Assignment2: Linear regression

Saeed Kalateh, MSc Student, Robotic Engineering,

Abstract—In this assignment, linear regression has been implemented on two different data sets. First, one-dimensional problem without intercept on the Turkish stock exchange data is implemented. Then, solutions on different random subsets of the whole data set are compared graphically. Next, one-dimensional problem with intercept and multi-dimensional problem are tested on the Motor Trends car data.

Index Terms—Linear Regression

1 Turkish Data set

Two data sets are used for this assignment. For the first part, Turkish data set is used. It has 536 observations and 2 features.

1.1 Importing Data

To import the data, "," is used as delimiter and the members of data set are imported as floats. Again, the imported data is stored at a matrix named data.

1.2 One-dimensional without intercept

This problem obeys the following equation.

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

where x_l is the input data and t_l is the target. w is the weight of the linear regression or in fact, the slope of our fitted line. One column of data is stored in x matrix, while the other stored in t. The following code is used to code this problem.

$$w = (t' * x)/sum(x. \wedge 2)$$

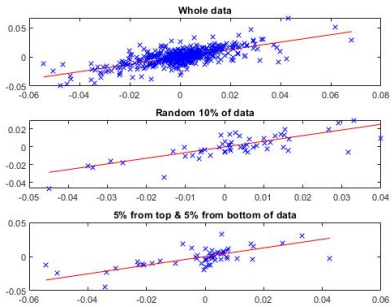

Fig. 1. Linear regresion of Turkish stock exchange

In figure 1, the fitted is plotted on three different subsets of data set. The first plot in Figure 1 show the fitted line on the whole data. The second one, shows the fitted line on 10% of data which was chosen randomly from the whole data set. The last one depicts the fitted line on the first and last 5% of the data set.

1.3 Test regression model

In this section, the objective (mean square error) is computed on both 95% and 5% of the data. Figure 2 shows the bar chart of the objective on both splits of the data set.

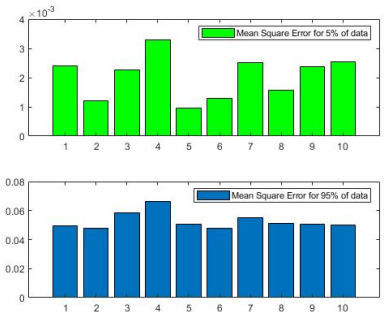

Fig. 2. Objective (mean square error) on 95% and 5% of the data

2 Motor Trends car data

The data has 32 observations and 4 attributes. The first column has been omitted to create a numeric matrix. In the following section, one dimensional regression with intercept and multi dimensional problem are tested on the data set.

2.1 One-dimensional with intercept

The line is fitted using mpg and weight columns, which are the first and fourth columns in imported data matrix. In this task, two values for w should be calculated. These formulas have been used to calculate these weights.

$$w_1 = \frac{\sum_{l=1}^{N} (x_l - \bar{x})(t_l - \bar{t})}{\sum_{l=1}^{N} (x_l - \bar{x})^2} \tag{2}$$

$$w_0 = \bar{t} - w_1 \bar{x} \tag{3}$$

First, the mean value for x and t should be calculated. Using these values, w_1 and w_2 are obtained, respectively. The line equation looks like this:

$$y = w_0 + x w_1 \tag{4}$$

where w_1 is the slope andw_0 is the intercept (bias) for the regression. Next, the data and the fitted line are plotted as shown in Figure 3.

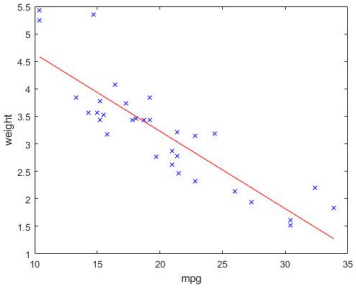

Fig. 3. One-dimensional problem with intercept on the Motor Trends car data

2.1.1 Test one-dimensional regression model

The data is separated using into two randoms subsets. As a result, 5% of the data is stored at $rand5perOfData$ variable in the code. The rest of observations are stored at $rand95perOfData$. The previous task is implemented using only 5% of data and tested on the other 95%. This task is repeated 10 times and the plot is as shown in figure 4.

