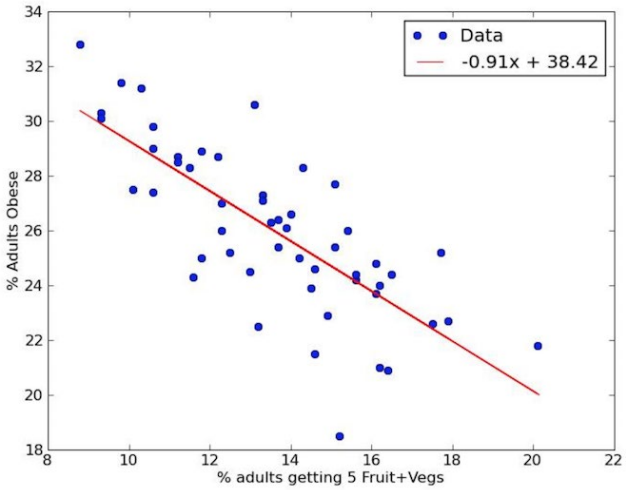


Fig 2: Negative correlation

If dependent variables(y-axis) values increases and independent variables(X-axis) values decreases than it is known as negative correlation

From above equation we get to know that in simple linear regression we can do predictive analysis for two continuous variables only. What if we have multiple independent variables which is more than one variable in such case, we use multiple linear regression.
To find best fit for multiple linear regression we use below equation:

$$y = \beta_0 + \beta_1.X_1 + \beta_2.X_2 + \beta_3.X_3 \ldots\ldots + \beta_n.X_n$$

Where:
n - > denotes no. of variables
β_0 -> denotes intercept of line
β_n -> denotes slope of line or linear regression' coefficient of X_n
y -> denotes dependent variable
X_n -> denotes independent variable

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

Answer:

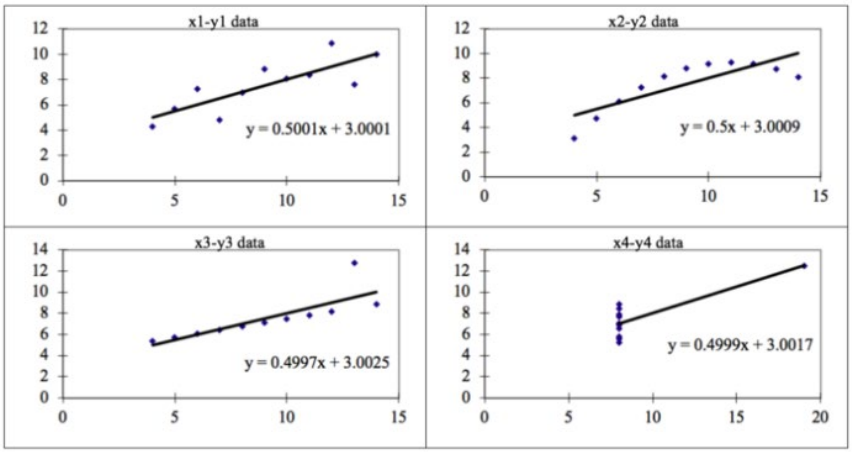
Anscombe's quartet explains that we should not just limit our self just on a data statistics value i.e., not just relay on mean, standard deviation of various variables. At first glance the values may look similar but actually they are not depicting what is visible to us. Anscombe's quartet say's once we are done with statistics analysis then we should also perform data visualization, by doing so we will be able to clearly understand the correlation of the data set's variables.

For example:

In below fig we can see that there is total 4 data sets, and their values look similar so we may think that they are similar and on graph (scatter plot) they will look similar.

Anscombe's Data									
Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
8	4	4.26		4	3.1		4	5.39		19	12.5
9	12	10.84		12	9.13		12	8.15		8	5.56
10	7	4.82		7	7.26		7	6.42		8	7.91
11	5	5.68		5	4.74		5	5.73		8	6.89
				Summary Statistics							
N	11	11		11	11		11	11		11	11
mean	9.00	7.50		9.00	7.500909		9.00	7.50		9.00	7.50
SD	3.16	1.94		3.16	1.94		3.16	1.94		3.16	1.94
r	0.82			0.82			0.82			0.82	

But when we plot above data set in scatter plot than we get to know the actual correlation between variables. Below image depicts the data visualization.



From above fig we get to know that if the statistics values may look same but the correlation may not be same. From above figure first plot shows that line fits for linear regression model and for remaining plots linear regression can not be handled because 3rd and 4th plot have outliers and 2nd plot is not fit for linear regression.

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

Answer:

- Pearson's R is also known as Pearson Correlation Coefficient (PCC).
- Pearson's R determines the strength of association, The stronger the association of the two variables, the closer the Pearson's R
- Pearson's R will be to either +1 or -1 depending on whether the relationship is positive or negative, respectively
- Pearson's R does not represent the slope of the line of best fit. Therefore, if you get a Pearson correlation coefficient of +1 this does not mean that for every unit increase in one variable there is a unit increase in another.

$$\rho = Cov(x,y)/ \sigma_x . \sigma_y$$

where:
Cov(x,y) is covariance of x and y
σ_x is variance of x variable
σ_y is variance of y variable

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

Answer:

Scaling is method where we scale down an independent variable to a limit so that it can be easy to find correlations and the execution also reduces because of scaling. Sometimes independents variables are of different units such as variable A's unit is kilometers and variable B's unit is in meters so there might be difficulty in measuring correlation and execution time may increase. So, to resolve this we use scaling, and scaling is done while we are processing data.

There are two types of scaling

1. Normalized scaling
2. Standardized scaling

Normalized scaling

It is also known as min max scaling/normalization. Here variables are scaled in range 0 to 1.

Formula is:

$$X' = (X - X_{min})/(X_{max}-X_{min})$$

Where:

X is current value which is to be scaled
X_{min} is lowest values in that column
X_{max} is greatest values in that column

Standardized scaling

Standardization is another scaling technique where the values are centered around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation. In standardization outlier in data will not be affected as in normalized. They still will be outlier but in scaled value.

$$X' = (X- \mu)/ \sigma$$

Where:

μ is mean of variable, σ is deviation and X is value which is to be scaled.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

Answer:
VIF stands for Variable Inflation Factors, determines the strength of the correlation between the independent variables.
To calculate VIF below equation is used:

$$VIF = 1/1-R^2$$

Where R^2 represent how well an independent variable is described by the other independent variables.

If R^2 value is exactly 1 than the VIF value will be infinite.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

Answer:
Q-Q plot is known as Quantile-Quantile plot, it is a graphical method of knowing whether samples of data came from the same population or not. Q-Q plot is very useful to determine:

- If two populations are of the same distribution
- 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.
- Skewness of distribution