LINEAR REGRESSION ASSIGNMENT

Assignment Based Subjective Question

Shashank Pawaskar
shashanksp@msn.com

Module 2 : Linear Regression Assignment

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?**

Response:

- **Category Variable Season :**
  - Number of rentals are lowest in the Spring Season.
  - Number of rentals are highest in Fall Season, followed by Summer.
  - This Indicates that Season can be good predictor for the dependent variable.
- **Category Variable Month :**
  - The trend noticed for the Season, aligns with Month data, where Sep & Oct (Fall Season) has higher rental, followed by May to July (Summer Season)/
  - The median of Months April to September (Summer and Fall) have median value above 4000.
  - This indicates that Month can be a good predictor for the dependent variable.
- **Category Variable weekday :**
  - The Median across all weekdays is approximately aligned.
  - We will use this variable and check again when we build the model.
- **Category Variable weathersit :**
  - Rentals are highest on a Clear day, followed by Misty day.
  - This indicates that weathersit has relation with the dependent variable and can be good predictor.
- **Category Variable workingday :**
  - Higher number of rentals on working day, indicates it can be good predictor for the dependent variable.
- **Category Variable holiday :**
  - Very High number of rentals when it is a not a holiday, indicates some bias nature.
  - We will use this variable and check again when we build the model.variable.

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

Response:

- We use drop_first=True during dummy variable creation to avoid Dummy Variable Trap.
- Dummy Variable Trap occurs when we are when two or more dummy variables and one variable can be predicted from others, which means it will be difficult for regression model to interpret predicted coefficient variables.
- In Bike Rental Assignment, we have Season which has 4 status
  - Fall, Spring, Summer, Winter
- When we create dummy variables without 'drop_first=True' then the dummy variables will be created for each of the status values, as follows:

Season_Fall	Season_Spring	Season_Summer	Season_Winter
0	1	0	0
0	0	1	0
1	0	0	0
0	0	0	1

When you have a categorical variable with say 'n' levels, the idea of dummy variable creation is to build 'n-1' variables, indicating the levels

- Season : Season_Fall can be explained as 1 – (Season_Spring + Season_Summer + Season_Winter)
    - If the values of Season_Spring, Season_Summer, Season_Winter is 0, that implies that it is Fall Season.
- You can clearly see that there is no need of defining **Four** different levels. If you drop a level, say 'Fall', you would still be able to explain the 4 levels.
- Let's drop the dummy variable 'Single' from the columns and see what the table looks like:

Season_Spring	Season_Summer	Season_Winter
1	0	0
0	1	0
0	0	0
0	0	1

- Which shows that If both the dummy variables namely ‘Spring', ‘Summer’ and 'Winter' are equal to zero, that means that it is ‘Fall’ Season.
- This reduces the multi-Collinearity between these dummy variables created.
- In a broader sense, we can conclude that if there are n dummy variables, n-1 dummy variables will be able to predict the value of the nth dummy variable, so one dummy variable should be dropped to avoid multicollinearity.

3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**
   Response:
    - Following are the NUMERICAL ASSOCIATIONS with Bike Rentals
        - atemp : 0.63
        - temp : 0.63
        - windspeed :-0.24
        - humidity : 0.10
    - atemp and temp both have same correlation with target variable of 0.63 which isthe highest among all numerical variables.

4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**
   Response:
    - The simple way to determine if this assumption is met or not is by creating a scatter plot x vs y. If the data points fall on a straight line in the graph, there is a linear relationship between the dependent and the independent variables, and the assumption holds.

