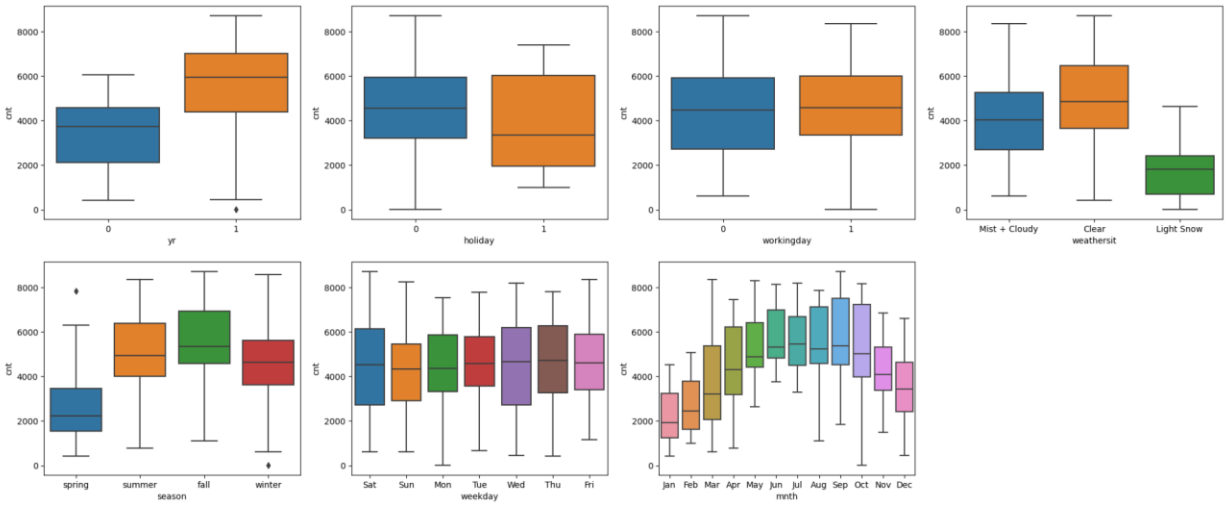
Linear Regression Subjective Questions

By Himanshu Dongre

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)

   Ans:



   Insights:

   1. Bike rentals seem to be more popular when weather is clear and seem to decrease as weather gets worse. Weathersit looks like a good indicator for predicting bike rentals.
   2. Bike rentals seem to be doing better in 2019 as compared to 2018.
   3. Surprisingly, bike rentals seem to be lower on holidays. Most people may be using bikes for commuting to work.
   4. Bike rentals seem to pick up during months 5,6,7,8,9 and then taper off during the end and beginning of the year. Looks like a good indicator for predicting bike rentals.
   5. Bike rentals also seems to be the highest for season 3 (fall) followed by season 2 (summer). Looks like a good indicator for predicting bike rentals.
   6. Weekday doesn't seem to show much change and may not be useful for predicting bike rentals.

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

   Ans:

   During the creation of dummy variable, new columns are created for each unique value in the original categorical variable, which are then assigned a binary value. However, to represent n categories in a binary format, we do not require all n variables. The information in the table can be explained even with n-1 variables. Therefore, to reduce redundancy and multi-collinearity we drop the 1st column using drop_first=True.

Example: Weathersit variable has 3 categories: Clear, Mist Cloudy and Light Snow.

These 3 categories can be represented using 2 columns:

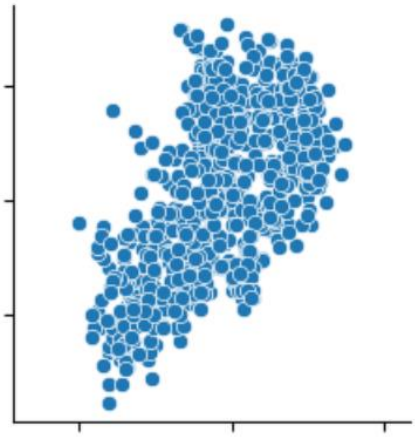
0 0 -> clear

0 1 -> Mist Cloudy

1 0 -> Light Snow

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

   Ans:



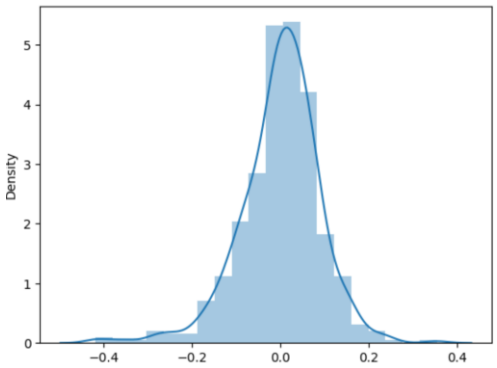
   Temp variable seems to have the highest correlation with cnt(target variable)

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

   Ans:

   Assumptions of Linear Regression are as follows:

