Bike Sharing System

Linear Regression Assignment Subjective Questions

- Sreedhar Gunda

Assignment-based Subjective Questions

- From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?
    - **Weather**
        - The rentals are more when the weather is clear or slightly cloudy. When there is heavy storm there are no rentals at all
    - **Weekdays**
        - There is no significant difference between the weekdays. Working Day vs Non-Working Day, there is no significant difference between these two.
    - **Months**
        - Bike rental from months June to September are very high. Bike rentals during January is the least among all the months.
    - **Season**
        - Highest number bike rentals happened in Fall and least in spring.

Assignment-based Subjective Questions

- Why is it important to use **drop_first=True** during dummy variable creation?
  - While using get_dummies for a categorical column having 3 values we will get 3 new columns. These 3 columns will be of 0 and 1. For eg – Categorical value – ***Summer, Winter ,Spring***.

  After creating dummy columns for summer, it will be 100, for winter it is 010 and Spring 001. Now if we observe the same thing can be observed with 2 columns also, like Summer it will be 00, winter 10 and Spring 01. We can express the other columns also in terms of remaining columns.

  Issue- Because of the above mentioned one, it leads to a high chance of multi collinearity between columns and effects our model prediction. So, we need to drop one of these columns and we drop the first one.

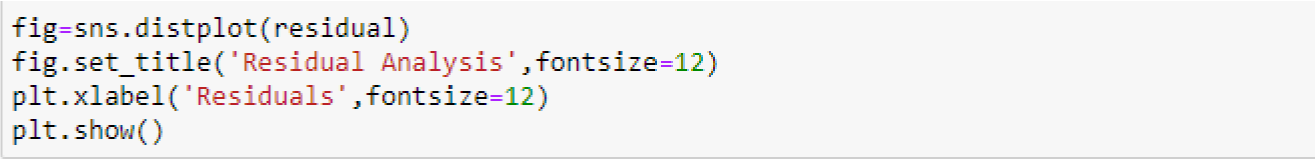
Assignment-based Subjective Questions

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

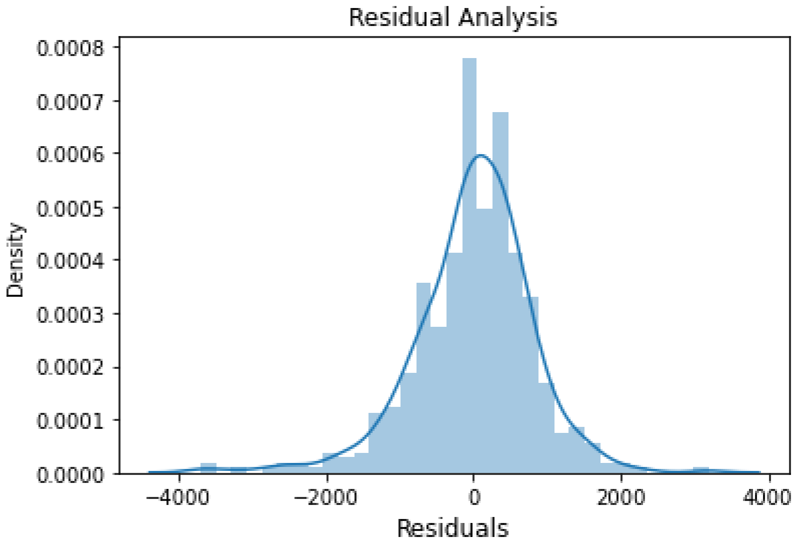
    - **temp** and **atemp** has the highest correlation with cnt(target variable).

Assignment-based Subjective Questions

- How did you validate the assumptions of Linear Regression after building the model on the training set?
  - Plotted a distribution plot with the residuals obtained (actual y_train values – predicted y_train values) and observed that the residuals(error terms) followed a normal distribution with mean nearly equal to zero.



```
fig=sns.distplot(residual)
fig.set_title('Residual Analysis',fontsize=12)
plt.xlabel('Residuals',fontsize=12)
plt.show()
```



- Things to be considered while moving to a MLR :

  - OverFitting – Train and test accuracy are nearly equal (83 and 80 respectively). This makes the model is not overfitted and correct.

  - Multicollinearity – Using the VIF table we observed that collinearity for each column wrt other columns is less than 5 which is ok.