5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?**
   Response:
   - **T**op two Positive Impact Features / predictor variables
     - aTemp (Temperature) : With Coefficient of 0.445, indicates that for unit increase in Temperature the Bike Rentals are increasing by 0.445 units.
     - Year : with Coefficient of 0.239, indicates that for unit increase in Year the Bike Rentals are increasing by 0.239 units.
       - Which is likely to be the projected growth of Bike Rental Year-on-Year.
   - Top two Negative Impact Features / predictor variables
     - Weather Light_Snowrain : with Coefficient of -0.229, indicates that for a unit increase in Light_SnowRain, the Bike rentals are decreasing by -0.229 units.
       - Which can be interpreted as whenever the forecast indicates Rain or Snow, the Bike Rentals decrease.
     - Spring Season : With Coefficient of -0.154, indicates that for unit increase in Spring Season the Bike Rentals are decreasing by -0.154 units.
       - Which can be interpreted as, the Bike Rentals are are lower during the Spring Season.

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**

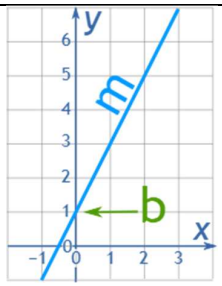
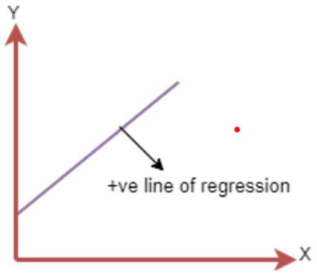
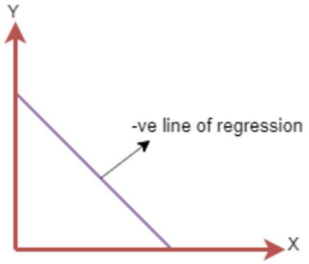
Response:
Machine Learning can identify patterns that normally cannot be seen or cannot find in huge amounts of data. There are different Machine Learning algorithms which are well suited for many different types of situations, such as Supervised and Unsupervised Learning, as well as Semi-Supervised and Reinforcement learning, which are somewhere between the former two. Machine learning models can be classified into the following three types based on the task performed and the nature of the output:

1. **Supervised Learning**
   1. **Regression**: The output variable to be predicted is a **continuous variable**, e.g. scores of a student based on their last year data.
      1. Output variable to be predicted is a continuous / numeric variable.
      2. Linear Regression, Ridge Regression, Neural Network Regression, Lasso Regression, Decision Tree Regression are some of the ML Algorithms under Regression type.
   2. **Classification**: The output variable to be predicted is a **categorical variable**, e.g. classifying incoming emails as spam or ham
      1. Output variable to be predicted is a categorical variable.
      2. Logistic Regression, Naïve Bayes, K-Nearest Neighbors, Decision Tree, Support Vector Machines are some of the Classification type of ML Alogrithms.
2. **Un-Supervised Learning**
   1. **Clustering**: **No pre-defined notion of a label** is allocated to groups/clusters formed, e.g. customer segmentation
      1. E.g Customer Market Segmentation - You segment customers in categories (say A, B, C) and then provide say 50% discount to A, 60% to B and 70% to C.

Module 2 : Linear Regression Assignment

Linear regression is Regression machine learning where we train a model to predict the behaviour of data based on some variables. Linear regression is a **linear model**, e.g. a model that assumes a linear relationship between the input variables (x) and the single output variable (y). More specifically, that y can be calculated from a linear combination of the input variables (x).

Linear Regression is represented by mathematical equation y = mx + b

<table>
<tr><td> 
</td><td>$y = mx + b$
Slope or Gradient
Value of y when x=0</td></tr>
<tr><td colspan="2">For ML, the equation is denoted as $y = \beta_0 + \beta_1 x$</td></tr>
<tr><td>Positive Linear Relationship
 
</td><td>Negative Linear Relationship
 
</td></tr>
</table>

Linear regression can be further divided into two types of the algorithm:

- **Simple Linear Regression:** If a single independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Simple Linear Regression.
- **Multiple Linear regression:** If more than one independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Multiple Linear Regression. The equation is denoted as $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 ... + \beta_n x_n$, where X_1, X_2 , X_3,…. Represent independent variables.

