Subjective Questions

Assignment Based Questions:

1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?**
    - yr, workingday, Sep(mnth=9) and Mon (weekday=6) have a positive correlation on the dependent variable that is count of total shared bikes.
    - windspeed, Light Snow (weathersit=3), Mist (weathersit=2), Spring (season=1), Dec (mnth=12), Jan (mnth=1) and Nov (mnth=11) have a negative correlation with the dependent variable.
2. **Why is it important to use drop_first=True during dummy variable creation?**
    - When we create dummy variables, one variable for each category under the column is created and the value it holds is binary. Only one of the categories is 1 and rest all are 0. When n-1 categories are 0 then automatically n^{th} category becomes 1. Hence, the first category can be deleted assuming when all others are 0, deleted category is 1.
3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**
    - Looking at the pair plot, temp and atemp seems to have the highest correlation with the target variable. Just by looking at the plot it is difficult to say out of these two which has the highest correlation with the target variable. But on calculating the correlation coefficient is same for both temp and atemp which is 0.65.
4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**
    - The assumptions were validated as follows:
        i. 1^{st} assumption: Linear relationship between X and y – For this I created pairplot where we could see a linear relationship between ‘cnt’ (y) and ‘temp’ and ‘atemp’ (X)
        ii. 2^{nd} assumption: Error terms are normally distributed – For this I performed Residual Analysis by creating a distplot on the error terms/residual for both training and test dataset. For both, the distribution was normal and centered around 0
        iii. 3^{rd} assumption: Error terms are independent of each other – For this I plotted a scatter plot between the dependent variable (cnt) and the error terms. There was no visible pattern between the two.
        iv. 4^{th} assumption: Error terms have constant variance – For this I plotted a scatter plot between the y_train data and the error terms. Visibly it looks that the error terms have constant variance. The variance is not showing any increasing or decreasing trend.
5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?**
    - Based on the final model created, following are the top 3 features contributing significantly towards explaining the demand of the shared bikes:
        i. Weathersit=3(Light Snow) (coeff = -0.3145)
        ii. Yr (coeff = +0.2465)

iii. Season =1(Spring) (coeff = -0.2209) – This is the dummy variable created from Season

General Questions

1. **Explain the linear regression algorithm in detail.**
    - Step 1: Read the data from the file
    - Step 2: Understand the data using the data dictionary
    - Step 3: Summarize and Visualize the data
        i. Using describe, info function
        ii. Performing EDA
        iii. Checking the correlation
        iv. Relation of dependent variable with other variables (is there a linear relationship with any variable?)
    - Step 4: Adding the dummy variables for the categorical variables and dropping the first one (removing redundant variables)
    - Step 5: Concatenating the dummy variables to the main dataframe
    - Step 6: Droping the categorical variable for which dummy variables were created (removing redundant variables)
    - Step 7: Splitting the data into train and test with 70:30 ratio (or 75:25 or 80:20)
    - Step 8: Scaling the features using Min-Max scaler or Standard scaler
    - Step 9: Separating dependent variable (y_train) from the dataframe and labelling the remaining as the X_train
    - Step 10: There are multiple ways to build a model.
        i. Method 1: Using OLS method, start with one variable (highest correlation) and keep adding more variables. Assess the variables and model using p-value, VIF, Adj R-sq and F-stat.
        ii. Method 2: Using OLS method, build the model with all the variables and keep removing the variables with high p-value (>0.05) or high VIF(>5) or both. Assess the model using Adj R-sq and F-stat.
        iii. Method 3: Use RFE by providing the number of variables you need to build the model. Use OLS to build the model and then assess the model and the variables using p-value, VIF, Adj R-sq and F-stat. Remove variables manually with high p-value (>0.05) or high VIF (>5) or both.
        iv. Finalize the model when all the variables are significant and the vif is <5 for all the variables.
    - Step 11: Perform residual analysis. Plot the error (difference between predicted and actual values for train dataset) distribution plot to check with the error terms are normally distributed.
    - Step 12: Predict using the model on the test data set
    - Step 13: Evaluate the model by plotting a scatter plot between the actual values of y from test set and the predicted value of the test dataset
    - Step 14: Plot the error terms (difference between predicted and actual values for test dataset) distribution plot to check with the error terms are normally distributed

- Step 15: Calculate the r2 score of the actual y test values and the predicted y test values

2. **Explain the Anscombe's quartet in detail.**
   - Anscombe's quartet is a group of dataset where the descriptive statistics (mean, standard deviation and linear regression line) are almost the same but when plotted (x,y) on the graph, they look very different. Francis Anscombe conducted the experiment with 4 data set of 11 data point each. This experiment was done to explain the importance of looking at the plotted data as well and not just the descriptive statistics. Anscombe quartet also demonstrates the impact of the outliers on the data.

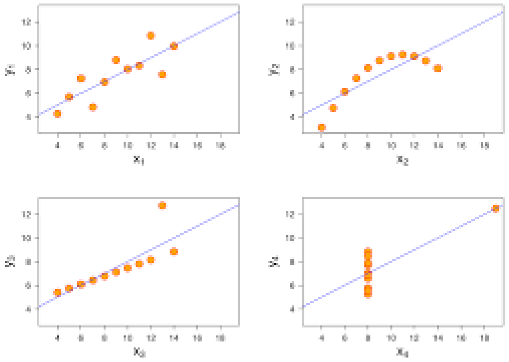

Fig 1: Anscombe's quartet (copied from Wikipedia)

3. **What is Pearson's R?**
   - Pearson's R is the correlation coefficient which measures the correlation between 2 variables. The value of the correlation coefficient varies from -1 to 1. Following is the interpretation of the values:
     i. 1: Very strong +ve correlation
     ii. -1: Very strong -ve correlation
     iii. 0: No correlation
     iv. 0<r<1: Represents +ve correlation. Higher the value, higher the correlation
     v. -1<r<0: Represents -ve correlation. Higher the absolute value, higher the correlation
   - +ve correlation means if one variable increases, the other also increases
   - -ve correlation means if one variable increases, the other decreases

4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?**
   - Scaling is a very important step in preparing the data. Scaling helps in transforming and bringing the data from all the variables to a common scale.
   - The various independent variables have values at different scale. Some might be binary, some range from 1-10 and some might be in thousands. So, in order to avoid the bias of these variables and give equal importance to all the variables, scaling is performed. Scaling brings the variables in the same range. There are 2 types of scaling: Normal and Standard
   - The major difference between the two scaling technique is
     i. Normalization: The data is distributed between 0 and 1 by using the formula (x-xmin)/(xmax-xmin).

ii. Standardization: The data is created with mean = 0 and standard deviation = 1. Formula used is (x-mean)/std deviation

5. **You might have observed that sometimes the value of VIF is infinite. Why does this happen?**
   - VIF = 1/(1-Rsq) where Rsq is the correlation between the variables. If Rsq is 1, then VIF becomes infinite, which means that the variables are in perfect correlation. So, whenever there are variables which are in perfect correlation, VIF for these variables will be infinite.
6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.**
   - Q-Q plots are used to find the type of distribution of the dataset. In linear regression we make an assumption that the error terms are normally distributed, which can be visualized using the Q-Q plot. If there is a fairly straight line, in that case we can be sure that the error terms are normally distributed.

Submitted by

Trapti Agarwal