  - Feature Selection – p values obtained for all the columns is 0 which means all are highly significant

Assignment-based Subjective Questions

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

  - Temperature(Positively Correlated, Coeff: 3905)
  - Year(Positively Correlated, Coeff: 2038)
  - Light Rain (Negatively Correlated, Coeff: -2484)

General Subjective Questions

1. Explain the linear regression algorithm in detail

Linear Regression is based on supervised learning. It performs a regression task.

Regression models a target prediction value based on independent variables. In linear regression we have 1 dependent variable and 1 independent variable.

It is mostly used for finding out the relationship between variables and forecasting.

Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, LR technique finds out a linear relationship between x and y.

The formula is given as y=mx+c where ,

y is the dependent variable

x is the dependent variable

m is the coefficient of x

c is intercept

The best fit regression line can be found by minimizing the cost function (RSS in this case, using the Ordinary Least Squares method). This can be achieved by two methods-

1.Differentiation

2.Gradient Descent method

The strength of our model can be explained by R^2(r-squared) (1- (RSS/TSS))

RSS-Residual Sum of Square

TSS-Total Sum of Squares

BEST FIT line-

The best-fit line is found by minimising the expression of RSS (Residual Sum of Squares) which is equal to the sum of squares of the residual for each data point in the plot.

Residuals for any data point is found by subtracting predicted value of dependent variable from actual value of dependent variable: ri=(yi-y_predi)

Sum of all residuals RSS= r12 +r22+….+r2n

TSS(Total sum of squares): Itis the sum of errors of the data points from mean of response variable., TSS =Σ(yi-ymean)2

R2 determines the accuracy of the model. Higher the R2 value the better the model has fitted our data.

R2 = 1-(RSS/TSS)

Assumptions of simple linear regression:

1. Linear relationship between X and Y.

2. Residuals/Error terms are normally distributed with mean nearly equal to zero

3. Error terms are independent of each other

4. Error terms have constant variance

Assessing the model – R2 value should be high and p value of x should be less than 5%(0.05), which indicated that x is significant in predicting the y value.

Building a model

1.We can build the model using statsmodel.api.

2.Split the data into test and train data(70-30 ratio respectively)

3. Add constant to x. Required while doing in statsmodel.

4.Call OLS on x and y variables and call fit() on the model.

5.Call summary() to see the details and check for R2 values and p value.

6. Predict the y values using predict() on the model.

7. Calculate the residuals (y_train-y_pred).

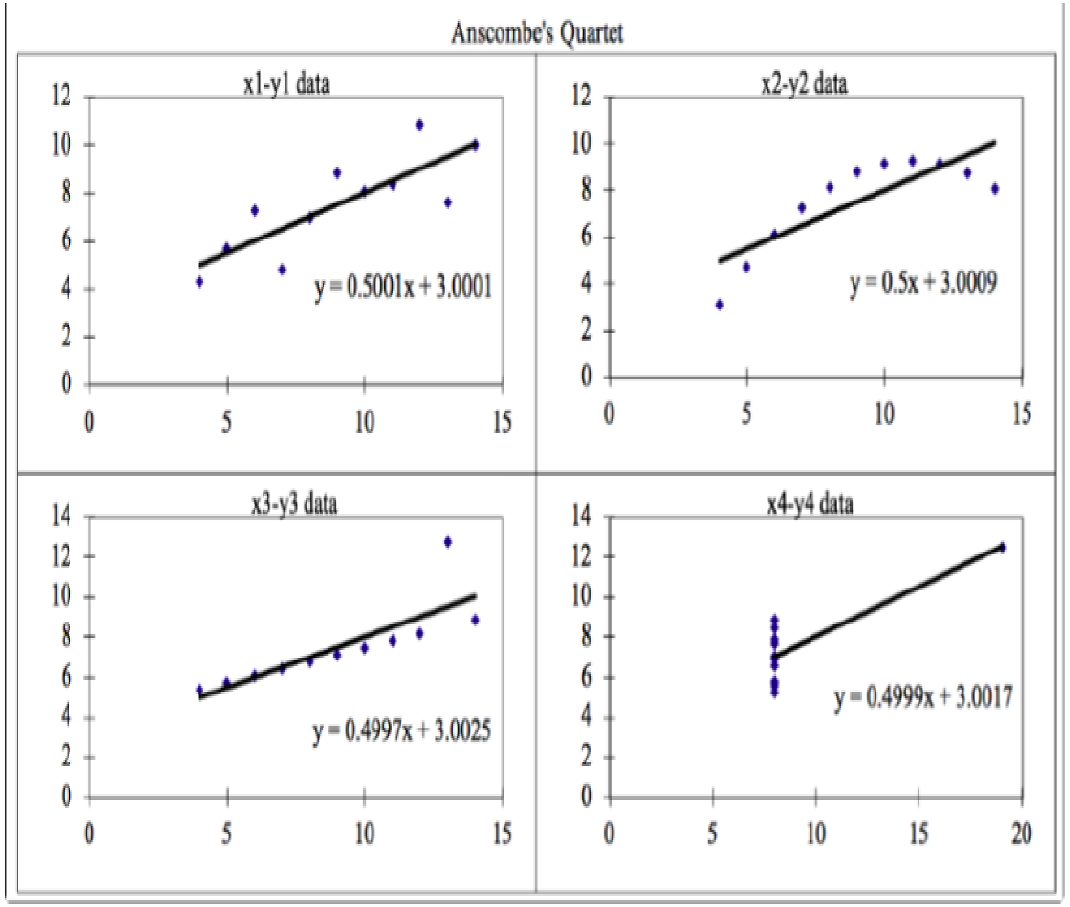
8.Plot the distribution plot with respect to residuals and check if it is normally distributed or not.

