Linear Regression

Alice Rivi
Robotics Engineering
University of Genoa, Genoa
S5135011@studenti.unige.it

***Abstract*—Implementing a linear regression using the provided quantity of data is the assignment's objective. Regression is the process of using measured data to approximate a functional dependence. The first task that needed to be completed involved gathering the data and putting them ready for the programme, in this case MATLAB, to read and use them. The first data set includes two columns, the first representing the input and the second the output; the second data set has four different columns, each corresponding to mpg (miles per gallon), disp (displacement), hp (horse power), and weight.**

Then we had to complete the second task:

- **first, on the first data set, we have to implement a one-dimensional issue without an intercept;**
- **then we needed to find a regression without an intercept and compare it graphically with a 10% random subset of the total dimension;**
- **in the third point, we used the weight and mpg columns from the second data set, respectively, and computed a new one-dimensional issue utilising the weight as the output and the mpg data as the input.**
- **the final bullet dealt with a multi-dimensional problem that involved predicting the mpg, the first column, using the final three columns of the second data set.**

The third assignment was redoing Task 2's Points 1, 3, and 4 using 5% of the total data provided this time. Then, using both the training data and the remaining 95% of the data, we had to determine the objective, mean square error.

In order to fully discuss the programme, all of the results had to be shown on a graph or in a table after the programme had been tested numerous times, such as 10 times.

1. Introduction

Regression analysis is a set of statistical procedures used in statistical modelling to estimate the associations between a dependent variable—often referred to as the outcome or response variable—and one or more independent variables. In linear regression, the most typical type of regression analysis, the line or a more complicated linear combination that most closely matches the data in terms of a given mathematical criterion is found.

This enables the researcher to calculate the conditional expectation of the dependent variable under precise mathematical assumptions when the independent variables have a particular set of values.

2. Getting data

The first task required us to gather the data from the provided data sets and prepare them for our software, MATLAB.

Two distinct data sets were provided to us, the first of which contained just two columns—the first for input and the second for output. We may read them using the $readmatrix$ function in MATLAB, which turns the data set into a usable matrix.

The second data set was slightly more challenging because it included three more numeric columns in addition to the first literal column. The data set was converted into a matrix using the same procedure as previously, and the first column was then removed because it was unnecessary for us and was unreadable by our software. So, at this point, all of the data was available and prepared for processing.

3. Linear Regression Mode

There were four separate points to achieve for this job. In the first, using the first data set provided, we had to build a linear regression of a one-dimensional problem. Because the data set's mean value was zero, we did not need to take the intercept of the linear regression into account in this instance.

We then divided all the outcome values (first column) and all the independent values to calculate the angular coefficient, which was stored in a vector with the appropriate dimensions (the second column). Then, by merely multiplying the angular coefficient obtained with the data set's x-values, we could compute the regression's result. It goes like this:

$$y = wx \tag{1}$$

A graph was created and contained all of the results. The second point was just a comparison of the prior finding with a problem of a similar nature computed with a subset of 10% the original set's dimension. The computation process was identical throughout, and the outcome in this instance was also plotted on a graph.

In the third stage of the work, the issue remained a one-dimensional one, but this time, the computation of the intercept was also required. We followed the same process, modifying the computation of X—all of the independent variable's values—only slightly. The formula used to calculate the result in this instance is as follows:

$$y = w_1 x + w_0 \tag{2}$$

where w_0 is the intercept of the line.

The result was then plotted in a graph, with the result attending, and the angular coefficient and computation of the outcome were the same as before.

The last point was the most challenging because it involved a multi-dimensional problem; in this case, the target column was the data set's first column, while the other three columns served as the input variables. In this instance, the intercept has to be calculated as well.

Because the software uses more understandable code, it uses a slightly modified formula to determine the angular coefficient. However, the outcome remains the same; the theoretical model is as follows:

$$\mathbf{w} = (X^T X)^{-1} X^T \mathbf{t} \tag{3}$$

where $\mathbf{t}$ represents the target vector.

When the w vector is calculated, the result is always the input times the angular coefficient.

The diagonal of the square with sides ranging from the lowest to the highest value is approximated by the graph in this situation; the diagonal is the exact diagonal used as a reference.

With each of the following graphs, we present a broad overview of the outcome:

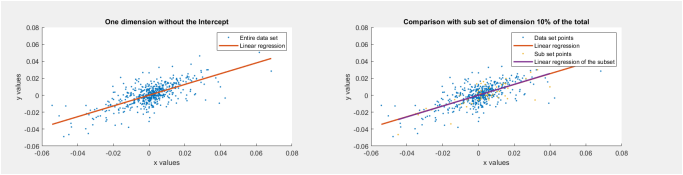

Figure 1. Plots pertaining to the first and second tasks points: The first is the complete data set, while the second is a comparison using a subset that is 10% of the original data set's dimension.

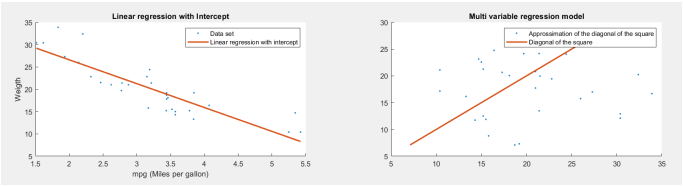

Figure 2. Plots pertaining to the first and second tasks points: The first is the complete data set, while the second is a comparison using a subset that is 10% of the original data set's dimension.

4. Test Regression Model

The final job required us to verify the regression model we had created by computing errors on two subsets of the data.

The set was first split into two parts: the training test was comprised of the first 5–10% of the data, and the remaining 95–90% was the test set. We have to execute this for a specific number of iterations—for instance, 10 distinct iterations—in order to test the behaviour because the division is randomly computed.

In order to complete the task, we calculated the angular coefficient using the test data, and using the resultant value, we estimated the outcome using all of the training set's goal values. Then, using the Mean Squared Error between the calculated outcome and the set's target, the error was calculated. Of course, both the test set and the training set were used.

A histogram graph was used to plot all the results at the conclusion to display the mistake [Figure 3].

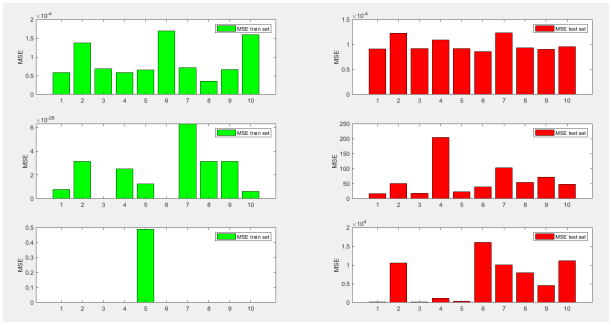

Figure 3. Histograms for linear regression obtained.

Naturally, as the image demonstrates, the training set's results are significantly inferior to those of the test set due to the division's low weighting.