# LINEAR REGRESSION

Student Gianluca Galvagni, S5521188, Università degli Studi di Genova

***Abstract*—How second assignment, I talked about linear regression. Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable. My goal for this assignment was created a code which can make by different data sets and then give me the error between the dates calculated by the algorithm and the test set.**

***Index Terms*—Linear regression, Square error.**

## I. Introduction

LINEAR REGRESSION means approximating a functional dependency based on measured data. With terms ”linear” I indicates that the approximation will be a linear one (a line in the one-dimensional case). In particular, given a certain data set, I try to understand how the target *t* is correlated with the observations *x*. With this prevision, I can estimate the target given some observations. The problem I have faced was to implement the algorithm on Matlab and then to analyze the results.

## II. One-dimensional input without intercept

In this first part, I want to build a linear model *y(x)* that predicts *t* given *x*, so:

$$t \approx y \ where \ y = wx$$

Where *t* is the target, *y* the prediction, *x* the observations that it is a vector because I am in the one-dimensional case.
I want *y(x)* to be similar to *t(x)* for any *x*. For instance, I would like to have $t_1 \approx y_1$ where $y_1 = w\,x_1$ or $t_2 \approx y_2$ where $y_2 = w\,x_2$. For that, with a certain observation $x_1$ and a certain related target $t_1$ I found the parameter $w$ as:

$$w = \frac{t_1}{x_1}$$

But given another couple $x_2, y_2$ the resulting *w* will be different: It is impossible to find a perfect value for *w*.
The goal is to find a correct parameter *w* that has the most suitable value based on a certain optimization. I need to quantify how wrong is each estimate based on certain measure and make this measure as small as possible on average.
The square error $(t$ - $y)^2$ is a good choice because is differentiable and even. It also give more weight to large error respect to smaller ones. Having N couples $t_l, y_l$ I have to minimize the mean value of the loss over the whole data set:

$$\frac{d}{dw}J_{MSE} = 0 \quad where$$

$$J_{MSE} = \frac{1}{N}\sum_{l=1}^{N}\lambda_{SE}(y_1, t_1) \quad and \quad \lambda_{SE}(y, t) = (y - t)^2$$

After some passages, I arrive to this final solution:

$$w = \frac{\sum_{l=1}^{N} x_l t_l}{\sum_{l=1}^{N} x_l^2}$$

So, with a new observation *x*, I will be able to predict the target as:

$$y = w\,x$$

## III. One-dimensional input with intercept

With the model $y = w\,x$ I build a line that intercepts the origin. I can improve the model adding one parameter $w_0$ to make the line intercepting another point on $y$ axis:

$$y = w_1 x + w_0$$

The solution in this case can be found by *centering* around the means $\overline{x}$ and $\overline{t}$ of $x$ and $t$:

$$\overline{x} = \frac{1}{N}\sum_{l=1}^{N} x_l \qquad \overline{t} = \frac{1}{N}\sum_{l=1}^{N} t_l$$

The wanted parameters $w_0$ *(intercept)* and $w_1$ *(slope)* are:

$$w_1 = \frac{\sum_{l=1}^{N}(x_l - \overline{x})(t_l - \overline{t})}{\sum_{l=1}^{N}(x_1 - \overline{x})^2} \quad w_0 = \overline{t} - w_1\overline{x}$$

## IV. Multi-dimensional inputs

The formulas can be generalized for the multi-dimensional case. In this case each observation $x_l$ is now a vector of *d* elements($X$); thus, I have *d* parameters $w_i$ with *j = 1...d* plus a parameter $w_0$ to the intercept:

$$X = \begin{pmatrix} x_1 \\ x_2 \\ . \\ . \\ . \\ x_N \end{pmatrix} \qquad w = \begin{pmatrix} w_1 \\ w_2 \\ . \\ . \\ . \\ w_d \end{pmatrix}$$

The output of the model for a single observation will be:

$$Y_N = x_{N,1}\,w_1 + x_{N,2}\,w_2 + ... + x_{N,d}\,w_d$$

Combining all $Y_N$ I obtain:

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_N \end{pmatrix} = \begin{pmatrix} x_{1,1} & x_{1,2} & . & . & . & x_{1,d} \\ x_{2,1} & x_{2,2} & . & . & . & x_{2,d} \\ & & & . & & \\ & & & . & & \\ & & & . & & \\ x_{N,1} & x_{N,2} & . & . & . & x_{N,d} \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \\ . \\ . \\ . \\ w_d \end{pmatrix} = X\,w$$

Note the difference between the $N \times d$ matrix $X$ and the $N\times$ $(d{+}1)$ matrix $X$. As in the one-dimensional case, I have to minimize the mean square error objective $J_{MSE}$; but now I have $d$+1 parameters so I must set the gradient of $J_{MSE}$ equal to zero. After some passages I obtain:

$$\bigtriangledown J_{MSE} = \frac{\partial}{\partial w} J_{MSE} = X^T X w - X^T t = 0$$

and so I obtain the value for $w$:

$$w = (X^T X)^{-1} X^T t = X^{\dagger} t$$

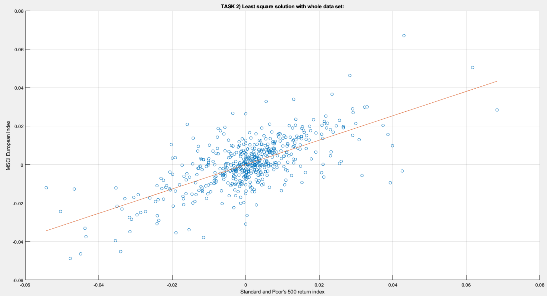
where $X^{\dagger}$ is the *Moore-Penrose pseudoinverse* of X. The computation of the *pseudoinverse* is not always numerical reliable. It is almost impossible that $X^T X$ is not invertible because it would mean that we have two identical observations on all columns (noise is always present). But I can also have problems if the variables are correlated (similar to each one): this would mean that $X^T X$ has high condition number and the pseudoinverse will be numerical unstable. In this case, there are solutions based on iterative computation by successive approximations.

## V. Lab activity

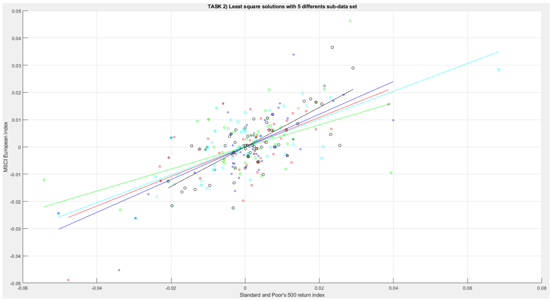
In this assignment I have used Matlab R2022b to make my program. I started from the dates processing with they download and some things to make them readable and then I solved the second and third tasks. The dates' name are *"turkish-se-SP500vsMSCI"* and *"mtcarsdata"*. I used the formulas show before in the points II, III and IV.

1) *TASK 2:*

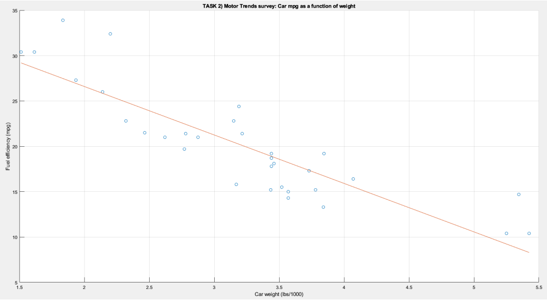
1 **One-dimensional problem without intercept on the Turkish stock exchange data.** I calculated the least square to the linear regression problem and then I plotted out the results.



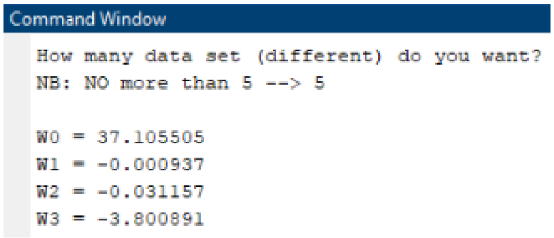
2 **Compare graphically the solution obtained on different random subsets (max 5) and then 10 percent of the whole data set.** I plotted up five different random subsets, with five different colors and with the same colors I plotted the linear model.



3 **One-dimensional problem with intercept on the Motor Trends car data, using columns mpg and weight.** I calculated $W0$ and $W1$ and then I plotted out the result.



