CSI5180. Topics in Artificial Intelligence Machine Learning for Bioinformatics Applications Fall2019

Assignment 2 Report

Question 1.1

1. Model.py
   - Implementing class linearRegression and methods as in the assignment
   - Global variables- Samples(list of samples),Labels(list of labels),theta0 and theta1
   - Implementing each method as mentioned in assignment
2. a2q11.py
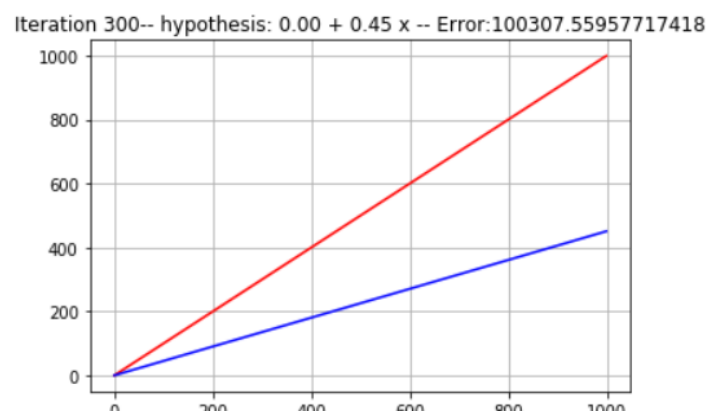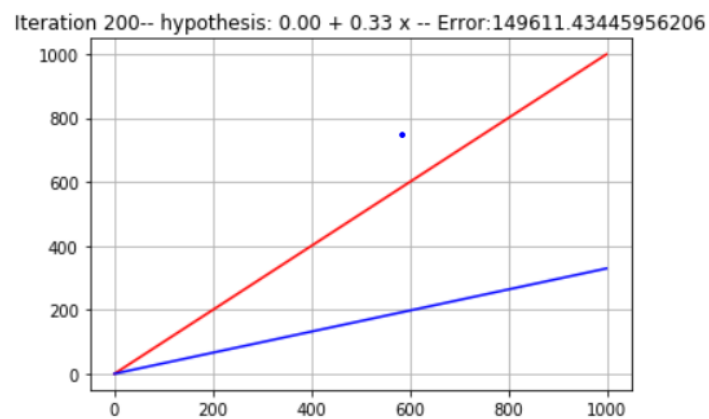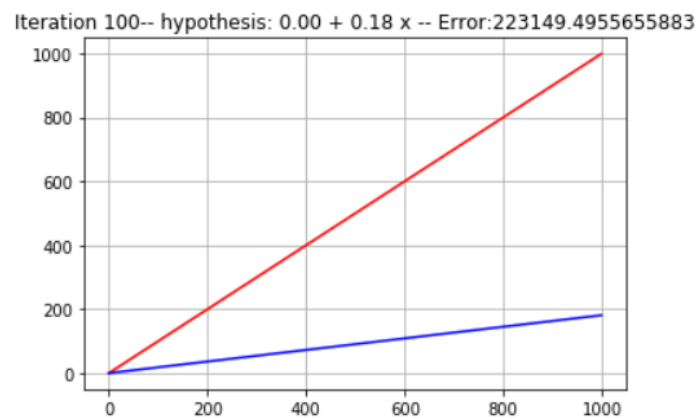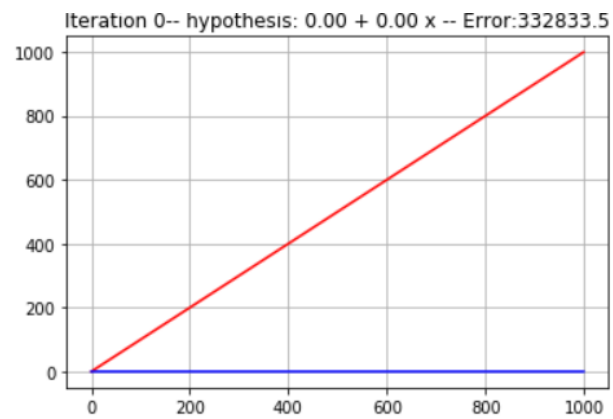   - Importing and creating an object of class linearRegression in setline method
   - X = 0 to X = 999
   - Y = X
   - Adding samples using addSample method – Adding features and labels
   - α = 0.000000003
   - numOfSteps = 100 (no of iterations of fit method per loop)
   - looped 50 times (no of times fit is called)
   - printing current hypothesis and cost function

```
importing Jupyter notebook from model.ipynb
Current hypothesis:  0.00 + 0.00 x ,  cost =  332833.5000
Current hypothesis:  0.00 + 0.18 x ,  cost =  223149.4956
Current hypothesis:  0.00 + 0.33 x ,  cost =  149611.4345
Current hypothesis:  0.00 + 0.45 x ,  cost =  100307.5596
Current hypothesis:  0.00 + 0.55 x ,  cost =  67251.5877
Current hypothesis:  0.00 + 0.63 x ,  cost =  45089.0847
Current hypothesis:  0.00 + 0.70 x ,  cost =  30230.1496
Current hypothesis:  0.00 + 0.75 x ,  cost =  20267.9197
Current hypothesis:  0.00 + 0.80 x ,  cost =  13588.7044
Current hypothesis:  0.00 + 0.83 x ,  cost =  9110.5989
Current hypothesis:  0.00 + 0.86 x ,  cost =  6108.2360
Current hypothesis:  0.00 + 0.89 x ,  cost =  4095.2902
Current hypothesis:  0.00 + 0.91 x ,  cost =  2745.7030
Current hypothesis:  0.00 + 0.93 x ,  cost =  1840.8671
Current hypothesis:  0.00 + 0.94 x ,  cost =  1234.2164
.
.
.
```
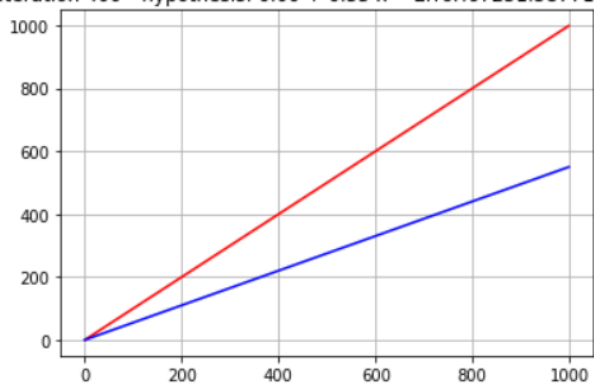
```
Current hypothesis:  0.00 + 1.00 x ,  cost =  0.0170
Current hypothesis:  0.00 + 1.00 x ,  cost =  0.0114
Current hypothesis:  0.00 + 1.00 x ,  cost =  0.0076
Current hypothesis:  0.00 + 1.00 x ,  cost =  0.0051
Current hypothesis:  0.00 + 1.00 x ,  cost =  0.0034
Current hypothesis:  0.00 + 1.00 x ,  cost =  0.0023
Current hypothesis:  0.00 + 1.00 x ,  cost =  0.0015
Current hypothesis:  0.00 + 1.00 x ,  cost =  0.0010
Current hypothesis:  0.00 + 1.00 x ,  cost =  0.0007
```

