Linear+Regression+Subjective+Questions.pdf

KHUSHBOO SINGH

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?**

Ans- There are following infer:-

- **In season boxplot increase in bike rental count in fall and summer season and decrease in spring season.**
- **There are few outlier**
- **In boxplot year we can see demand of bike rental in 2019 increase as compare to 2018.**
- **In weathersit boxplot we can see that clear, partly cloudy rental is higher.**

2. **Why is it important to use drop_first=True during dummy variable Creation?**

Ans- drop_first=True is important to use because it helps to reduce the extra Columns during creation of dummy variable. Hence it reduces the correlations created among dummy variables.

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

Ans- Looking at pair-plot among the numerical variable the variable temp and atemp seems to have highest correlation with the target variable cnt.

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

Ans- There are following assumptions of Linear regression after building the model on training set:-

1. **Linear relationship between X and Y.**

2. **Error terms are normally distributed.**
3. **Error term are independent to each other.**
4. **Error terms has constant variance (homoscedasticity).**

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

Ans-Based on the final model significant features contributing significantly towards explaining the demand of the shared bikes are-

1. **yr with coefficient 0.234 indicating that the bike demand has increased from last year and it is expected to increase in future.**
2. **temp with coefficient 0.451 indicating as temperature increases, the demand of bikes increase.**
3. **Snow with coefficient -0.2864 indicating there is snow ,less demand of bikes**

General Subjective Questions

1. Explain the linear regression algorithm in detail?

Ans- Linear regression algorithm is a machine learning algorithm. It is based on supervised learning. Linear regression is a part of regression analysis.

Linear regression- Linear regression is a one of the very basic forms of machine learning where we train a model to predict the behavior of data based on some variables.

In the case of linear regression means the two variable which is on x-axis and y –axis should be linearly correlated.

Example- The police department running a campaign to reduce the number of robberies, in this case graph will be linearly downward.

Linear regression equation-

$$Y=a+bx$$

Where a and b equation is

b(slope)= $n\sum xy-(\sum x)(\sum y)/n\sum x^2-\sum x^2$

a(intercept)= n $n\sum y-b(\sum x)$ **/n**

Here x and y is two variables on the regression line

b=slope of the line

a=y-intercept of the line

x=independent variable from data set.

y=dependent variable from the dataset

Advantages of Linear regression

- **Linear regression provides a powerful statistical method to find the relationship between variables.**

- **Linear regression produces the best predictive accuracy for linear relation.**
- **Risk management**
- **Price prediction**

Steps in linear regression algorithm-

- **Reading, understanding and visualizing the data**
- **Processing the**
  - *** Encode the categorical variables using dummy variables.**
  - ***Divide into Train and Test data set.**
  - ***Scaling on training data set**
- **Building the model**
  - *** Create X and Y variables**
  - *** Use RFE and build initial model with RFE recommended features.**
  - ***Feature selection and iteratively work on better model by dropping features based on p-value and VIF**
- **Model Evaluation**
  - *** Prove that the assumptions of linear regression are true**
  - ***Residual analysis**
  - *** Scaling on test data set**
  - ***Predict target variable using the fitted model**
  - ***Evaluate the prediction on the test data set**
- **Conclusion**