2.1.2 Multi-dimensional problem on the MTcars data

In this section, the regression should predict mpg with the other three columns. That means, according to our code, the first column as the output (t) and the columns 2, 3, and 4 as inputs (x), where x is 32x3 and t is 32x1. This time, the w is calculated like this:

$$\mathbf{w} = \left(X^T X\right)^{-1} X^T \mathbf{t} \tag{5}$$

After w is calculated, y is obtained using:

$$y = wx \tag{6}$$

2.1.3 Test multi-dimensional regression model

In this section, we have to re-run the problem using only 5% of the data. Therefore, 5% of the data is stored at $rand5perOfData$ variable in the code.The rest of observations are stored at $rand95perOfData$. Calculating error for these partitions of data will result in the following bar graph. Figure 5 shows the mean square error for 5% and 95% of the data. The error is large for the 95% of data since we have only used 5% percent of the total observations and we have 3 attributes (multi-dimensions).
Previous test on one-dimensional data showed less mean square error since the algorithm had to learn fewer features. As a result, the error becomes huge.

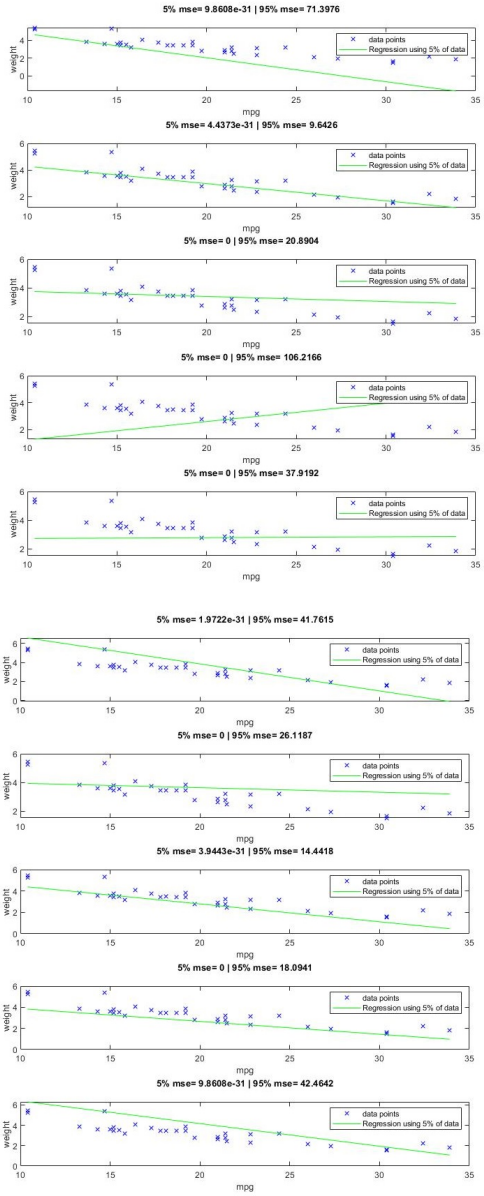

Fig. 4. Testing one-dimensional regression model using random 5% of data

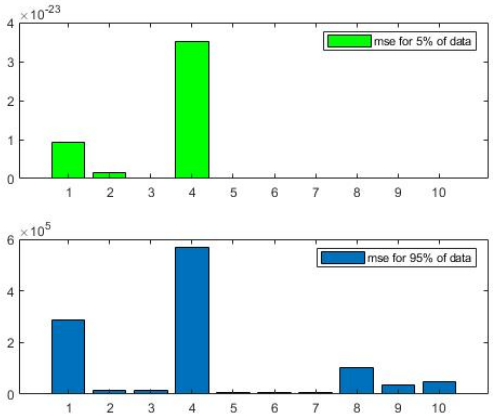

Fig. 5. Mean square error for 5% and 95% of the data