BONUS:
The blue line in the below graphs represents the red line in the assignment figures

Iteration 0-- hypothesis: 0.00 + 0.00 x -- Error:332833.5

Iteration 100-- hypothesis: 0.00 + 0.18 x -- Error:223149.4955655883

Iteration 200-- hypothesis: 0.00 + 0.33 x -- Error:149611.43445956206

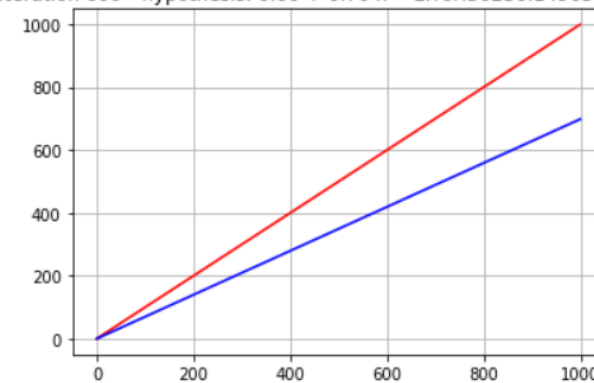Iteration 300-- hypothesis: 0.00 + 0.45 x -- Error:100307.55957717418

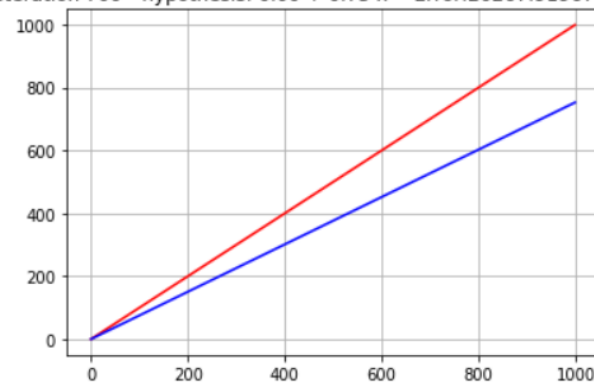Iteration 400-- hypothesis: 0.00 + 0.55 x -- Error:67251.5877190991