4 **Multi-dimensional problem on the complete MTcars data, using all four columns (predict mpg with the other three columns).** Here I used all data base the t for the mpg and on the x with the rest of dates. For that point I used a Matlab's function to calculate $W0$, $W1$, $W2$ and $W3$.



```
How many data set (different) do you want?
NB: NO more than 5 --> 5

W0 = 37.105505
W1 = -0.000937
W2 = -0.031157
W3 = -3.800891
```

2) *TASK 3:* ***Re-run 1,3 and 4 from task 2 using a % of the data.***

5-6-7 I created the possibility to chose the percent of the two dates set and the numbers of different data set. And plotted all results in a graph for each points (Fig. 1).

## VI. Conclusion

At the end of this second assignment, there are some points of view to analyze.

- In the three first graphics you can see the linear regression with different dates set. In particularly, the plot with five different lines accentuates this behavior and how you can see all lines go for the point $(0, 0)$, we were waiting that result.
- For the fourth point I only printed in the command window the results of the operation ($W0$, $W1$, $W2$, $W3$).

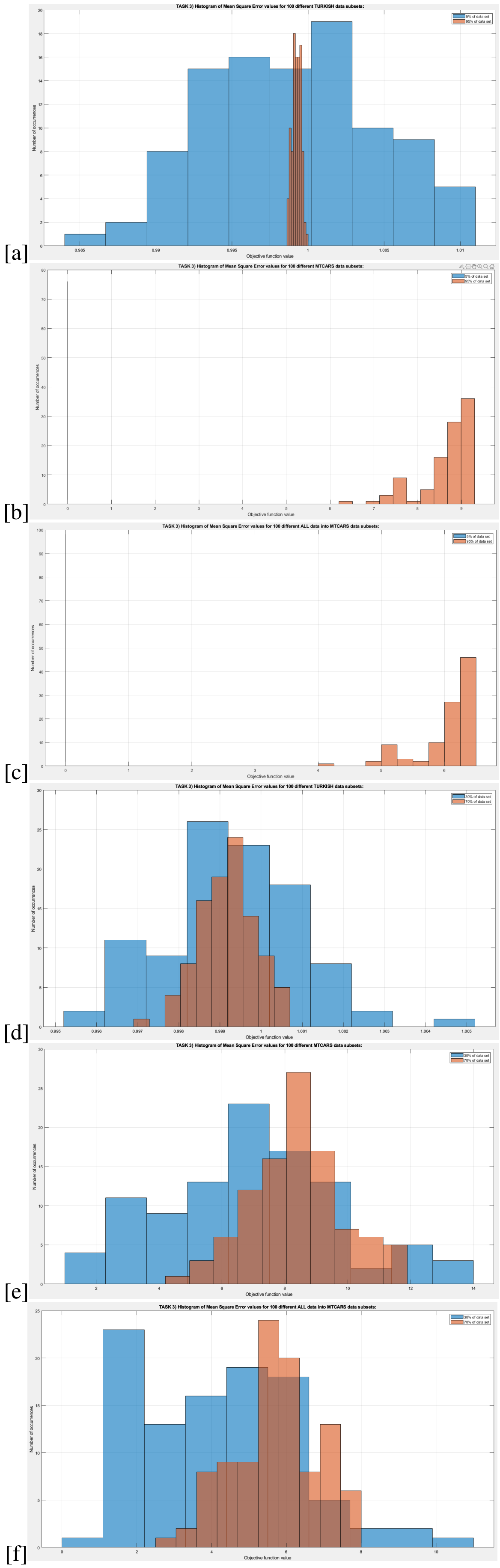


Fig. 1.
(a-b-c) BLUE data set = 5% and ORANGE data set = 95% .
(d-e-f) BLUE data set = 30% and ORANGE data set = 70%.
With 100 different dates set.

- In the last six bars there are many differences; Firstly, with only 5% of the data set the graphs [b] and [c] have not enough dates to print a good result. Secondly, in the plots [d],[e] and f there are the typical ”bell” and it grows up on the most repeated date. Finally, you can see the *”Objective function value”* is similar more or less always in each comparison [a]-[d], [b]-[e] and [c]-[f].

I reported only two of the possible running. Up to now, I assume that results acceptable for this assignment.