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.

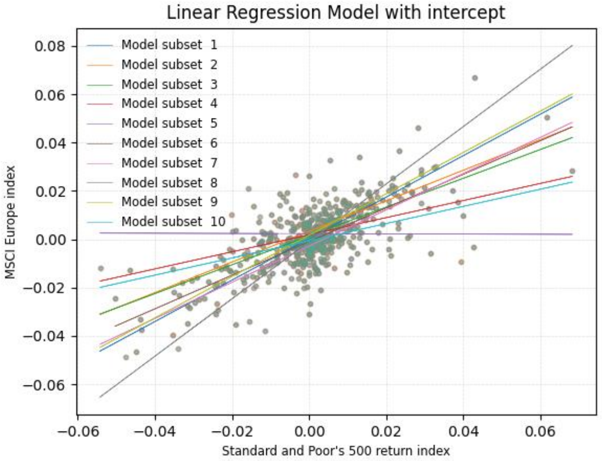

Fig. 1. Graphical representation of the model trained on ten subsets with interception, each corresponding to 5% of the Turkish dataset. In the background are visualized the corrispective test subsets (95% of the dataset).

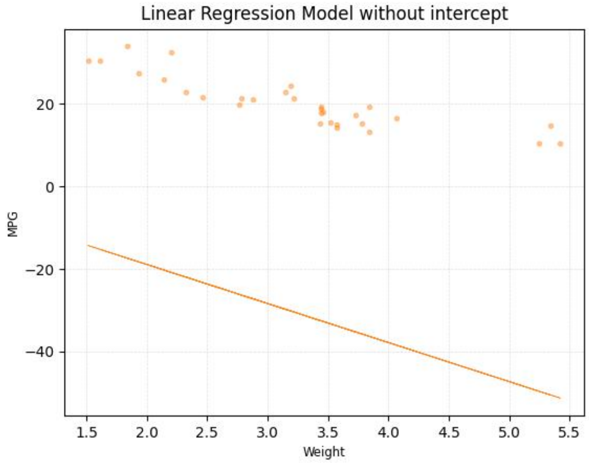

Fig. 2. Graphical representation of the model trained without interception, which corresponds to 5% of the MTK dataset. In the background is visible the test subset (95% of the dataset).

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

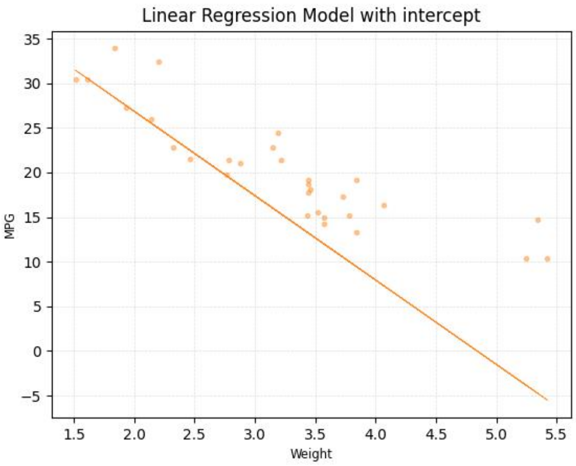

Fig. 3. Graphical representation of the model trained with interception, which corresponds to 5% of the MTK dataset. In the background is visible the test subset (95% of the dataset).

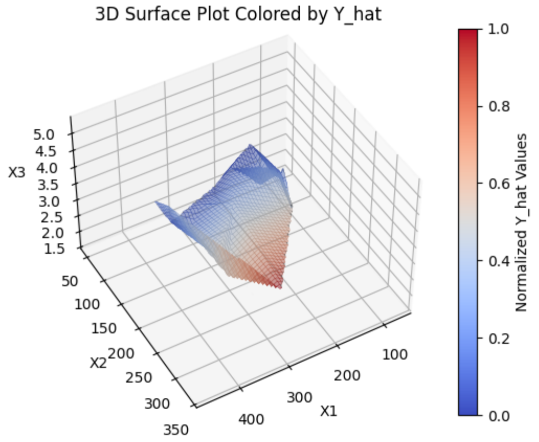

Fig. 4. Graphical representation in fourth dimensional representation, where the fourth one is the predicted output. The model, trained with interception, corresponds to 5% of the MTK dataset.

two point in the domain the previous relation can be found only introducing an error term ε.

$$y = \beta_0 + \beta_1 x + \varepsilon \tag{1}$$

The intercept β_0 in linear regression is the value of y when all predictors x are zero. Including the intercept allows the model to better fit the data, especially when it's unrealistic to assume that $y = 0$ in the absence of independent variables, which forces the model to pass through the origin. It is useful only if there are theoretical reasons for y to be zero when the predictors are zero. In linear regression, ϵ represents the *model's error* or *residual term* [1]. Epsilon captures all the variations in y that cannot be predicted by the independent variables used in the model. Mathematically, epsilon is the difference between the observed value of y and the value predicted by the linear model:

$$\varepsilon = y_{\text{observed}} - y_{\text{predicted}}$$

The ε term is crucial because it highlights the model's limitations and suggests that, despite efforts to model the data, there will always be unpredictable elements or unconsidered factors that influence the outcome.

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).

$$y = X\mathbf{w} \tag{2}$$

where X are the independent variables represented by matrix form and $\mathbf{w}$ is the slope's vector. If it considered the intercept must be added a unit column at the beginning of the X matrix.
Another important point considered to implement the model is the computation of the pseudo-inverse in case of exceptions. In particular, the pseudo-inverse (which is a generalization of the matrix inverse, without the condition of invertibility) is computed when, during the computation of the gradient, the matrix is not invertible. This means that the system of linear equations would not have solutions, and the pseudo-inverse is used to provide an approximate solution.

C. Model evaluation

Mean Squared Error (MSE) is a key metric that quantifies the overall accuracy of the model. It is computed as the average of the squared differences between the observed values and the predicted values across all data points:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i^2 \tag{3}$$

MSE is used as a *loss function* to minimize the model error during the training, and it provides an aggregate measure of the prediction error across the dataset.

III. Results

From Figure 2 and 3 it is visible that, training the model with only two classes (i.e. *weight* and *mpg*) of the MTK data is notably different the use of the intercept. In the case of the training model without intercept the linear regression line of the predicted values does not intercept any point. In contrast, Figure 4 presents challenges in obtaining readily comprehensible results. This may be attributed to the limited data set utilized for the regression analysis. Analyzing the Turkish dataset, a median of $2.44 \cdot 10^{-4}$ was calculated for the mean squared error (MSE) obtained from the ten subsets. The first quartile was $2.20 \cdot 10^{-4}$, and the third quartile was $2.65 \cdot 10^{-4}$. These values were obtained for both cases analyzed: with and without the inclusion of an intercept.

IV. Conclusion

In this report, we presented a linear regression analysis of two datasets: Turkish and MTK. Our results show that the linear regression model is able to capture the relationship between the variables in both datasets. For the Turkish dataset, we found that the model is able to explain a significant portion of the variance in the data, with a mean squared error (MSE) of $2.44 \cdot 10^{-4}$. For the MTK dataset, we analyzed the data in two different ways: considering only the first two columns (*weight* and *mpg*), and considering all four columns (*disp*, *hp*, *weight*, and *mpg*). Our results show that the model performs better when considering all four columns, with an MSE of 68.40 and 225.28. This suggests that the additional features in the dataset provide valuable information for predicting the target variable. Overall, our results demonstrate the effectiveness of linear regression in modeling the relationships between variables in these datasets. However, we also note that the MSE values are relatively small, indicating that the models may not be capturing all of the underlying patterns in the data.

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.