Linear Regression

Gian Marco Balia
Robotic Engineering - University of Genoa
s4398275@studenti.unige.it

***Abstract*—Linear regression is a statistical technique used to model the relationship between a dependent variable and one or more independent variables by fitting a linear equation to observed data. Widely used across fields like economics, biology, and engineering, linear regression is valued for its simplicity and interpretability in predictive modeling. While it assumes a linear relationship and is sensitive to outliers, enhancements like regularization and multivariate approaches improve its robustness. This paper reviews the basics of linear regression, its applications, and methods to strengthen model accuracy.**

***Index Terms*—Linear regression, multivariable, mean squared error**

I. Introduction

Linear regression is a useful tool for predicting a quantitative response, is still a useful and widely used statistical learning method [2]. In essence, it aim to predict a continuous target variable by finding the best-fitting line through the data points, minimizing the distance (or error) between the predicted values and the actual data points. This is achieved by adjusting the slope and intercept of the line to reduce the error, typically measured by a cost function such as Mean Squared Error (MSE) [**?**]. Linear regression is widely used not only for predictive analysis but also as a diagnostic tool to understand relationships between variables, particularly in areas like economics, biology, and social sciences. It serves as a baseline in machine learning to assess more complex models, and although it has limitations [**?**].

II. Material and methods

A. Data processing

There where analysed two different datasets:

- *turkish-se-SP500vsMSCI* contains, in the first column, historical data on the returns of the S&P 500 index (USA) and, in the second, the MSCI Europe index, representing the U.S. and European stock markets.
- *mtcarsdata-4features* contains four features from the mtcars dataset, which collects data on various car models. The four main variables include *mpg* (Miles Per Gallon, as an indicator of fuel efficiency), *weight* (vehicle weight), *hp* (horsepower), and *disp* (engine displacement). This dataset is used to analyze how a car's mechanical characteristics affect fuel consumption.

Each dataset is initialized to store training and test subsets. The data is randomly shuffled before splitting into training and test subsets. For the Turkish dataset, ten subsets are created for cross-validation purposes, while the MTK data is split into training (5%) and test (95%) sets. To show the difference, is selected data from the beginning and end of the Turkish dataset rather than random samples, as nearby data points are often more similar.

B. Linear Regression Model

The presented Linear Regression Model, works in two different methods. For the Turkish dataset the LRM's efficiency is evaluated using ten random subsets, each one a different 10% of the dataset (Figure 1). Instead, with the MTK dataset the model is implemented in two variables and multivariable linear regression represented in Figures 2, 3, and 4. Another evaluation taken to account is the applications of a intercept value.

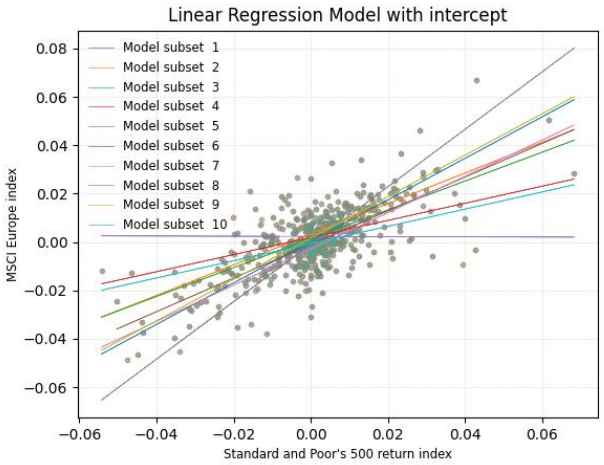

Fig. 1.