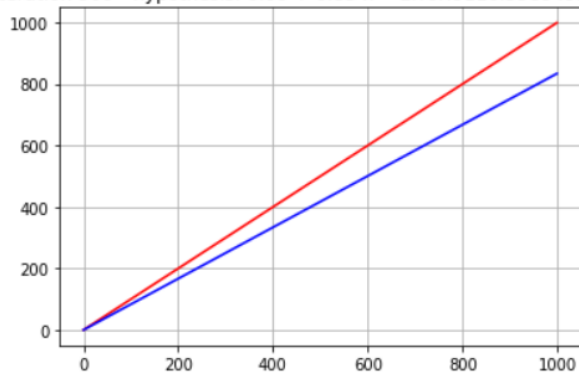Iteration 500-- hypothesis: 0.00 + 0.63 x -- Error:45089.08470917492
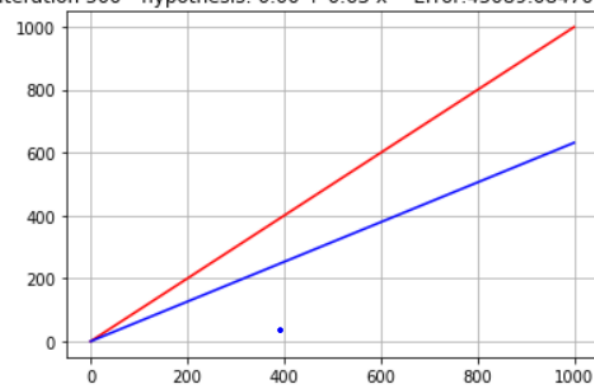
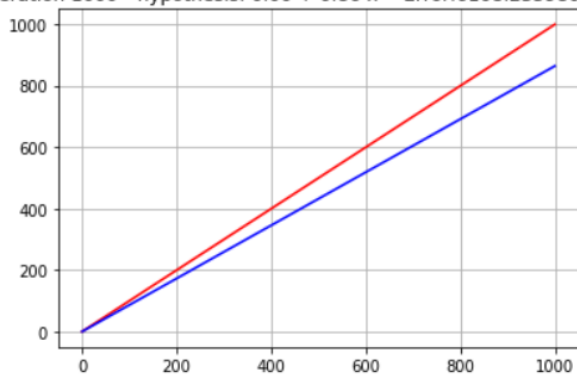Iteration 600-- hypothesis: 0.00 + 0.70 x -- Error:30230.149634636906

Iteration 700-- hypothesis: 0.00 + 0.75 x -- Error:20267.91967124011

4

Iteration 900-- hypothesis: 0.00 + 0.83 x -- Error:9110.598949992851

Iteration 1000-- hypothesis: 0.00 + 0.86 x -- Error:6108.235980006333

Iteration 1100-- hypothesis: 0.00 + 0.89 x -- Error:4095.2902210705947
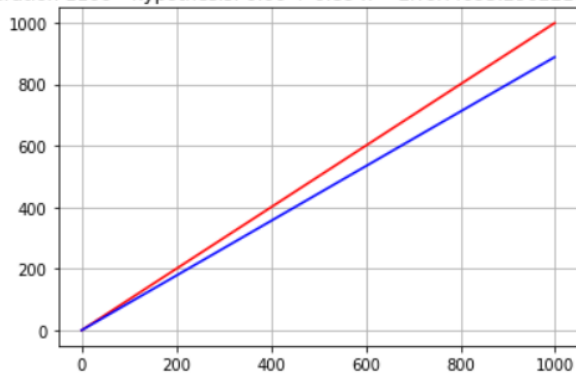
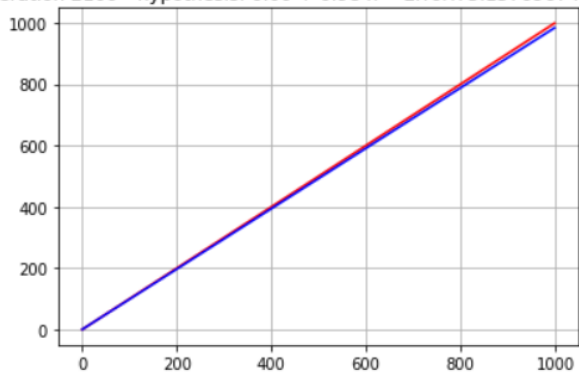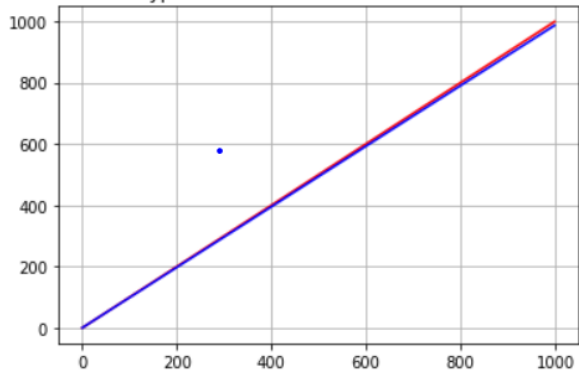Iteration 1200-- hypothesis: 0.00 + 0.91 x -- Error:2745.70302294598
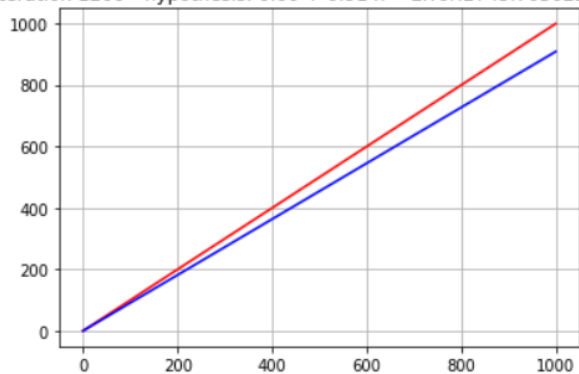
Iteration 2000-- hypothesis: 0.00 + 0.98 x -- Error:112.09973445673137
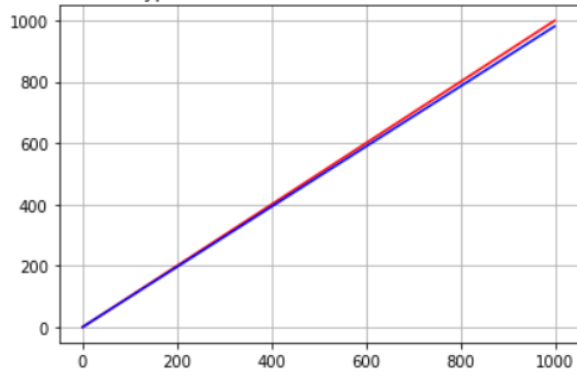
Iteration 2100-- hypothesis: 0.00 + 0.98 x -- Error:75.15769674292328

