Report for Machine learning

Assignment 2 - S5089493 - Raghuveer Siddaraboina

Introduction

Linear Regression is a statistical supervised learning technique to predict the quantitative variable by forming a linear relationship with one or more independent features.

It helps determine:

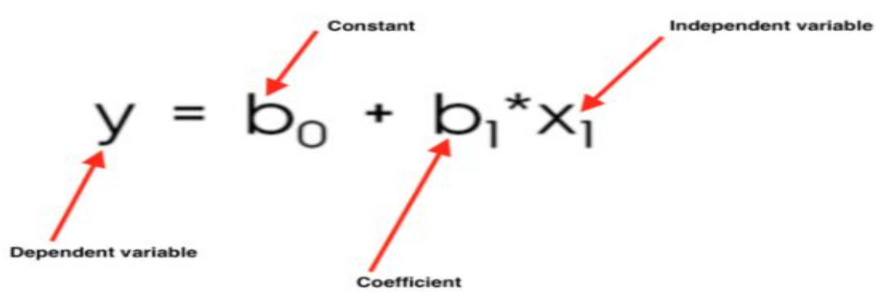
- If a independent variable does a good job in predicting the dependent variable.
- Which independent variable plays a significant role in predicting the dependent variable.

Types of Linear Regression

Simple Linear Regression:

Simple Linear Regression helps to find the linear relationship between two continuous variables,One independent and one dependent feature.

Formula can be represented



Multiple Linear Regression:

Multiple linear Regression is the most common form of linear regression analysis. As a predictive analysis, the multiple linear regression is used to explain the relationship between one continuous dependent variable and two or more independent variables.

The independent variables can be continuous or categorical (dummy coded as appropriate).

We Often use Multiple Linear Regression to do any kind of predictive analysis as the data we get has more than 1 independent features to it.

Formula can be represented as

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \epsilon$$

where, for $i = n$ observations:

$y_i =$ dependent variable

$x_i =$ expanatory variables

$\beta_0 =$ y-intercept (constant term)

$\beta_p =$ slope coefficients for each explanatory variable

$\epsilon =$ the model's error term (also known as the residuals)

In a regression context, the slope is very important in the equation because it tells you how much you can expect Y to change as X increases

Assignment tasks

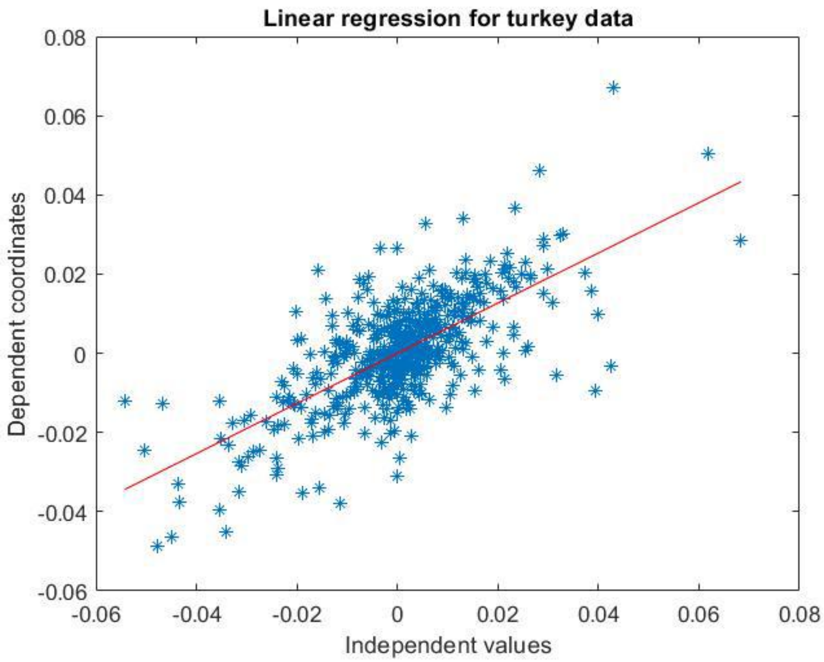
Task 2 : Filt a linear regression model

1. One dimensional problem without intrcept of turkish stock exchange data.

The basic equation for a line with slope and intercept is used for this task.But,as per instructions intercept is set to zero i.e the equation reduces to

