**LINEAR REGRESSION FROM SCRATCH**

Linear regression is a fundamental statistical method used to model the relationship between a dependent variable and one or more independent variables. In this report, I have implemented linear regression from scratch using the Auto-MPG dataset, a popular dataset that provides various features of automobiles and their corresponding miles per gallon (MPG) values. Our objective is to predict the MPG values based on the available features.

### Dataset Description

The Auto-MPG dataset contains 398 instances of automobiles with the following attributes:

1. Cylinders
2. MPG (Miles per Gallon)
3. Displacement
4. Horsepower
5. Weight
6. Acceleration
7. Model Year
8. Origin

The goal is to predict the MPG (dependent variable) using the other features (independent variables).

### Data Preprocessing

Before applying linear regression, the data has to be preprocessed. This includes handling missing values, normalizing the data, and splitting the dataset into training and testing sets.

1. Handling Missing Values: The dataset contains some missing values in the 'Horsepower' attribute. I have dropped these missing values.
2. Normalization: To ensure that all features contribute equally to the model, I have normalized the features to have zero mean and unit standard deviation.
3. Splitting the Dataset: I have split the dataset into a training set (80%) and a testing set (20%) to evaluate the model's performance.

### Implementation of Linear Regression from Scratch

Linear regression can be represented by the following equation:

where:

* ŷ is the predicted value.
* *X* is the matrix of input features.
* *w* is the vector of weights.
* ​*b* is the bias term.

The goal is to find the optimal weights and bias that minimize the mean squared error (MSE) between the predicted and actual MPG values. The MSE is defined as:

#### Gradient Descent Algorithm

We will use gradient descent to minimize the MSE. The updates for weights and bias are given by:

Where

* is the cost function.
* 𝛼 is the learning rate.
* The partial derivatives are:

### Results and Discussion

After training the linear regression model from scratch, I have obtained the following mean squared errors (MSE):

* Training MSE: This value indicates how well the model fits the training data.
* Testing MSE: This value indicates how well the model generalizes to new, unseen data.

Lower MSE values indicate better model performance. In practice, we aim to minimize both training and testing MSE to ensure that our model generalizes well.

*(The MSE values can be viewed in the .ipynb file attached with this report)*

### Conclusion

In this report, I have implemented linear regression from scratch using the Auto-MPG dataset. I have preprocessed the data, normalized the features, and used gradient descent to minimize the mean squared error. The model's performance was evaluated using training and testing MSE. This exercise demonstrates the fundamental concepts of linear regression and the importance of data preprocessing and evaluation metrics.

### Future Work

Future improvements could include:

1. Feature Engineering: Creating new features or selecting the most relevant features.
2. Regularization: Applying techniques like Lasso or Ridge regression to prevent overfitting.
3. Advanced Models: Exploring more complex models like polynomial regression or machine learning algorithms to improve predictions.