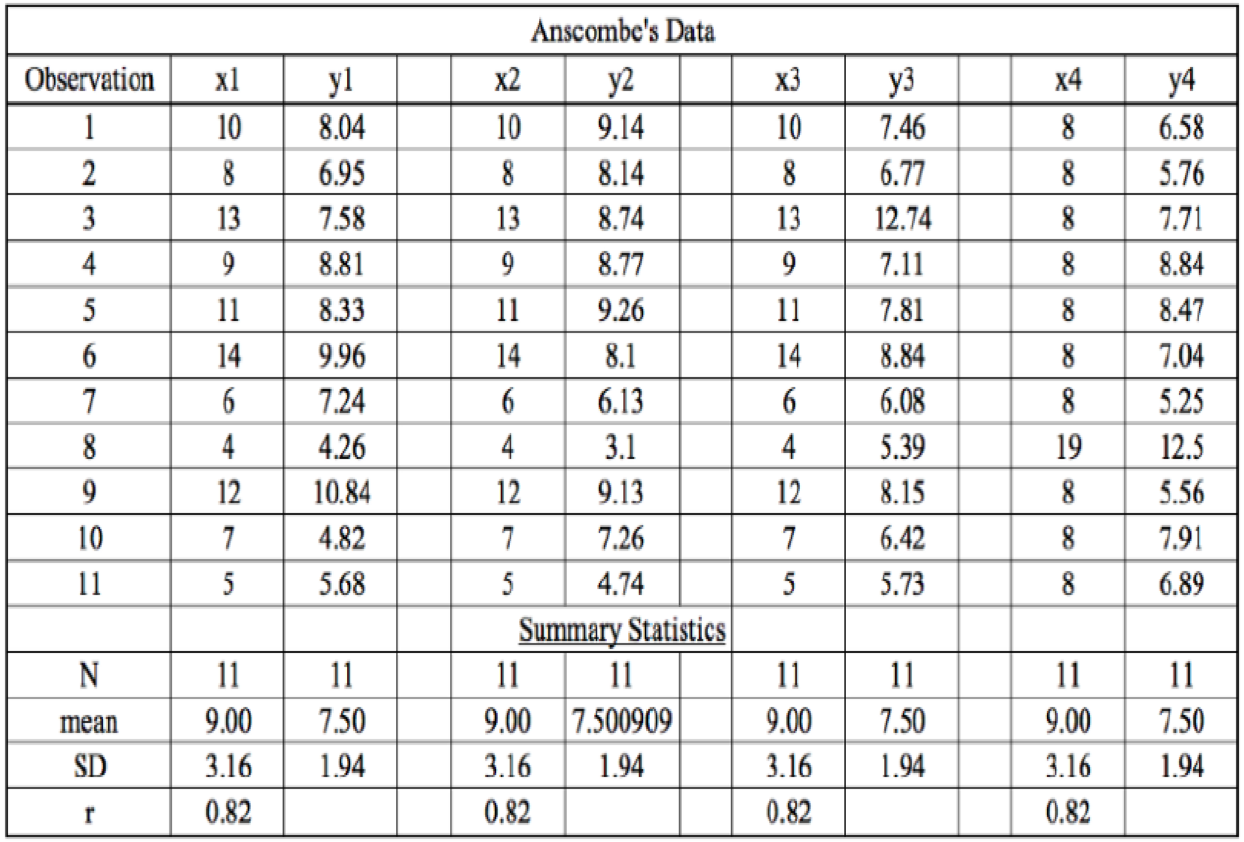
9.Predict the y values on the test data and calculate R2 on that. If both the R2 obtained (from model and from test data) are nearly equal, we can say the model is working fine.

General Subjective Questions

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

Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties yet appear very different when graphed. Each dataset consists of eleven (x,y) points.





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	

Data Points with mean and SD values

They have very different distributions and appear differently when plotted on scatter plots.

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.

This tells us about the importance of visualising the data before applying various algorithms to build models.

These suggest that the data features must be plotted in order to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers.

Linear Regression can only be considered a fit for the data with linear relationships**.**

Data Points with mean and SD values

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

Conclusion: These 4 datasets explain the importance of data visualisation and before applying any algorithm to build a model.

General Subjective Questions

- 3.**What is Pearson’s R**?

Pearson correlation coefficient is a measure of the strength of a linear association between two variables — denoted by r.

Assumptions-

1. For the Pearson r correlation, both variables should be normally distributed.

2. There should be no significant outliers. Pearson’s correlation coefficient, r, is very sensitive to outliers, which can have a very large effect on the line of best fit and the Pearson correlation coefficient. This means including outliers in our analysis can lead to misleading results.

3. Each variable should be continuous.

4. Two variables should have a linear relationship. Plotting a scatter plot will help to know if the relation between them is linear or not.

5. The observations are paired observations ie for every independent observation there must be a corresponding dependent observation.

6. Homoscedasticity is the residual terms have constant variance. The variance should not follow any pattern as the residual values change.

Properties-

1.Value lies between +1 and -1. +1 positively correlated, -1 negatively correlated, 0 no correlation.

2.Independent on the unit of measurements.

3.Correlation coefficient between the two variables is symmetric

General Subjective Questions

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

Scaling:

It is a step-in data Pre-Processing which is applied to independent variables to normalize the data within a particular range. To have all the columns in similar range.

Reasons for scaling :

Most of the times, collected data set contains features of different ranges, units and magnitude. If scaling is not performed, then algorithm only takes magnitude and not units hence incorrect modelling will happen. To solve this issue, we have to do scaling to bring all the variables to the same range. Scaling just affects the coefficients and parameters like p-values, R-squared, etc won’t be affected. It also helps in speeding up the calculations in an algorithm.

Normalised Scaling

Also called as minmax scaling. It brings all the data in between 0 and 1. It is part of sklearn.preprocessing.

MinMaxScaling is calculated by (X-Xmin)/(Xmax-Xmin)

Here Xmax and Xmin are maximum and minimum values of the feature respectively.

Standardization Scaling:

Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean zero and standard deviation one. All the values in this won’t be between 0 and 1.

Formula- X’ = (X-(mean))/(sd)

General Subjective Questions

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

  If there is perfect correlation between the columns, then we get VIF = infinity.

  This shows a perfect correlation between two independent variables. Infinite VIF value indicates that the corresponding variable can be expressed exactly by a linear combination of other variables. In the case of perfect correlation, we get R2 =1, which lead 1/(1-R2) to infinity. To solve this problem, we need to drop one of the columns which is causing this perfect multicollinearity, we can start dropping by seeing the p values(which is the least significant) if multiple columns have VIF as infinite

General Subjective Questions

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

Quantile-Quantile plots (Q-Q Plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile.

For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it.

Purpose-

The purpose of Q-Q plots is to find out if two sets of data have come from the same distribution or not.

A 45-degree line is plotted on the graph and

1.if the 2 data sets come from a common distribution the points will fall on that line.

2. if all the point lie away from the line they come from a different distribution.

Q-Q Plot in Linear Regression:

We fit a linear regression model, check if the points lie approximately on the line, and if they don't, our residuals are not normally distributed. This implies that for small sample sizes, we can't assume your estimator x is significant in predicting the dependent variable