The data set is divided into two parts during Linear Regression (supervised learning) method:

- **Training data** is used for the model to learn during modelling
- **Testing data** is used by the trained model for prediction and model evaluation

In Linear regression, there is a notion of a best-fit line — the line which fits the given scatter-plot in the best way. The below plot shows the difference (depicted by Red Line) between the actual and the linear line is called Residual.

Module 2 : Linear Regression Assignment

A linear regression algorithm helps in predicting the value of a dependent variable, finding a best fit line helps in minimizing the difference between the Predicted Value and Actual Values. Some of methods used to arrive at best fit line are as follows:

- Ordinary Least Squares (OLS) - is common technique for estimating coefficients of linear regression equations which describe the relationship between one or more independent / explanatory variables and a dependent variable
    - Square(E_1) + Square(E_2) +……..+ Square(E_n) -→ Residual Sum of Squares (RSS)
        - Where E_1 = y – y predicted value
- Gradient Descent is an optimisation algorithm which optimises the objective function (for linear regression it's cost function) to reach to the optimal solution.
    - When there are one or more inputs you can use a process of optimizing the values of the coefficients by iteratively minimizing the error of the model on your training data.

Linear Regression Model Performance – Finding the best model out of various models (which is also called Optimization)

- The R-squared value indicates how much of the variation in the dependent variable can be explained by the explanatory variable
    - The R-squared values range between 0 and 1.
    - A value of 0.8 means that the independent / explanatory variable can explain 80 percent of the variation in the observed values of the dependent variable.
- The p-value explains how reliable that explanation is.
    - p-value higher than 0.05 means that the independent / explanatory variable does not affect the dependent variable.

The aspects to consider when moving from simple to multiple linear regression are:

- Overfitting : As you keep adding the variables, the model may become far too complex. It may end up memorising the training data and will fail to generalise. A model is generally said to overfit when the training accuracy is high while the test accuracy is very low
- Multicollinearity - Associations between predictor variables, which you will study later
- Feature selection - Selecting the optimal set from a pool of given features, many of which might be redundant becomes an important task
- Normal distribution of error terms

The model finalized with training data set, is then applied to Test Data set. The R-Squared is verified to measure the accuracy of the model. If the R-Squared value of Test Data is in close proximity of the R-Squared value of model with Training Data Set, indicates a good model.

Module 2 : Linear Regression Assignment

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

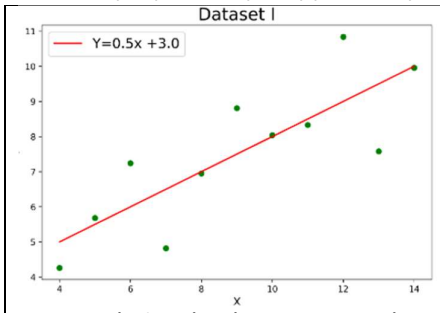
Response:

Anscombe's quartet comprises a set of four dataset, having identical descriptive statistical properties in terms of means, variance, R-Squared, correlations, and linear regression lines but having different representations when we scatter plot on graph. The datasets were created by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data and to show that summary statistics alone can be misleading.

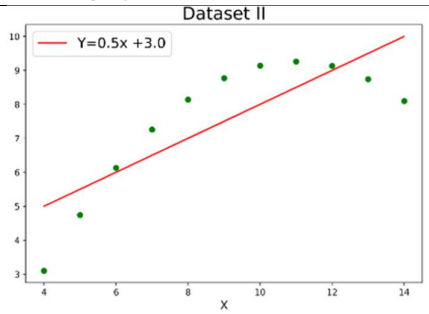
- The four datasets that make up Anscombe's quartet each include 11 x-y pairs of data. When plotted, each dataset seems to have a unique connection between x and y, with unique variability patterns and distinctive correlation strengths. Despite these variations, each dataset has the same summary statistics, such as the same x and y mean and variance, x and y correlation coefficient, and linear regression line.
- **Anscombe's quartet** is used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on summary statistics. It also emphasizes the importance of using data visualization to spot trends, outliers, and other crucial details that might not be obvious from summary statistics alone.
- The four datasets of **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

