Linear Regression

Kohei Tateyama, S6427214, Università di Genova

***Abstract*—**

***Index Terms*—Linear regression, Machine Learning, AI, regression problem**

I. Introduction

LINEAR regression is a foundational statistical technique used to model the relationship between variables and make predictions based on those relationships. In this report, we delve into the principles of 1-dimensional (simple) and multi-dimensional (multiple) linear regression. Through practical applications, we aim to make a better understanding of regression methods. Using two distinct datasets to explore the effectiveness of these models, visualize the results, and employ Mean Squared Error (MSE) for evaluation.

II. Linear Regression

A. 1-Dimensional Linear Regression

In 1-dimensional linear regression, we have a single independent variable X and a single dependent variable Y. 1-dimensional linear regression can be represented as:

$$Y = mx + b$$

where m is the slope of the line, and b is the intercept. The objective of linear regression is to determine the values of m and b that minimize the difference between the predicted values and the actual data points.

B. Multi-Dimensional Linear Regression

In multiple linear regression, we have multiple independent variables $(X_1, X_2, \ldots, X_n)$ and a single dependent variable Y. Multi-dimensional linear regression can be represented as:

$$Y = b + m_1X_1 + m_2X_2 + \ldots + m_nX_n$$

Where b is the intercept, and $m_1, m_2, \ldots, m_n$ are the slopes for each independent variable. The goal is to find the values of b, $m_1, m_2, \ldots, m_n$ that minimize the difference between the predicted values and the actual data points.

III. Lab Work

Task1: Data Preprocessing

In this assignment, we used 2 different datasets. The first dataset shown in table 1 is the dataset of turkish stocks. And the second dataset shown in table 2 is the dataset of cars. Although the datasets are more extensive, for the purpose of this report, only the initial portions of each dataset are shown.

Task 2: Fit a Linear Regression Model

-0.004679315	0.012698039
0.007786738	0.011340652
-0.030469134	-0.017072795
0.003391364	-0.005560959
-0.021533208	-0.010988634
-0.022822626	-0.012451259
0.001756552	-0.012220196

Fig. 1: Turkish stock dataset

Model	mpg	disp	hp	weight
Mazda_RX4	21	160	110	2.62
Mazda_RX4Wag	21	160	110	2.875
Datsun_710	22.8	108	93	2.32
Hornet_4_Drive	21.4	258	110	3.215
Hornet_Sportabout	18.7	360	175	3.44

Fig. 2: Car dataset

1) 1 dimensional regression without intercept: In this task, we used the entire turkish dataset 1 to apply a 1-dimensional linear regression with the intercept of 0. The result is shown in figure 3

2) Comparison with subsets: In this task, we used 10% of the turkish dataset 1 as subsets to apply the linear regression. The result is shown in figure 4. As the data is sparse, with every iteration the slope changed with different subsets.

3) 1-dimensional regression with intercept: In this task, we used the columns 'mpg' and 'weight' in the car dataset 2 to do a 1-dimensional regression with intercept. The result is shown in figure 5.