Iteration 2200-- hypothesis: 0.00 + 0.99 x -- Error:50.38976598777308

Iteration 2300-- hypothesis: 0.00 + 0.99 x -- Error:33.78401189688285
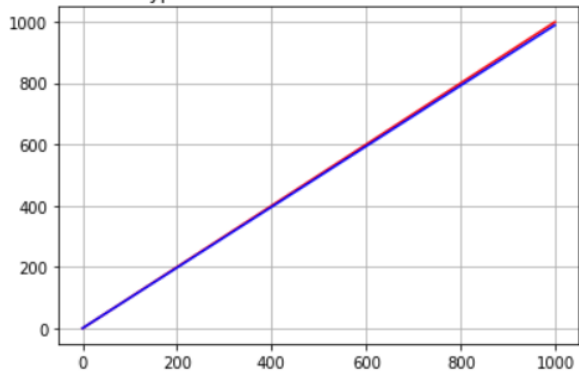
Iteration 3000-- hypothesis: 0.00 + 1.00 x -- Error:2.0572803384761995
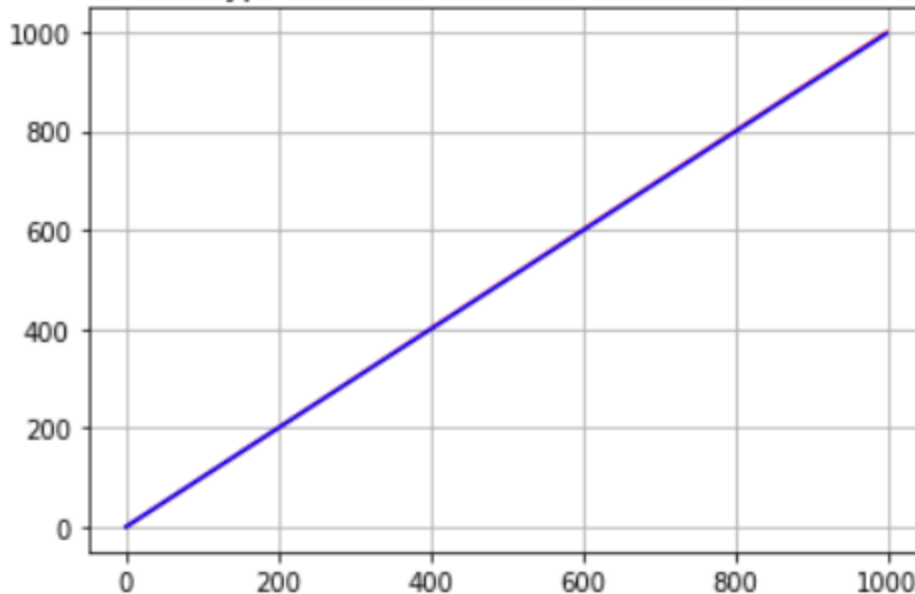
After 3000 it remains the same

