**Projects on Advanced Statistical Methods**

Linear Regression Project -1

Business Project – Car Sales

**Dummy Variables:-** If we include a separate dummy variable for each category, we will introduce multicollinearity to the regression.

**Weight interpretation:**

1. **Continuous Variables:**
   1. A **positive weight** shows that as a feature increases in value, so do the log\_price and Price respectively.
   2. A **negative weight** shows that as a feature increase in value, log\_price and Price decrease
2. **Dummy Variables:-** 
   1. A **positive weight** shows that the respective category (Brand) is more expensive than the Benchmark (Audi)
   2. A **negative weight** shows that the respective category (Brand) is less expensive than the Benchmark (Audi)

**Plt.scatter(x,y,[,alpha])** :- creates a scatter plot

**Alpha**: specify the opacity

**Steps:**

1. Importing the relevant Libraries
2. Loading the Raw Data
3. Exploring the descriptive Statistics
4. Determining the variables of Interest
5. Dealing with Missing Values
6. Exploring the PDF (Probability Distribution Function)
7. Dealing with Outliers (exclude on the basis of quantile more than 0.99) for every variables or some basic understanding about the other parameters like Engine Volume
8. Relaxing the assumptions and convert the price into log for linearity
9. Again draw the plot against all the variables to check the linearity
10. Check the VIF for multicollinearity and remove if have high multicollinearity
11. Now create dummy variables for Models (It’s work only on Categorical Variable)
12. Now Start Linear Regression model while declaring inputs and targets
13. Import the scaling module
14. Divide the module into Train Test split
15. Create the Regression
16. Fit the Module
17. Predict and plot for check the accuracy
18. Find the Bias and weights
19. Now reverse back the log function using exponential function and check the final predicted value and Targets
20. You can also import in the excel file

**Libraries Used**

|  |
| --- |
| Import numpy as np  Import pandas as pd  import matplotlib.pyplot as plt  import seaborn as sns  sns.set()  import statsmodels as sm |
| From sklearn.preprocessing import StandardScaler  From sklearn.linear\_model import LinearRegression  From sklearn.stats.outliers\_influence import variance\_inflation\_factor  From sklearn.model\_selection import train\_test\_split |

**Important Syntax:**

|  |
| --- |
| sns.distplot(data\_2['Total']) |
| f, (ax1,ax2,ax3,ax4,ax5) = plt.subplots(1,5, sharey=True, figsize=(15,3))  ax1.scatter(data\_2['Month'],data\_2['Open Price'])  ax1.set\_title('Price and Month')  ax2.scatter(data\_2['Issue Price'],data\_2['Open Price'])  ax2.set\_title('Price and Issue Price')  ax3.scatter(data\_2['QIB'],data\_2['Open Price'])  ax3.set\_title('Price and QIB')  ax4.scatter(data\_2['Total'],data\_2['Open Price'])  ax4.set\_title('Price and Total')  ax5.scatter(data\_2['Issue Size'],data\_2['Open Price'])  ax5.set\_title('Price and Issue Size')  plt.show() |
| from statsmodels.stats.outliers\_influence import variance\_inflation\_factor  variables = predictors  vif = pd.DataFrame()  vif["VIF"] = [variance\_inflation\_factor(variables.values, i) for i in range(variables.shape[1])]  vif["features"] = variables.columns  vif |
| reg\_summary = pd.DataFrame(x.columns.values, columns=['Features']) |

**Also attached the PDF of IPO Analysis v2.**



Logistic Regression Project -1



Cluster Analysis – Projects