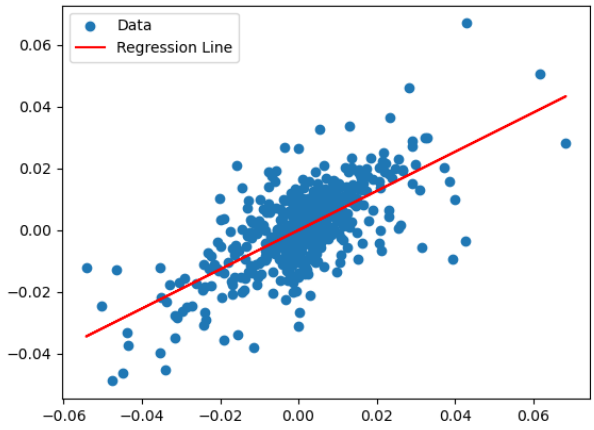

Fig. 3: One dimensional linear regression with no intercept

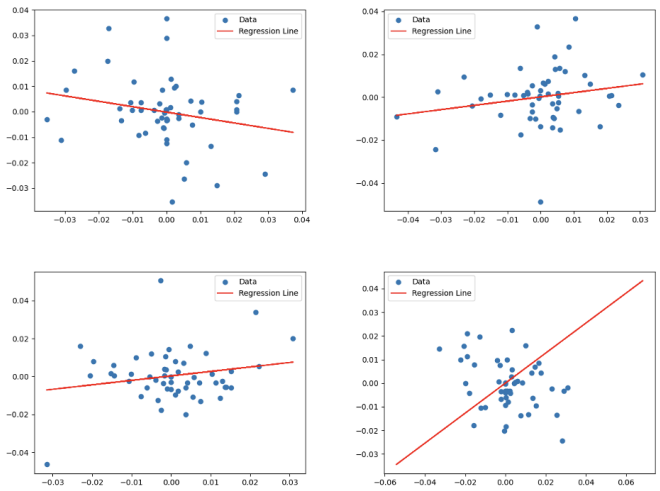

Fig. 4: Linear regression with the subset of 1

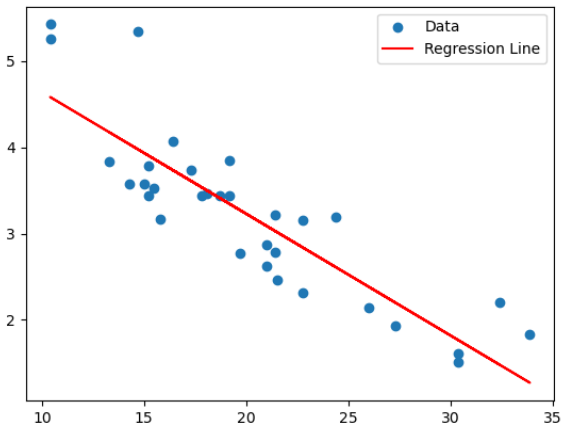

Fig. 5: Data

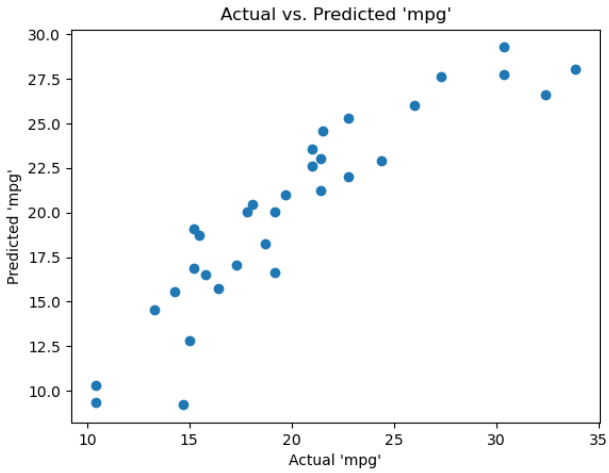

Fig. 6: Data

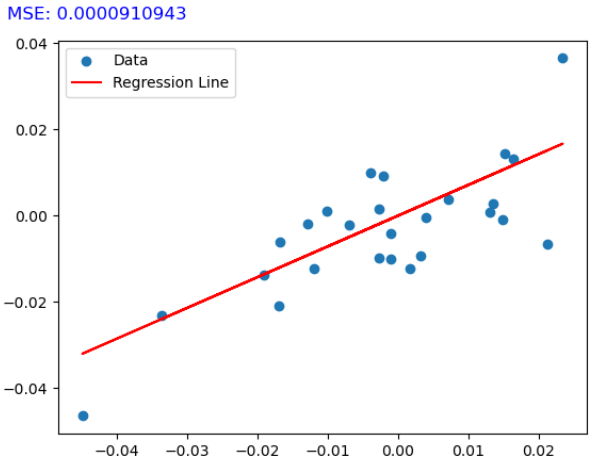

Fig. 7: Data

4) Multi-dimensional regression: In this task, I used all the columns in the car dataset to implement a multi-dimensional regression. The result is shown in figure 6. The horizontal axis shows the actual 'mpg' and the vertical axis shows the predicted 'mpg'. All the plots are around $y = x$ indicating that the prediction is close to the actual 'mpg'.

Task 3: Test regression model For this task, we re-ran the previous tasks with 5% of the training data 10 times with each different subsets. For each subsets, I calculated the mean squared error (MSE).

A. 1

The result for the first task is shown in the table I. And the figure 7 shows one subset of this task. Because the data had very small numbers, the MSE became a very small number.

Iteration	Training Data MSE	Remaining Data MSE
1	3.2734×10^{-4}	1.6056×10^{-4}
2	2.7298×10^{-4}	1.6333×10^{-4}
3	1.5222×10^{-4}	1.6950×10^{-4}
4	1.1276×10^{-4}	1.7152×10^{-4}
5	3.4780×10^{-4}	1.5951×10^{-4}
6	1.3476×10^{-4}	1.7039×10^{-4}
7	1.7772×10^{-4}	1.6819×10^{-4}
8	2.3757×10^{-4}	1.6514×10^{-4}
9	1.2399×10^{-4}	1.7094×10^{-4}
10	1.1390×10^{-4}	1.7146×10^{-4}

TABLE I: MSE Values for Different Iterations 4-1

B. 3

The result for the third task is shown in the table II.

Iteration	Training Data MSE	Remaining Data MSE
1	14.2478	10.5929
2	8.3588	11.9519
3	9.7604	11.6284
4	12.8432	10.9170
5	12.6583	10.9597
6	9.2100	11.7554
7	16.3636	10.1046
8	14.2575	10.5906
9	11.5615	11.2128
10	11.8015	11.1574

TABLE II: MSE Values for Different Iterations 4-3

C. 4

The result for the forth task is shown in the table III.

Iteration	Training Data MSE	Remaining Data MSE
1	5.98	34.62
2	6.39	11.40
3	0.77	16.98
4	0.36	17.89
5	5.76	7.20
6	0.19	9.54
7	1.77	56.41
8	0.57	22.85
9	6.34	13.83
10	5.71	17.05

TABLE III: MSE Values for Different Iterations 4-4

IV. How to run the code

The source code(LinearRegression.py) and the datasets (mtcarsdata-4features.csv) and (turkish-se-SP500vsMSCI.csv) is submitted in 1 folder. To run the code LinearRegression.py, first change the directory to the directory of this folder in terminal. Now, if by running the code, all the tasks should be running.

V. Conclusion

In conclusion, this report has provided the results of the practical use case of both one-dimensional and multi-dimensional linear regression. I've illustrated the process of fitting linear regression models, assessing their performance by employing metrics like the Mean Squared Error (MSE), and extracting valuable insights from the obtained results. As our experiments and discussions have shown, linear regression emerges as a valuable instrument suitable for endeavors such as forecasting stock prices or gaining insights into the elements that impact car fuel efficiency.