Question 1.2

3. a2q12.py
   - Importing and creating an object of class linearRegression in RandomLine method
   - Variables are randomly generated using ramdom.uniform in python and seed(1)
   - X is randomly generated between [-4,6],500 examples
   - a is randomly generated between [-5,10]
   - b is randomly generated between [-4,5]
   - Y = a*X + b
   - Adding samples using addSample method – Adding features and labels
   - α = 0.000000003
   - numOfSteps = 100 (no of iterations of fit method per loop)
   - looped 50 times (no of times fit is called)
   - adding gaussian noise (0 is the mean of the normal distribution you are choosing from,1 is the standard deviation of the normal distribution ,500 is the number of elements you get in array noise)
   -
   - printing current hypothesis and cost function

Note: Chosen values

   1. alpha = 0.000003

2. numOfSteps = 1000
3. numOfLoops = 250

```
Current hypothesis:  0.00 + 0.00 x ,  cost =  757.4229
Current hypothesis:  0.07 + 0.49 x ,  cost =  674.1221
Current hypothesis:  0.14 + 0.96 x ,  cost =  600.0072
Current hypothesis:  0.20 + 1.40 x ,  cost =  534.0650
Current hypothesis:  0.26 + 1.82 x ,  cost =  475.3942
Current hypothesis:  0.32 + 2.21 x ,  cost =  423.1929
Current hypothesis:  0.37 + 2.58 x ,  cost =  376.7476
Current hypothesis:  0.42 + 2.93 x ,  cost =  335.4234
Current hypothesis:  0.47 + 3.25 x ,  cost =  298.6556
Current hypothesis:  0.52 + 3.56 x ,  cost =  265.9417
Current hypothesis:  0.56 + 3.86 x ,  cost =  236.8346
Current hypothesis:  0.61 + 4.13 x ,  cost =  210.9364
Current hypothesis:  0.65 + 4.39 x ,  cost =  187.8934
Current hypothesis:  0.68 + 4.64 x ,  cost =  167.3906
Current hypothesis:  0.72 + 4.87 x ,  cost =  149.1479
Current hypothesis:  0.75 + 5.09 x ,  cost =  132.9160
Current hypothesis:  0.79 + 5.29 x ,  cost =  118.4731
Current hypothesis:  0.82 + 5.49 x ,  cost =  105.6221
Current hypothesis:  0.85 + 5.67 x ,  cost =  94.1872
Current hypothesis:  0.87 + 5.84 x ,  cost =  84.0124
Current hypothesis:  0.90 + 6.00 x ,  cost =  74.9587
Current hypothesis:  0.93 + 6.16 x ,  cost =  66.9024
Current hypothesis:  0.95 + 6.30 x ,  cost =  59.7336
Current hypothesis:  0.97 + 6.44 x ,  cost =  53.3544
.
.
.

Current hypothesis:  1.98 + 8.62 x ,  cost =  1.0852
Current hypothesis:  1.98 + 8.62 x ,  cost =  1.0839
Current hypothesis:  1.98 + 8.62 x ,  cost =  1.0827
Current hypothesis:  1.98 + 8.62 x ,  cost =  1.0815
Current hypothesis:  1.99 + 8.62 x ,  cost =  1.0803
Current hypothesis:  1.99 + 8.62 x ,  cost =  1.0791
Current hypothesis:  1.99 + 8.62 x ,  cost =  1.0780
Current hypothesis:  1.99 + 8.62 x ,  cost =  1.0768
Current hypothesis:  1.99 + 8.62 x ,  cost =  1.0757
Current hypothesis:  2.00 + 8.62 x ,  cost =  1.0745
Current hypothesis:  2.00 + 8.62 x ,  cost =  1.0734
Current hypothesis:  2.00 + 8.62 x ,  cost =  1.0723
```

```
Current hypothesis:  2.00 + 8.62 x ,   cost =  1.0712
Current hypothesis:  2.00 + 8.62 x ,   cost =  1.0701
Current hypothesis:  2.00 + 8.62 x ,   cost =  1.0691
Current hypothesis:  2.01 + 8.62 x ,   cost =  1.0680
Current hypothesis:  2.01 + 8.62 x ,   cost =  1.0669
Current hypothesis:  2.01 + 8.62 x ,   cost =  1.0659
Current hypothesis:  2.01 + 8.62 x ,   cost =  1.0649
Current hypothesis:  2.01 + 8.62 x ,   cost =  1.0638
```
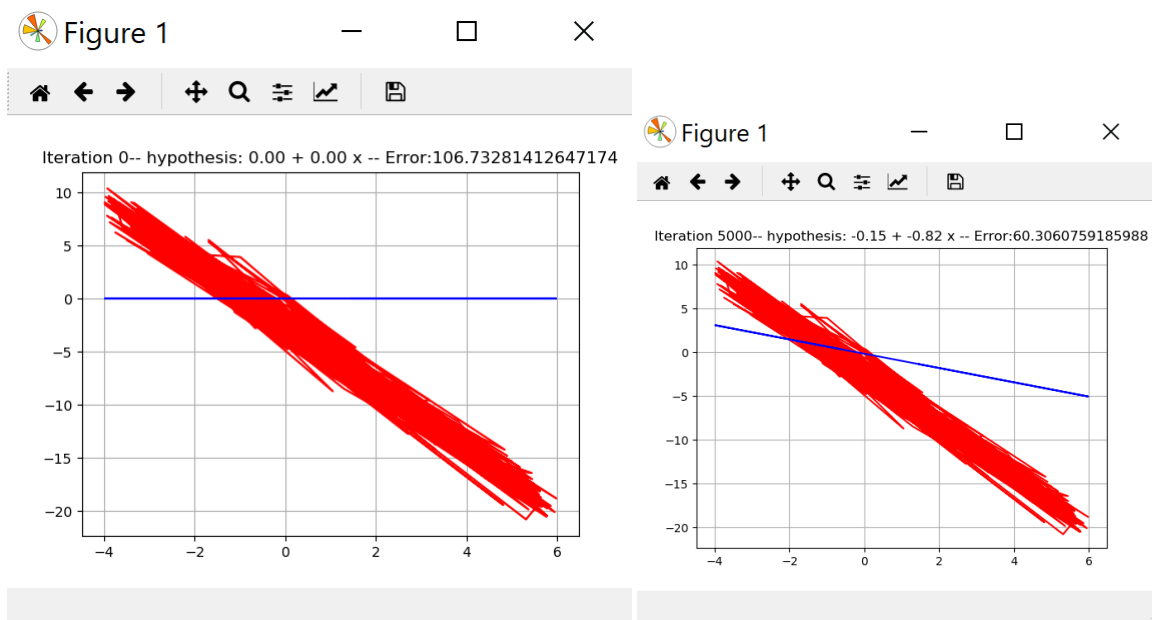
Bonus:

The blue line in the below graphs represents the red line in the assignment figures

1. alpha = 0.000003
2. numOfSteps = 1000
3. numOfLoops = 250
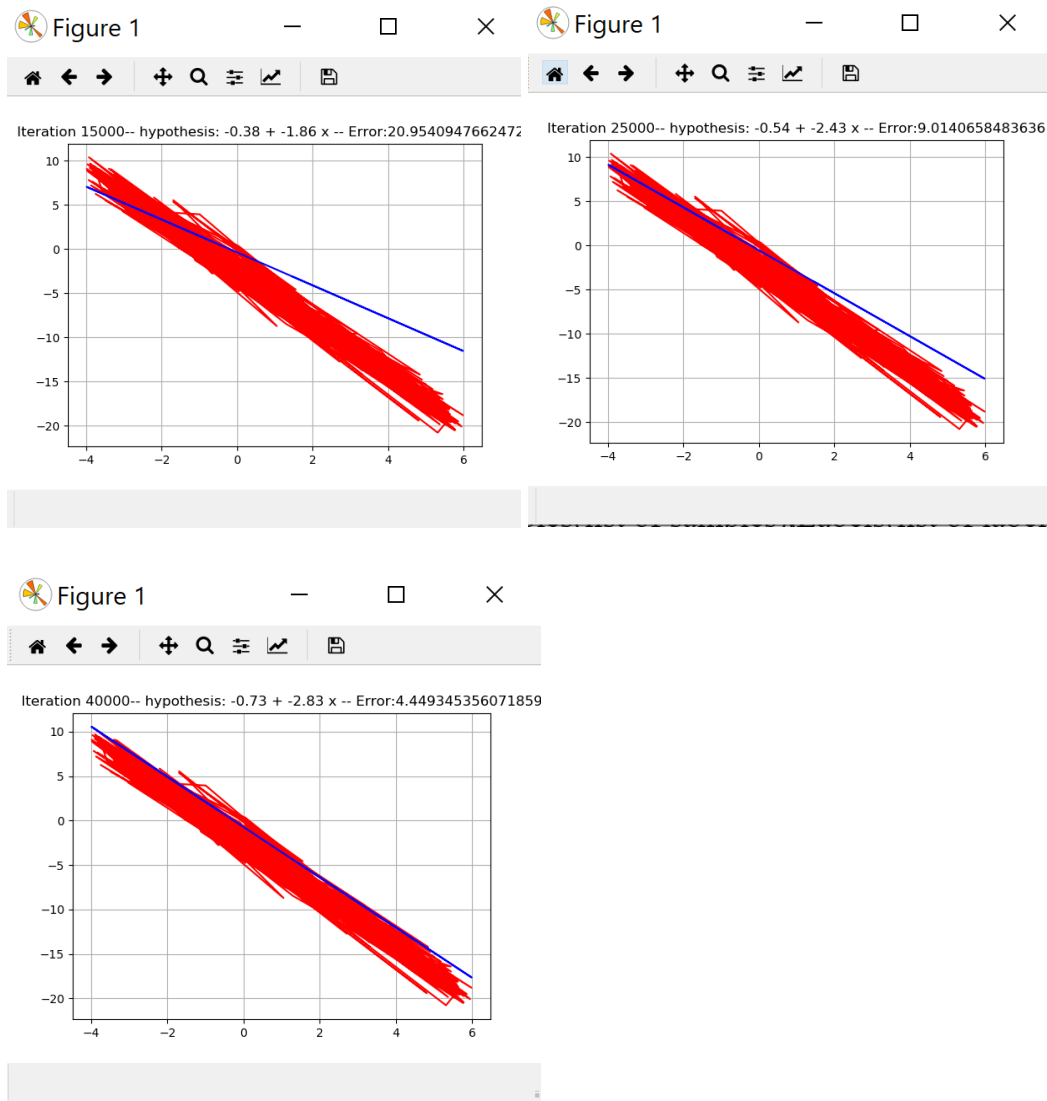
It remains the same after 40,000.