y = m*x

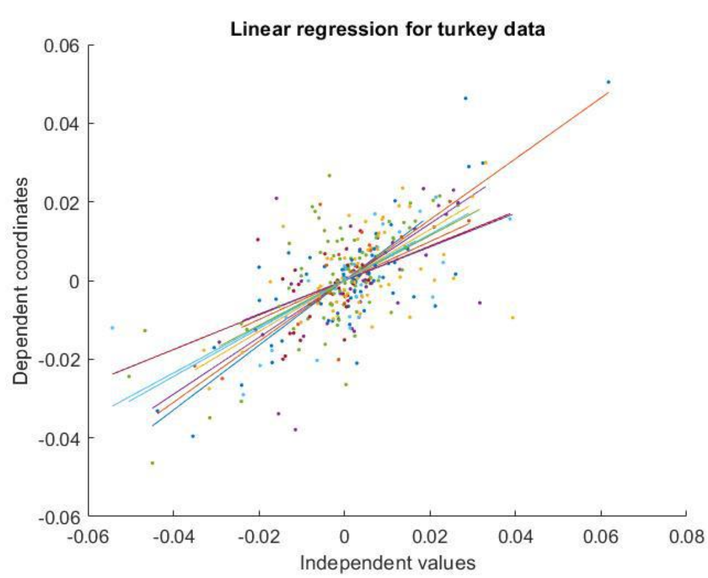
From equation and the data x is taken as first collumn and y is taken as second column and the below result is obtained

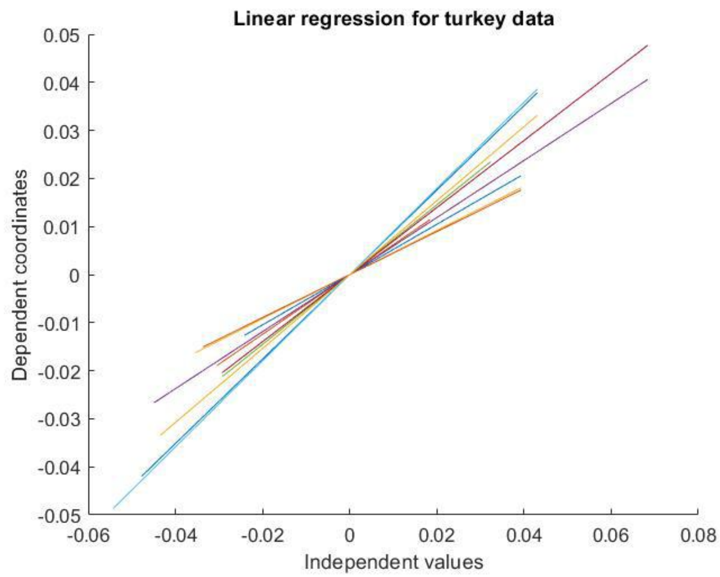


The slope obtained in the above task is **0.6339**.

2. Compare graphically the solution obtained on different random subsets (10%) of the whole data set

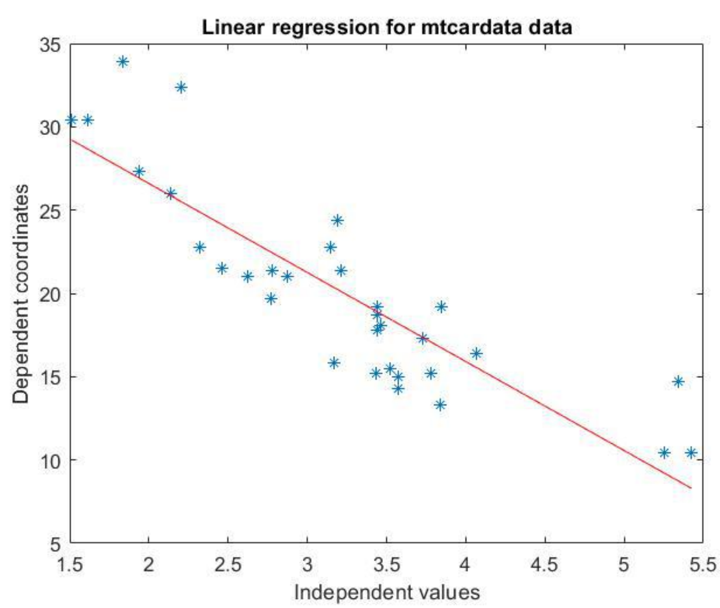
The graph obtained by performing this task we gat the below graph. First graph is with the points and lines for the second graph only lines are shown for clarity.