   - Error terms are normally distributed with zero mean, should be independent and have constant variance (homoscedasticity):
     - This was validated by plotting the residual error



- There should be a linear relationship with some of the variables in the dataset and cnt(target variable):
  - This was validated by analyzing pair plots of the numeric variables.
- Model is not created by chance
  - This was validated by looking at the probability of F-statistic
- Model is not overfitting the dataset
  - This was validated by looking at the train and test dataset R2 and Adjusted R2 scores.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

   Ans:

   top 3 features contributing significantly towards explaining the demand of the shared bikes is as follows:

   - Temp: 0.4917
   - Light Snow: -0.2847
   - Yr: 0.2339

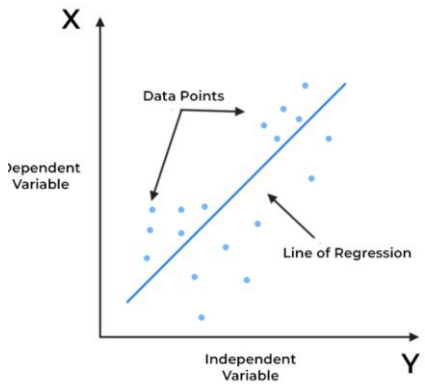
General Subjective Questions

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

   Ans:

   Linear Regression: It is a statistical technique that attempts to explore and model the relationship between two or more variables using a straight line. It is used to predict the value of a variable based on the value of another variable. The variable that we want to predict is known as the dependent variable and the variable/s that we use to predict the dependent variable is called independent variable/s.

   In this analysis we estimate the coefficients of the linear equation, involving one or more independent variables that best predict the value of the dependent variable. It predicts a straight line (1 variable) or surface (multiple variables) that minimizes the error between the true value and precited value given ground truth data.



In General, there are two types of linear regression methods:

1. Simple Linear Regression: Where there is one target variable and one independent variable.
   It takes the form of β0+β1X

   Where β1 is the slope of the line and β0 is the intercept
2. Multiple Linear Regression: Where the number of independent variables is more than one.
   It takes the form of β0+β1X +β2X2+β3X3 …

   Where β0 is the intercept and β1, β2, β3… are coefficients pf X1, X2, X3 respectively.

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

   Ans:

   Anscombe's quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but have very different distributions and can appear very different when plotted on a graph. It can easily fool a regression model if built.

   It was constructed in 1973 by a statistician Francis Anscombe to illustrate the importance of plotting graphs before analyzing and building models, and the effect of other observations on statistical properties. Each dataset consists of eleven (x,y) points.

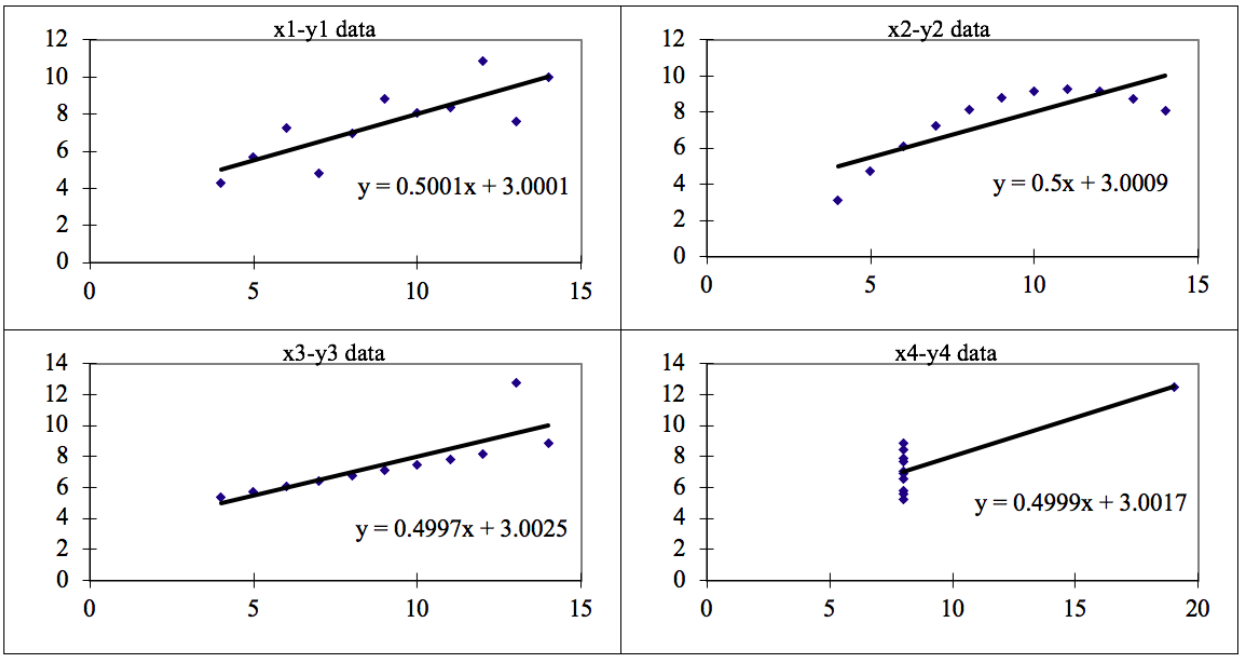
   These four plots can be defined as follows:

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

The statistical information for all these four datasets is approximately similar and can be computed as follows:

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	

However, when these models are plotted on a scatter plot, they all generate different plots that is not interpretable by any regression algorithms.