Question 2.1

1. Model.py
   - Implementing class linearRegression and methods as in the assignment
   - Global variables- Samples(list of samples),Labels(list of labels),theta(List of all thetas for D featues)
   - Implementing each method as mentioned in assignment and adding another method named summation(to find the summation with respect to theta value required for all examples)
   - Modifying for D features

2. a2q21.py
   - Importing and creating an object of class linearRegression in setPlane method
   - X = 0 to X = 999
   - z = x+2y
   - N = no of featues
   - Adding samples using addSample method – Adding features and labels
   - Adding "n" in setSamples method (here n = 2)
   - α = 0.000000003
   - numOfSteps = 1000 (no of iterations of fit method per loop)
   - looped 10 times (no of times fit is called)
   - printing current hypothesis and cost function
   - calculating points for points((x,2x),5x) and ((2x,x),4x) and storing in variables (x,y,z) respectively
   - array of features as (x1,x2)

```
importing Jupyter notebook from model.ipynb
Current hypothesis:  0.00 + 0.00 x_1 + 0.00 x_2 ,  cost =  6823086.7500
Current hypothesis:  0.00 + 1.18 x_1 + 1.81 x_2 ,  cost =  11390.9365
Current hypothesis:  0.00 + 1.07 x_1 + 1.93 x_2 ,  cost =  1538.5161
Current hypothesis:  0.00 + 1.02 x_1 + 1.98 x_2 ,  cost =  207.7996
Current hypothesis:  0.00 + 1.01 x_1 + 1.99 x_2 ,  cost =  28.0664
Current hypothesis:  0.00 + 1.00 x_1 + 2.00 x_2 ,  cost =  3.7908
Current hypothesis:  0.00 + 1.00 x_1 + 2.00 x_2 ,  cost =  0.5120
Current hypothesis:  0.00 + 1.00 x_1 + 2.00 x_2 ,  cost =  0.0692
Current hypothesis:  0.00 + 1.00 x_1 + 2.00 x_2 ,  cost =  0.0093
Current hypothesis:  0.00 + 1.00 x_1 + 2.00 x_2 ,  cost =  0.0013
Current hypothesis:  0.00 + 1.00 x_1 + 2.00 x_2 ,  cost =  0.0002
```

Question 2.2

3. a2q22.py
   - Importing and creating an object of class linearRegression in RandomPlane method
   - Variables are randomly generated using ramdom.uniform in python and seed(1)
   - X1 is randomly generated between [0,1],5000 examples (feature 1)
   - X2 is randomly generated between [0,1],5000 examples (feature 2)
   - a is randomly generated between [-100,100]
   - b is randomly generated between [-100,100]
   - c is randomly generated between [-100,100]
   - **noise d** is randomly generated between [-20,20]
   - y = a*x1+ b*x2+ c + δ
   - Adding "n" in setSamples method (here n = 2)
   - Adding samples using addSample method – Adding features and labels
   - α = 0.000000003
   - numOfSteps = 100 (no of iterations of fit method per loop)

- looped 50 times (no of times fit is called)
- printing current hypothesis and cost function
- array of features as (x1,x2)

Note: Have tried with all different values for numOfSteps (from 100 to 10000) , loops (from 10 - 2000) , α (from 0.000000003 to 0.00000000000001) but it was not converging. For some values it reduced till 140 and did not reduce.

