**Implementation and Performance Analysis of Linear Regression**

***Data Preparation:*** Start by collecting and preparing your dataset. Ensure that you have a suitable dataset with a continuous dependent variable and one or more independent variables.

***Data Split:*** Split the dataset into training and testing subsets. Typically, the training set is used to train the model, and the testing set is used to evaluate its performance.

***Model Training***: Fit the linear regression model to the training data. Most programming languages and libraries provide built-in functions or classes for linear regression. Train the model by providing the independent variables and the corresponding dependent variable.

***Model Evaluation***: Evaluate the trained model's performance using various evaluation metrics. Common metrics for linear regression include R-squared, mean squared error (MSE), and mean absolute error (MAE). Calculate these metrics by comparing the predicted values to the actual values from the testing dataset.

***Interpretation of Coefficients***: Analyse the coefficients of the linear regression model to understand the relationship between the independent variables and the dependent variable. Positive coefficients indicate a positive correlation, while negative coefficients indicate a negative correlation. The magnitude of the coefficients represents the strength of the relationship.

***Residual Analysis***: Examine the residuals (the differences between the predicted and actual values) to assess the model's goodness of fit. Plotting the residuals against the predicted values can help identify patterns, such as heteroscedasticity or outliers.

***Performance Visualization***: Visualize the performance of the linear regression model using appropriate graphs and plots. For example, you can create scatter plots to visualize the relationship between the independent and dependent variables, or plot the predicted values against the actual values.

***Further Analysis***: If necessary, you can explore additional aspects such as feature selection, regularization techniques (e.g., ridge regression, lasso regression), or cross-validation to enhance the model's performance or address specific requirements.

***Performance Comparison***: To assess the linear regression model's performance, you can compare it with other regression techniques or variations of linear regression (e.g., polynomial regression). This comparison can help determine the effectiveness and suitability of the linear regression approach for your specific dataset.

Remember to interpret the results with caution and consider the assumptions of linear regression, such as linearity, independence of errors, constant variance, and normal distribution of residuals. Additionally, always cross-validate your findings and consider potential limitations or sources of bias in the data.

Overall, implementing and analysing the performance of linear regression involves data preparation, model training, evaluation, interpretation, and visualization. It is crucial to follow a systematic approach and use appropriate evaluation metrics and visualization techniques to draw meaningful conclusions from the analysis.

Look at the data set below; it includes some car-related information.

**Table 1 data.csv**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Car | Model | Volume | Weight | CO2 |
| Toyota | Aygo | 1000 | 790 | 99 |
| Mitsubishi | Space Star | 1200 | 1160 | 95 |
| Skoda | Citigo | 1000 | 929 | 95 |
| Fiat | 500 | 900 | 865 | 90 |
| Mini | Cooper | 1500 | 1140 | 105 |
| VW | Up! | 1000 | 929 | 105 |
| Skoda | Fabia | 1400 | 1109 | 90 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Mercedes | A-Class | 1500 | 1365 | 92 |
| Ford | Fiesta | 1500 | 1112 | 98 |
| Audi | A1 | 1600 | 1150 | 99 |
| Hyundai | I20 | 1100 | 980 | 99 |
| Suzuki | Swift | 1300 | 990 | 101 |
| Ford | Fiesta | 1000 | 1112 | 99 |
| Honda | Civic | 1600 | 1252 | 94 |
| Hundai | I30 | 1600 | 1326 | 97 |
| Opel | Astra | 1600 | 1330 | 97 |
| BMW | 1 | 1600 | 1365 | 99 |
| Mazda | 3 | 2200 | 1280 | 104 |
| Skoda | Rapid | 1600 | 1119 | 104 |
| Ford | Focus | 2000 | 1328 | 105 |
| Ford | Mondeo | 1600 | 1584 | 94 |
| Opel | Insignia | 2000 | 1428 | 99 |
| Mercedes | C-Class | 2100 | 1365 | 99 |
| Skoda | Octavia | 1600 | 1415 | 99 |
| Volvo | S60 | 2000 | 1415 | 99 |
| Mercedes | CLA | 1500 | 1465 | 102 |
| Audi | A4 | 2000 | 1490 | 104 |
| Audi | A6 | 2000 | 1725 | 114 |
| Volvo | V70 | 1600 | 1523 | 109 |
| BMW | 5 | 2000 | 1705 | 114 |
| Mercedes | E-Class | 2100 | 1605 | 115 |
| Volvo | XC70 | 2000 | 1746 | 117 |
| Ford | B-Max | 1600 | 1235 | 104 |
| BMW | 2 | 1600 | 1390 | 108 |

The size of the engine can be used to estimate a car's CO2 emissions, but multiple regression allows us to include additional variables, such as the car's weight, to improve the prediction's accuracy.

* 1. **How does it function?**

We have modules in Python that will carry out the work for us. Import the Pandas module first.

***import pandas***

We can read CSV files using the Pandas module and get a DataFrame object in response. The name of file is data.csv. Data is defined above in the table.

***Input: df = pandas.read\_csv("data.csv").***

Create a list of the independent values, and then designate this list as variable X. Add the dependent values to the y variable.

Show Python code for reading dataset

**Summarizing the Dataset**

Read the basic Information about the dataset

Dimensions of Dataset

Listing all top 10 data,

Listing all bottom 10 data,

View the Statistical Summary

**Multiple Regression**

linear regression, multiple regression attempts to predict a value based on two or more factors, but with more than one independent value.