Linear Regression Assignment

Name: Prashant Mohan Sinha **Date: 13-04-2022**

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?

Ans: In the provided 'day.csv' data set, we have several independent categorical variables such as: ‘season’, ‘yr’, ‘mnth’, ‘holiday’, ‘weekday’, ‘workingday’, ‘weathersit’. The variables like ‘mnth’, ‘weekday’, ‘workingday’ does not affect the dependent variable or may be not needed in presence of other significant variables, for example four seasons represent the whole month time period, thus having the collinearity between them.

It is observed that demand of the bike is positively correlated with clear weather(strongly), mistyweather & with year 2019.

Also it is observed that demand of the bike is negatively affected by the season spring(strongly), winter, holiday & summer season.

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

Ans: It is necessary to reduce the extra column(s) created during dummy variable creation. By encoding the categorical variable as dummy variable, the same level of information can be represented by lesser number of encoded dummy variables. Further, it reduces the correlation among created dummy variables. Before using the **drop_first** straightway, we should understand the logical conclusion of the dropping the dummy variables. In such case, we may drop dummy variable manually.

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

Ans: The target variable (‘TotalUsers’ i.e. ‘cnt’ = ‘registered‘+ ‘casual’) is highly correlated with temperature variable ‘temp’ & ‘atemp’.

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

Ans: The assumption of Linear Regression after the model building is based on the residual analysis on the training data set. Followings are the assumption need to be satisfied:

- Normal Distribution of Residual/Error: Error values (ε) should be normally distributed for any given value of X
- Homoscedasticity: The errors should have constant variance
- Independence of Errors: Error values are statistically independent

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

Ans: If it is not the spring season, variables feature clear Weather, Year it falls, and misty + cloudy weather affects positively the target variable most. If the season falls in spring, it potentially affects the rest of variable in the negative sense. Reason for drop in demand in spring season could be external.

Linear Regression Assignment

Name: Prashant Mohan Sinha **Date: 13-04-2022**

____

General Subjective Questions

1. Explain the linear regression algorithm in detail.

A linear regression model algorithm is a supervised machine learning regression task program which determines the best fit line. It estimates the relationship between a dependent continuous variable, and one or more independent variables. The dependent variable is also called the target/response variable. Independent variables are also called predictor or explanatory variables. The input predictor variables are of two types: a). continuous numerical & b). categorical variables (should be converted into dummy variables). The best fit line is modelled on basis of minimizing the cost function based on least-square method. The method requires to scale the data set via minmax method or standardised based on mean & standard deviation.

In case of multi linear regression, the method requires to select the significant & non-multicollinear features in iterative way to perform the 1. statistical method significance probability test and/or 2. automated recursive feature elimination & 3. Variance Inflation Factor.

Algorithm divides the input data into training & test data set. The residual is calculated from the trained data regression model which assumes the following condition to satisfy:

- Normal Distribution of Residual/Error at any given point of X
- Homoscedasticity: The errors should have constant variance
- Errors Independence: Error values are statistically independent

Prediction of model is finally evaluated on the test data set.

Linear Regression Assignment

Name: Prashant Mohan Sinha **Date: 13-04-2022**

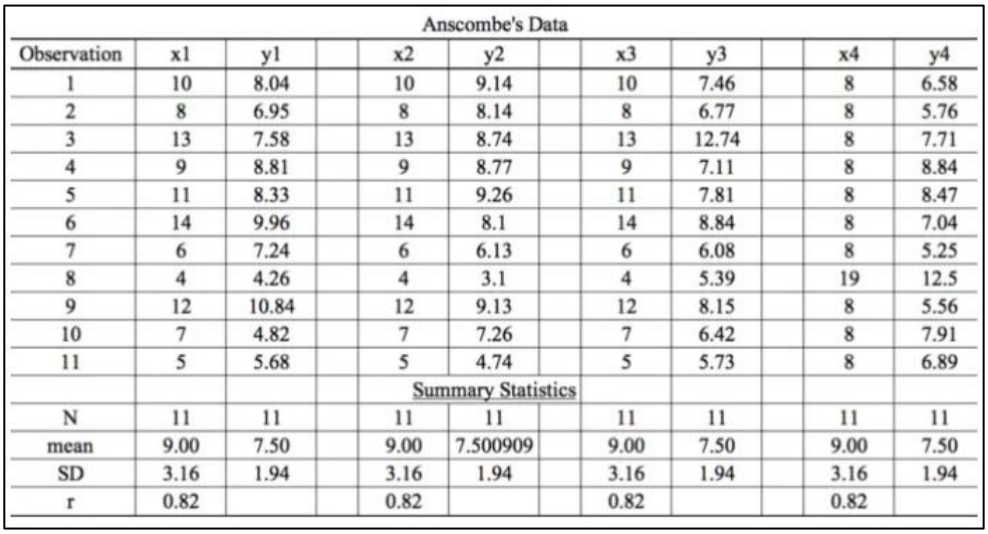
____

2. Explain the Anscombe's quartet in detail.

Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots.