4. a2q23.py
   - Importing and creating an object of class linearRegression in RandomPlane method
   - Variables are randomly generated using ramdom.uniform in python and seed(1)
   - D features = numOfFeatures – passed in the argument of RandomPlane method
   - 5000 examples
   - X is randomly generated between [0,1],5000 examples
   - Adding samples using addSample method – Adding features and labels
   - α = 0.000000003
   - numOfSteps = 100 (no of iterations of fit method per loop)
   - looped 50 times (no of times fit is called)
   - printing current hypothesis and cost function
   - randomly sample coefficients t0,t1,...,tn from [−100,100]
   - array of coefficient - t = []
   - array of all features of each 5000 examples - x = []
   - array of r - label - r = []
   - array of product of coefficient and x for each dimension - tx = []
   - array of features as (x1,x2...xn) - features = []
   - generating t1…tn and x1…xn in a loop for the given number of dimensions and multiplying corresponding t and x and storing in tx array later summing these values ofr each sample and adding t0 and noise to produce r as

$$r = t_0 + \sum_{j=1}^{D} t_j x^{(j)} = t_0 + t_1 x^{(1)} + t_2 x^{(2)} + \ldots + t_D x^{(D)}.$$

Note: Chosen values

1. alpha = 0.003
2. numOfSteps = 1000
3. numOfLoops = 100

```
Current hypothesis:  0.00 + 0.00 x_1 + 0.00 x_2 + 0.00 x_3 + 0.00 x_4 + 0.
00 x_5 ,  cost =  21663.9174
Current hypothesis:  -59.89 + -16.80 x_1 + -48.52 x_2 + -36.41 x_3 + -32.0
7 x_4 + -36.19 x_5 ,  cost =  122.3338
Current hypothesis:  -56.84 + -7.47 x_1 + -58.57 x_2 + -39.25 x_3 + -31.83
x_4 + -38.75 x_5 ,  cost =  50.4398
Current hypothesis:  -54.36 + -1.95 x_1 + -64.91 x_2 + -41.30 x_3 + -31.96
x_4 + -40.55 x_5 ,  cost =  21.8776
Current hypothesis:  -52.35 + 1.27 x_1 + -68.96 x_2 + -42.78 x_3 + -32.27
x_4 + -41.83 x_5 ,  cost =  10.1162
Current hypothesis:  -50.72 + 3.11 x_1 + -71.57 x_2 + -43.87 x_3 + -32.62
x_4 + -42.76 x_5 ,  cost =  5.0243
Current hypothesis:  -49.40 + 4.13 x_1 + -73.29 x_2 + -44.68 x_3 + -32.98
x_4 + -43.44 x_5 ,  cost =  2.6751
Current hypothesis:  -48.35 + 4.67 x_1 + -74.44 x_2 + -45.29 x_3 + -33.30
x_4 + -43.95 x_5 ,  cost =  1.5112
Current hypothesis:  -47.49 + 4.93 x_1 + -75.22 x_2 + -45.75 x_3 + -33.58
x_4 + -44.33 x_5 ,  cost =  0.8929
Current hypothesis:  -46.81 + 5.03 x_1 + -75.76 x_2 + -46.11 x_3 + -33.82
x_4 + -44.62 x_5 ,  cost =  0.5443
Current hypothesis:  -46.26 + 5.04 x_1 + -76.15 x_2 + -46.38 x_3 + -34.01
x_4 + -44.85 x_5 ,  cost =  0.3386
Current hypothesis:  -45.82 + 5.01 x_1 + -76.43 x_2 + -46.59 x_3 + -34.18
x_4 + -45.02 x_5 ,  cost =  0.2133
.
.
.
.
Current hypothesis:  -44.18 + 4.63 x_1 + -77.24 x_2 + -47.32 x_3 + -34.80
x_4 + -45.62 x_5 ,  cost =  0.0016
Current hypothesis:  -44.15 + 4.62 x_1 + -77.25 x_2 + -47.33 x_3 + -34.82
x_4 + -45.64 x_5 ,  cost =  0.0010
Current hypothesis:  -44.13 + 4.61 x_1 + -77.26 x_2 + -47.34 x_3 + -34.83
x_4 + -45.64 x_5 ,  cost =  0.0007
Current hypothesis:  -44.11 + 4.60 x_1 + -77.27 x_2 + -47.35 x_3 + -34.83
x_4 + -45.65 x_5 ,  cost =  0.0004
Current hypothesis:  -44.09 + 4.60 x_1 + -77.28 x_2 + -47.36 x_3 + -34.84
x_4 + -45.66 x_5 ,  cost =  0.0003
Current hypothesis:  -44.08 + 4.59 x_1 + -77.28 x_2 + -47.37 x_3 + -34.85
x_4 + -45.66 x_5 ,  cost =  0.0002
Current hypothesis:  -44.07 + 4.59 x_1 + -77.29 x_2 + -47.37 x_3 + -34.85
x_4 + -45.66 x_5 ,  cost =  0.0001
Current hypothesis:  -44.06 + 4.59 x_1 + -77.29 x_2 + -47.37 x_3 + -34.85
x_4 + -45.67 x_5 ,  cost =  0.0001
```