Exercise 5.1 (May 15, 2020)

-B9TB1707

Question:

We were tasked to plot the data according to the table given below.

	Nobel laureates per capita	Chocolate consumption per capita (kg/y/head)
Sweden	31.855	6.6
Switzerland	31.544	10.8
Denmark	25.255	8.6
Austria	24.332	7.9
Norway	23.368	9.8
UK	18.875	10.3
Ireland	12.706	8.8
Germany	12.668	11.4
USA	10.706	5.1
Hungary	9.038	3.5
France	8.99	7.4
Belgium	8.622	6.8
Finland	7.6	7
Australia	5.451	6
Italy	3.265	3.3
Poland	3.124	4.5
Lithuania	2.836	6.1
Greece	1.857	4.5
Portugal	1.855	4.5
Spain	1.701	3.3
Japan	1.492	2.2
Bulgaria	1.421	2.2
Brazil	0.05	2.5

After plotting the data, we were tasked to draw the line of best fit calculated by the method of least squares. Then we were asked to add a data point generated by using our student ID, and find how the line of best fit changes.

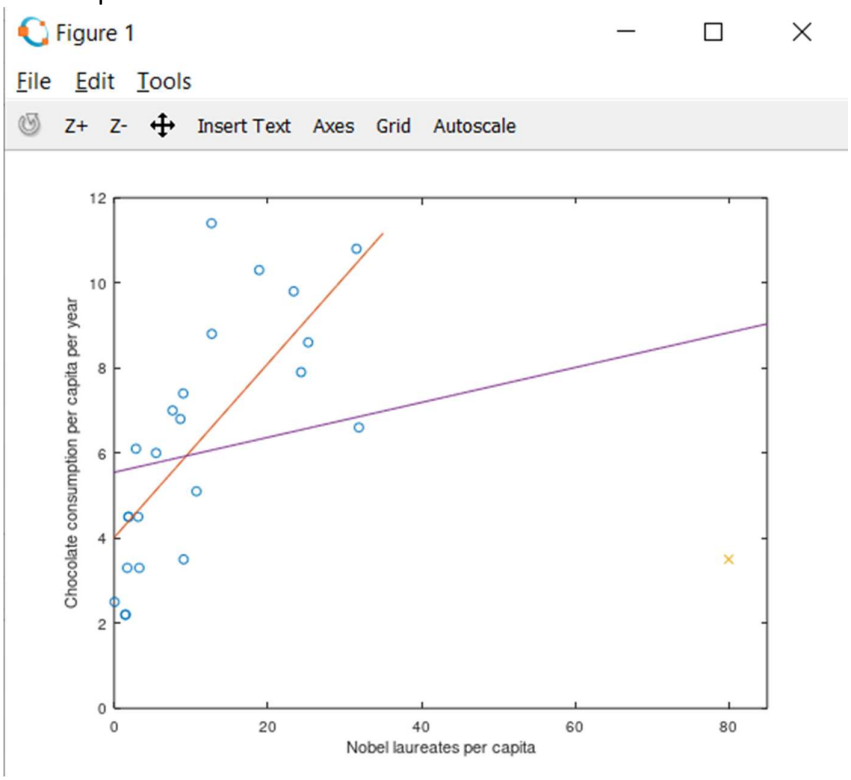
Solution:

My code for the solution is as follows:

CAPS_05_B9TB1707_5.1.m

```
load ('Nobel_vs_choco.txt')
nobel=data(:,1);
choco=data(:,2);
plot(nobel,choco,"o")
axis([0,85,0,12])
xlabel('Nobel laureates per capita')
ylabel('Chocolate consumption per capita per year')
X=ones(length(nobel),2);
X(:,1)=nobel;
Y=choco;
p=pinv(X)*Y;
xx=0:1:35;
hold on;
plot(xx,xx*p(1)+p(2))
my_nobel = 10*(1+7);
my_choco = .5*(0+7);
hold on;
plot(my_nobel,my_choco,"x");
X = [X; my_nobel 1];
Y = [Y; my_choco];
p=pinv(X)*Y;
xx=0:1:85;
hold on;
plot(xx,xx*p(1)+p(2))
```

The output is as follows:



Where X is the new custom generated data point and O is the data points from the given table. The red line is the line of regression of the original data set and the violet line is the new line of best fit after the introduction of the new data point.

How it works:

1. Line 1 imports the data set from the text file using the function load()
2. Lines 2 and 3 declares and initializes two arrays with the column 1 to “nobel” and “choco” respectively.
3. Line 4 plots each data point with the marker O
4. Line 5 specifies the size of the graph to be displayed.
5. Lines 6 and 7 names the axes.
6. Line 8 declares an array of size number of rows X number of columns in the data set. This is accomplished by length() function and specifying the number of columns, in this case 2.
7. Line 9 fills the first column of X with entries of nobel.
8. Line 10 declares a new array Y that contains the same values of choco
9. Line 11 computes the pseudo inverse
10. Line 12 declares a domain and interval increase.
11. Line 13 use the hold on command to retain plot data and draw over existing information.
12. Lines 15 and 16 generate the custom data entry.
13. Lines 17 and 18 plot the custom data point.
14. Lines 19 and 20 add the custom data point to X and Y

15. Lines 21 to 24 calculate and plot the new line of best fit in the same way as lines 11 to 14.

Conclusion:

The code above uses the method based on the Moore-Penrose pseudo-inverse method of regression. This method plots this line of best fit by finding the line that has least square of the difference between the points and the line. We assumed the relation between chocolate consumed and Nobel laureates is linear, but after seeing the plot it doesn't seem to be so. Therefore I believe we can produce a more useful analysis if we make a curved plot.