It was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting the graphs before analysing and model building, and the effect of other observations on statistical properties.

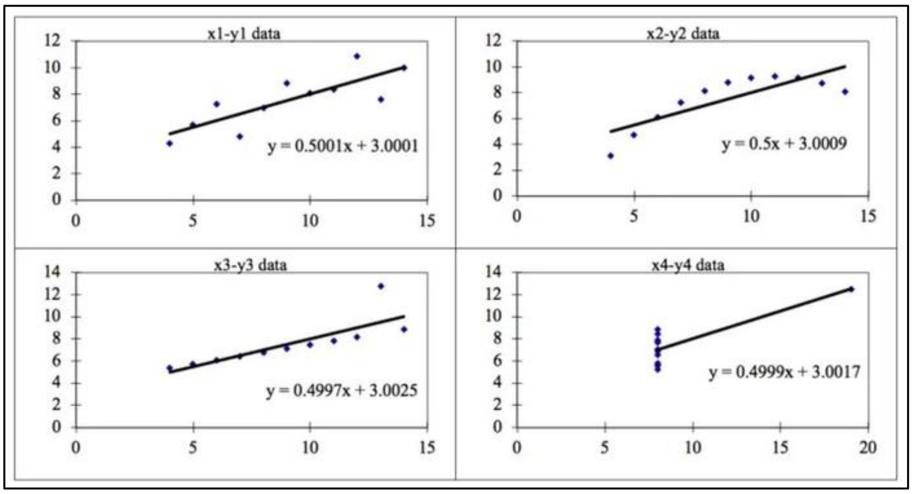
There are these four data set plots which have nearly same statistical observations, which provides same statistical information that involves variance, and mean of all x,y points in all four datasets.



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	

Model plot on scatter plot:



The four datasets can be described as:

- Dataset 1: this fits the linear regression model pretty well.
- Dataset 2: this could not fit linear regression model on the data quite well as the data is non-linear.
- Dataset 3: shows the outliers involved in the dataset which cannot be handled by linear regression model.
- Dataset 4: shows the outliers involved in the dataset which cannot be handled by linear regression model.

Hence, all the important features in the dataset must be visualised before implementing any machine learning algorithm on them which will help to make a good fit model.

3. What is Pearson's R?

In statistics, the Pearson correlation coefficient (PCC), also referred to as Pearson's r, is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.

The Pearson's correlation coefficient varies between -1 and +1 where:

i. r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)
ii. r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)
iii. r = 0 means there is no linear association
iv. 0< |r| < 0.5 means there is a weak association
v. 0.5< |r|< 0.8 means there is a moderate association
vi. |r| >0.8 means there is a strong association

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

Scaling is pre-processing of data which is applied to independent variables to normalize the data within a particular range.

In real-world; the collected data set of different parameters/attributes comes with different range of numerical values. If scaling is not done then it takes the magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

Normalized Scaling: It brings all of the data in the range of 0 and 1. If the data set is having large deviation/outliers, then outliers force to compress the rest of data significantly then nature of distribution lost.

Standardized Scaling: It brings all of the data into a standard normal distribution which has mean= zero and standard deviation of one SD unit. It retains the nature the original data distribution but may not scale all the feature in the same scale range.

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

If VIF is infinite, it means R^2 is exactly one or very much close to one. This means the independent feature is exactly having the collinearity with some other independent feature(s).

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

Q-Q(quantile-quantile) plots is to graphically analyse and compare two probability distributions by plotting their quantiles against each other. If the two distributions which we are comparing are exactly equal then the points on the Q-Q plot will perfectly lie on a straight-line y = x.

In natural events, Normal Distributions occur very frequently, thus capture the large scope of data distribution. Q-Q plot can check the normal distribution of the data set.

In linear regression Q-Q plot is used to validate the normal distribution of the residuals.