2. **Explain the Anscombe's quartet in detail?**

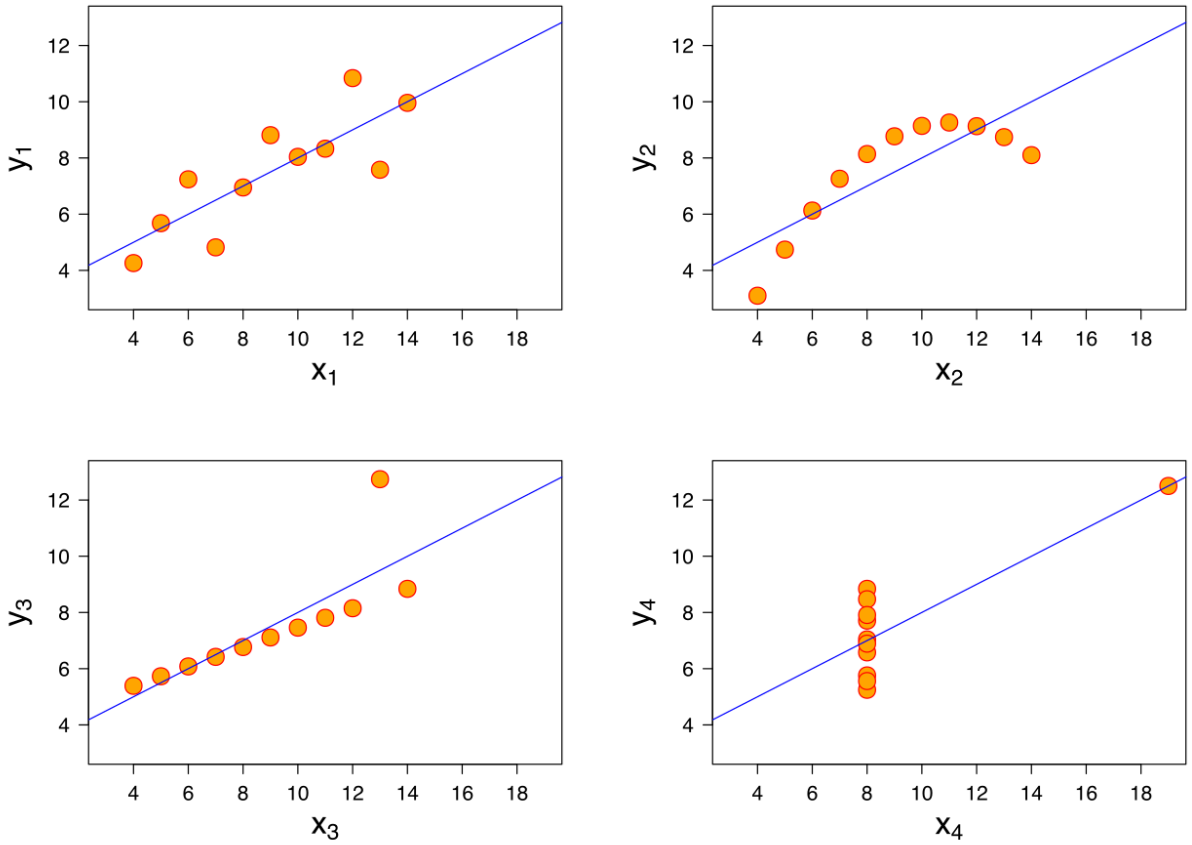
   Ans- Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, they have different distributed and appear very different when graphed. Each data set consists of eleven(x,y) points.

   It is demonstrated both the important of graphing data before analysis it and effect of outliers.

   .

- **The first scatter plot appears to be simple linear relationship.**
- **The second graph is not distributed normally,while the relationship between two variables is obvious.**
- **The third graph distribution is linear but should have a diiferent regression line.**
- **The forth graph is the example of high leverage point.**

- **All four sets are identical when examined using simple summary statistics but vary when graphed**



3.What is Pearson's R?

Ans- Pearson's R also called Pearson's correlation ,it is commonly used in linear regression ,Pearson's method is the most common method to use for numerical variables ,it assign value between -1 and 1,where 0 means no correlation ,1 means positive correlation and -1 means negative correlation.it measures linear correlation between two variable x and y.

$$r = \sum(x_i - \overline{x})\,(y_i - \overline{y}) / \sqrt{\sum(x_i - \overline{x})^2 \sum(y_i - \overline{y})^2}$$

where r= correlation coefficient

x_i**=values of the x-variable in the sample**

$\overline{x}$ **=mean of the values of the x-variable**

y_i=values of the y-variable in a sample

$\overline{y}$= mean of the values of the y-variable

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

Ans- Scaling is a step of data pre-processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculation in an algorithm.

It is important because collected data set contains features highly varying in magnitudes, units and range, if scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling, to solve this issue we have to do scaling to bring variable to the same level of magnitudes.

It do effect t-statistic, F-statistic ,p-value and R- squared

Normalized Scaling-

- **It is also called min-max scaling**
- **It brings all data in the range of 0 and 1**
- **MinMax Scaling : x=x-min(x)/max(x)-min(x)**

Sklearn.preprocessing.MinMaxScaler helps to implement in python.

Standardization Scaling-

- **Standardization replaces the values by their Z-score.**
- **It bring all of the data into a standard normal distribution which has mean(μ) zero and standard deviation(σ) is one.**
- **Standardization: x= x-mean/sd**
- **Sklearn.preprocessing.scale helps to implement standardization in python.**

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

Ans- Variance inflation factor(VIF) is a measure of the amount of multicollinearity in a set of multiple regression variables. This ratio is calculated for each independent variable. High VIF indicates that associated independent variable is highly collinear with the other variables in the models.

The selection of all the variables and proceeds by repeatedly deselecting variables showing high VIF. An infinite VIF indicates that the corresponding variable may expressed exactly by a linear combination of other variables, which is show an infinite VIF.

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

Ans-A Q-Q (quantile- quantile) plot is a probability plot,which is graphical method for comparing two probability distributions by plotting their quantile against each other.

USE and importance-

A Q-Q plot is used to compare the shapes of distribution and providing a graphical view of how properties are different or similar in the two distribution.

Q-Q plot can be used to compare collections of data.

We can use Q-Q plots to compare two samples of data .

Q-Q plot commonly used to compare a data set to a theoretical distribution to each other.

Q-Q plot is a graphical tool to asses if a set of data came from some theoretical distribution such as normal or exponential.

In Q-Q plot sample sizes do not need to be equal.