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

Ans:

Pearson's R also known Pearson correlation coefficient is a measure of linear correlation between two sets of data. It is the ratio between the covariance of two variables and the product of their standard deviations and is essentially a normalized measurement of the covariance. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.

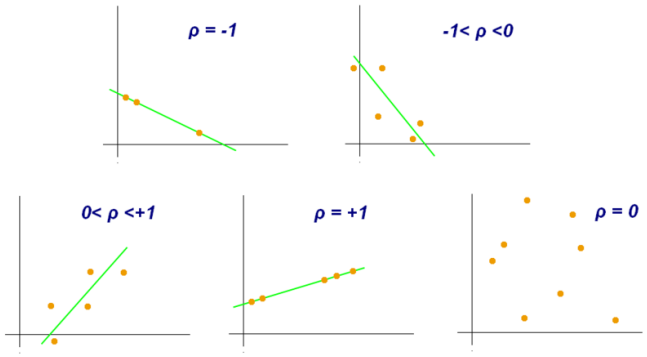
The formula for calculating Person's R is as follows:

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

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

- If there is a positive co-relation between two variables then Pearson's R lies between 0 and 1.
- If two variables have no relationship, the Pearson's R is 0 .
- If there is a negative co-relation then the Pearson's R lies between 0 and -1.

Given below are few plots along with their Pearson's R number.



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

   Ans:

   Scaling is a step which is generally done as part of pre-processing when building a model. It is a method to normalize the range of independent variables so that they are all of comparable scale. It also helps speed up the calculations when building a model.

   Initially, features in dataset have variety of magnitude, scale and units. If scaling is not performed, the model will ignore the scale and unit of the feature leading to coefficients which give too much or too little importance to certain features. This causes the model to be built incorrectly and takes longer to train. Scaling only affects the coefficients of and does not alter the p-values-statistics,r-squared etc.

   Normalized Scaling:

   - It brings all the data in the range of 0 and 1.
   - It should be used when we know that the distribution of data is not Gaussian.
   - It uses minimum and maximum values of a feature for scaling

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Standardized Scaling:

- It scales data such to have 0 mean and 1 as standard deviation
- This should be used when the data has a Gaussian distribution.
- Unlike normalization, standardization does not have a bound on range.
- It uses mean and standard deviation for feature scaling

$$X' = \frac{X - \mu}{\sigma}$$

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

Ans:

Variance Inflation Factor (VIF) is a measure of multi-collinearity for a feature in a dataset. Therefore, if there is a perfect co-relation between features, then R2 become 1 for such a case.

As VIF =1/1-Rsquare,

Therefore, for perfect co-relation VIF =1/1-1 =1/0=Infinity

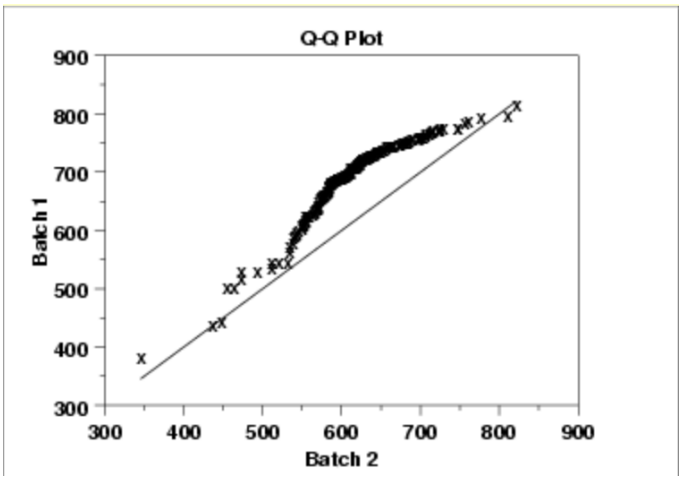
A VIF of infinity indicates that the corresponding variable can be expressed perfectly by a linear combination of other variables in the dataset. We usually drop such features from the dataset as they are redundant.

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

Ans:

In statistics, a Q–Q plot (quantile-quantile plot) is a probability plot, a graphical method for comparing two probability distributions by plotting their quantiles against each other. A quantile means the fraction (or percent) of points below the 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 reference line is plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions.



It helps us to determine if two datasets came from population with a common distribution. It can help determine if the test and train dataset for a model came from same distribution.

The advantages of the q-q plot are as follows:

- The sample sizes do not need to be equal.
- Many distributional aspects can be simultaneously tested.

  Shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.