From the above graph we can see that each result of a random subset has a **different slope** but all of them has same intercept i.e, **zero**.

3. One-dimensional problem with intercept on the Motor Trends car data, using columns mpg and weight



The above graph is the result of the task 2.3 from which we get the value of **slope** as **-5.3445** and the **intercept** of the line is **37.2851**.

4. Multi-dimensional problem on the complete MTcars data, using all four columns (predict mpg with the other three columns)

	Predicted values	Target values	error
1	23.5700	21	2.5700
2	22.6008	21	1.6008
3	25.2887	22.8000	2.4887
4	21.2167	21.4000	0.1833
5	18.2407	18.7000	0.4593
6	20.4722	18.1000	2.3722
7	15.5656	14.3000	1.2656
8	22.9115	24.4000	1.4885
9	22.0409	22.8000	0.7591
10	20.0411	19.2000	0.8411
11	20.0411	17.8000	2.2411
12	15.7693	16.4000	0.6307
13	17.0616	17.3000	0.2384
14	16.8715	15.2000	1.6715
15	10.3215	10.4000	0.0785
16	9.3598	10.4000	1.0402
17	9.2115	14.7000	5.4885
18	26.6135	32.4000	5.7865
19	29.2760	30.4000	1.1240
20	28.0391	33.9000	5.8609
21	24.6016	21.5000	3.1016
22	18.7549	15.5000	3.2549
23	19.0911	15.2000	3.8911
24	14.5488	13.3000	1.2488
25	16.6639	19.2000	2.5361
26	27.6204	27.3000	0.3204
27	26.0236	26	0.0236
28	27.7450	30.4000	2.6550
29	16.5025	15.8000	0.7025
30	20.9888	19.7000	1.2888
31	12.8168	15	2.1832
32	23.0296	21.4000	1.6296

For the multi dimensional problem the results are shown in the above table which shows the predicted values of mpg, target values and error value.

Task 3 : Test regression model

In this task the data is split in to random subsets as 10% of training data and 90% of test data (cause 5% is too small)

The objective function is calculated in this task.This is mean squared error. It tends to amplify the impact of outliers on the model's accuracy.

For the linear regression data the objective function value is calculated by using

$$J_{\mathrm{MSE}} = \frac{1}{N}\sum_{l=1}^{N}(t_l - y_l)^2 \ .$$

For the multiple linear regression data the objective function value is calculated by using

$$\begin{aligned} J_{\mathrm{MSE}} &= \frac{1}{2}\|\mathbf{t}-\mathbf{y}\|^2 \\ &= \frac{1}{2}\|\mathbf{t}-X\mathbf{w}\|^2 \\ &= \frac{1}{2}(\mathbf{t}-X\mathbf{w})^T(\mathbf{t}-X\mathbf{w}) \\ &= \frac{1}{2}(\mathbf{t}^T-\mathbf{w}^TX^T)(\mathbf{t}-X\mathbf{w}) \\ &= \frac{1}{2}\left(\mathbf{t}^T\mathbf{t}-\mathbf{t}^TX\mathbf{w}-\mathbf{w}^TX^T\mathbf{t}+\mathbf{w}^TX^TX\mathbf{w}\right) \\ &= \frac{1}{2}\|\mathbf{t}\|^2-\mathbf{w}^TX^T\mathbf{t}+\frac{1}{2}\|X\mathbf{w}\|^2 \end{aligned}$$

For the testing and training of the data the matlab codes are named as task3_1.m, task3_3.m, task3_4.m

For the last task i.e, Repeat for different training-test random splits.A matlab code is made named test.m and in the code in order to choose change the value of task as required i.e, chose values from 1,3 and 4.The test.m script runs one of the three functions named task1,task3 and task4.

The following figure shows the values of MSE task 4 for a random set repeated for 10 times.

	trainMSE values	testMSE values
1	-2.2737e-13	95.9279
2	-1.1369e-13	96.5124
3	2.2737e-13	88.7133
4	0.4900	89.9060
5	4.5475e-13	76.1471
6	1.1369e-13	92.7122
7	0	90.7547
8	5.6843e-14	91.0729
9	5.6843e-14	74.0427
10	-4.5475e-13	77.7603