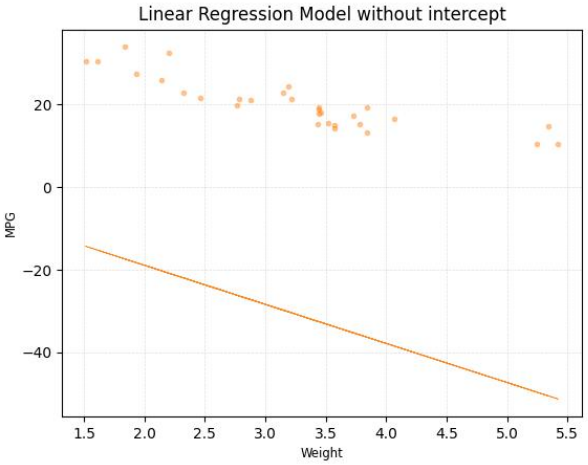

Fig. 2.

1) Turkish dataset: is shuffled and split in two subsets (one for training and one for testing). The training's subsets takes each the 5% of the entire dataset, meanwhile the testing's subsets the remainder 95%.

2) MTK dataset: is analyzed in two different ways: first, considering only the first two columns (*weight* and *mpg*), and then considering all four columns (*disp*, *hp*, *weight*, and *mpg*) where there is applied a multidimensional linear regression. This allows us to evaluate the impact of additional features on the linear regression model.

The performance of the linear regression model is also evaluated in two different scenarios: with and without intercept.

Linear regression use the concept of the mathematical relation

$$Y = \beta_0 + \beta_1 X$$

where Y is the dependent variable, X is the independent variable, β_0 is the intercept, β_1 is the slope of the line [1]. But, unless the independent and dependent variables are describable as only two point in the domain the previous relation can be found only introducing a error term ϵ.

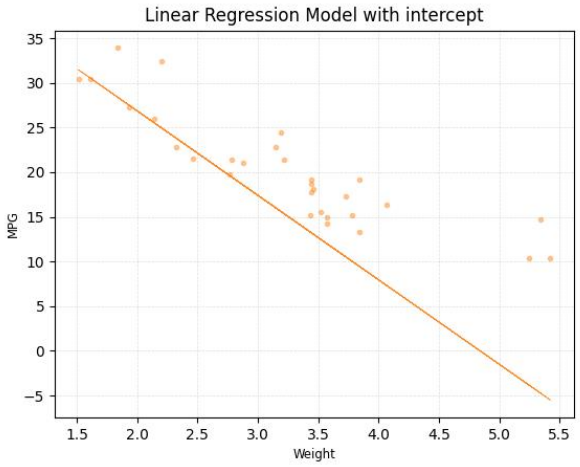

Fig. 3.

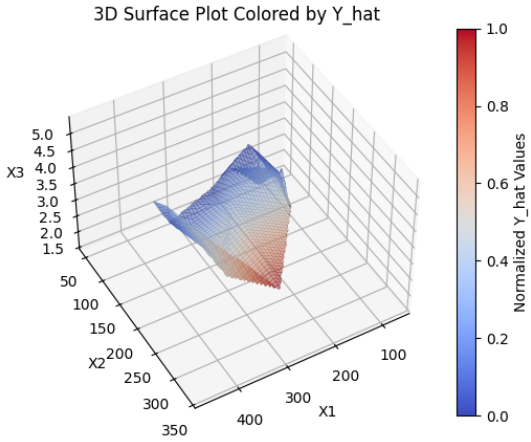

Fig. 4.

Multivariable linear regression is an extension of simple linear regression and is used to model the relation between a dependent variable (or target) and two or more independent variables (or predictors)

C. Model evaluation

III. Results and Conclusion

References

[1] G. James, D. Witten, T. Hastie, and R. Tibshirani, "Linear Regression," in *An Introduction to Statistical Learning: With Applications in R*, G. James, D. Witten, T. Hastie, and R. Tibshirani, Eds. New York, NY: Springer US, 2021, pp. 59–128.

[2] G. James, D. Witten, T. Hastie, R. Tibshirani, and J. Taylor, "Linear regression," in *An introduction to statistical learning: With applications in python*. Springer, 2023, pp. 69–134.