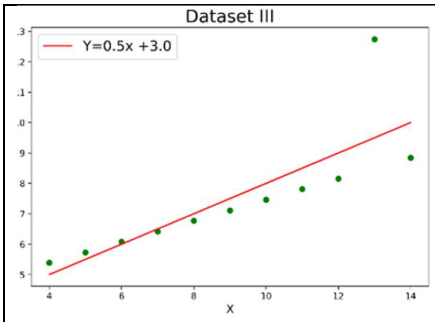
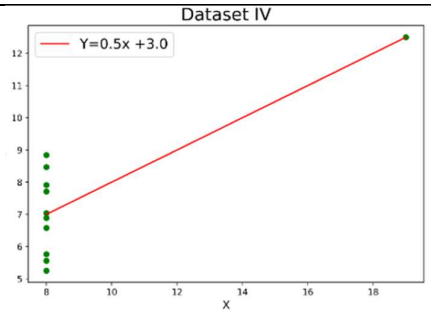
- **Anscombe's quarter** graphs : comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed.



Linear relationship between x and y



Non-linear relationship between x and y

 Linear Relationship with one Outlier(far from the line) at top right.	 One high-leverage point is enough to produce a high correlation coefficient.

- **Anscombe's quarter** is still often used to illustrate the importance of looking at a set of data graphically before starting to analyze according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets.

Source : https://www.geeksforgeeks.org/anscombes-quartet/

3. **What is Pearson's R?**

Response:

The Pearson correlation coefficient (r) is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables. When one variable changes, the other variable changes in the same direction.

- If *r* is positive, then as one variable increases, the other tends to increase.
- If *r* is negative, then as one variable increases, the other tends to decrease.
- A perfect linear relationship (*r*=-1 or *r*=1) means that one of the variables can be perfectly explained by a linear function of the other.

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

Response:

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 calculations in an algorithm.

- Scaling of variables is an important step because, as the different variables can be on a different scale with respect to all other numerical variables, which take very small values.
- Also, the categorical variables that take either 0 or 1 as their values.
- Hence, it is important to have everything on the same scale for the model to be easily interpretable.

It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

The two popular rescaling methods- Normalized (also called as Min-Max) scaling and Standardisation (mean=0 and sigma=1).

- **Normalized / Min-Max Scaling:** is the simplest method and consists in rescaling the range of features to scale the range in [0, 1]
  - ***sklearn.preprocessing.MinMaxScaler*** *helps to implement normalization in python.*
    - (x - Xmin) / (Xmax - Xmin)

- **Standardization Scaling:** replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (**μ)** zero and standard deviation one (**σ**).
    - Formula – (x – mean(X)) / SD(X)
    - This method is widely used for normalization in many machine learning algorithms
    - sklearn.preprocessing.scale helps to implement standardization in python.

The advantage of Standardisation over the other is that it doesn't compress the data between a particular range as in Normalization / Min-Max scaling,

- i.e., in Normalization / Min-Max Scaling we lose some data, especially the outliers.

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

   Response:

   VIF = infinity indicates that there is a perfect correlation between two independent variables.
   - In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity.
   - To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

   An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables, which will also have VIF = infinite.

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

   Response:

   Q-Q plots are also known as Quantile-Quantile plots. As the name suggests, they plot the quantiles of a sample distribution against quantiles of a theoretical distribution.
   - Quantile-Quantile plot or Q-Q plot is a scatter plot created by plotting 2 different quantiles against each other.
   - The power of Q-Q plots lies in their ability to summarize any distribution visually.
   - QQ plots is very useful to determine
       - If two populations are of the same distribution
       - If residuals follow a normal distribution. Having a normal error term is an assumption in regression and we can verify if it's met using this.
       - Skewness of distribution
   - In the Bike Sharing Assignment, I have used to Q-Q Plot to verify the Normal Distribution of the Residuals, plot image below: