Linear Regression Assignment - Shared Bikes Demand

By – Ishkhan Marzook

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?
   - Comparing the years 2018 and 2019, **most bike rentals happened in the year 2019**
   - Most bikes are rented during **Fall Season** and the least during **Spring Season**
   - Most bikes are rented during **Clear Weather** followed by **Mist + Cloudy**, **Light Snow**
   - Most bike rentals happen during the month of **September**
   - Most bikes are rented during **non-holiday** days
   - Most bikes are rented on **Thursday** compared with other weekdays.

2. Why is it important to use drop_first=True during dummy variable creation?
   - It helps by **reducing the creation of an extra variable** which contributes toward **reducing the correlations created among the dummy variables**

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

4. How did you validate the assumptions of Linear Regression after building the model on the training set?
   - Validated by performing **Residual Analysis** and during the Residual Analysis if error terms have a normal distribution, it’s a strong indication model fulfilling the Linear Regression assumption.

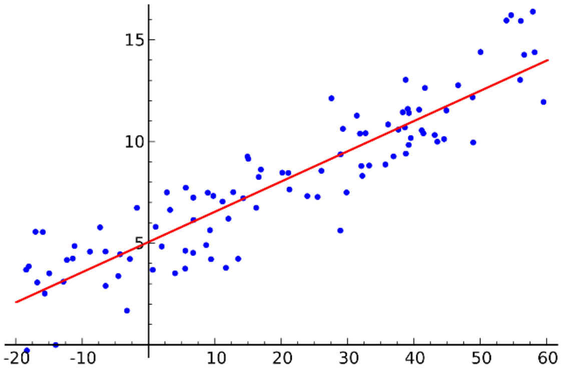
5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
   - **temp**
   - **weathersit (light snow)**
   - **2019 (Year)**

General Subjective Questions

1. Explain the linear regression algorithm in detail.

- A linear regression algorithm is a Machine Learning Algorithm used in predictive analysis to predict the dependent variables based on the given independent variable, linear regression algorithm is used to make predictions on continuous variables.

- Essentially linear regression algorithm displays the linear relationship between a dependent variable and an independent variable, which is why it is referred to as linear regression.



(Figure - 1)

- *Figure 1* shows the linear regression model, in the figure has data points distributed over the X and Y axis, by analyzing the data point distribution we can conclude that its sort of a linear distribution, which makes this dataset a good candidate for linear regression algorithm

- We apply the linear regression algorithm and find the best-fitted line for the distribution of the data points, in *Figure - 1* we can see the plotting of the best-fitted line. With the best-fitted line, we can predict the Y value for the Given X value.

- Below you can see the equation for a simple linear regression algorithm,

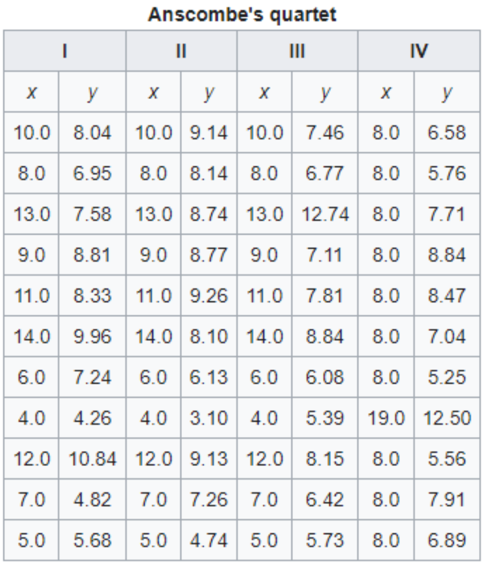
$$Y_i = B_0 + B_1X_i$$

Y_i = Dependent Variable
B_0 = Intercept
B_1 = Slope\Coefficient
X_i = Independent Variable

- Linear regression can be categorized into 2 major types,
  - Simple Linear Regression – Only one independent variable is used to predict the value of a dependent variable.
  - Multiple Linear Regression – More than one independent variable is used to predict the value of a dependent variable.

2. Explain the Anscombe's quartet in detail.

- Anscombe's Quartet contains four data sets that have nearly the same simple summary statistics (such as mean, variance, correlation, and best-fitted line), although the summary statistics look similar, yet have very different distributions and appear very different when plotted. This shows the importance of plotting the data to understand it better rather than relying solely on the summary statistics.

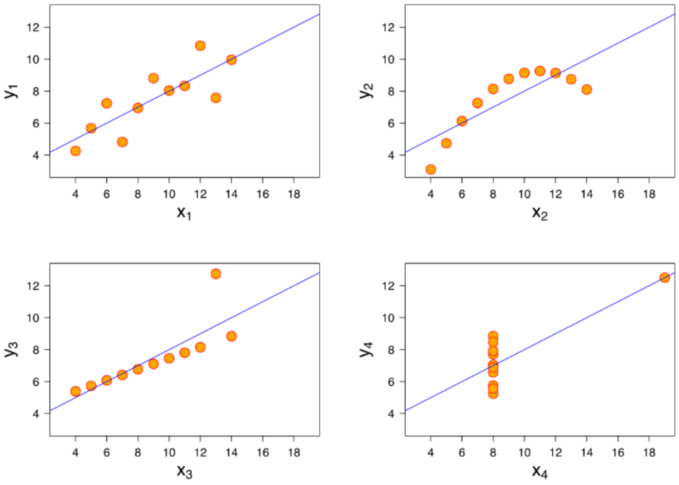


Anscombe's quartet

I		II		III		IV	
x	*y*	*x*	*y*	*x*	*y*	*x*	*y*
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

(Figure - 2)

- *Figure - 2* shows the Anscombe's Quartet contains four data sets that contain almost the same summary statistics, *Figure – 3* shows how the dataset looks after plotting it, we can clearly observe that the distribution of the data points not having any similarity**, this shows the importance of plotting the data to understand it better rather than relying solely on the summary statistics**.



(Figure - 3)

3. What is Pearson’s R?

- In statistics, Pearson’s R? else Pearson correlation coefficient is a tool to measure the strength of the linear relationship between 2 variables.

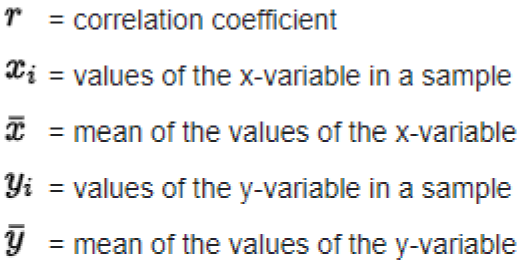
- It’s important to understand that Pearson’s R only measures the linear relationship, and it will be in the range of -1 and +1.

- Based on the following values of Pearson’s R we can conclude below fact regarding the relationships between two variables,
    - **R = 1** – There will be a positive slope between the two variables and when x increases by 1 unit y also increase by 1 unit.
    - **R = -1** – There will be a negative slope between the two variables and when x decreases by 1 unit y also decrease by 1 unit.
    - **R = 0** – This means there is no linear relationship between the two variables and data points are distributed randomly.

- *Figure 4* shows the formula for calculate the Pearson’s R,

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

(Figure - 4)



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

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

    - Scaling is a normalization process applied during the data processing stage to normalize the data within a particular range, which makes the respective algorithm perform faster and increases the interpretability of the data.

    - When analyzing a dataset there will be data values in different rages (for example certain columns only contain binary data, but some numerical columns data can spread across billions), if the scaling is not performed, the respective algorithm creates the model by considering the magnitude of the data leaving the unit of it makes the model inaccurate.

    - Normalized scaling is also called as Min-Max scaling and it's fitting all the data between scales of 0 to 1.

    - Standardization scaling is also called Z-score normalization which the data will be rescaled to ensure the mean and the standard deviation to be 0 and 1.

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

    - This is a scenario happening due to a perfect correlation exists between 2 independent variables, in this case since **R2 = 1**, which makes **1/(1-R2)** calculation to produce a result of infinity. We can solve this issue by removing variables causing multicollinearity from the dataset.

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

    - In statistics, a Q-Q plot also called the quantile-quantile plot is used to analyze two quantiles against each other, this is normally used for identifying the 2 sets of data coming from the same dataset.

References & Images

- https://en.wikipedia.org/wiki/Linear_regression
- https://en.wikipedia.org/wiki/Anscombe%27s_quartet
- https://en.wikipedia.org/wiki/Pearson_correlation_coefficient
- https://www.wallstreetmojo.com/pearson-correlation-coefficient/