Ishan Patel - 116169103
Nakul Patel - 116334877

Homework 1

ENPM 673
February 14, 2019

1 Problem 1

1.1 Eigenvalues and Eigenvectors

Given any matrix or linear tranformation A, there exists a vector, which when multiplied with this linear tranformation, does not change the direction of the vector, but simply scales it(magnifies or shortens the vector).

$$A * v = \lambda * v$$

In the above equation, v is the vector on which linear tranformation A is applied, which after multiplication, scales by the factor of lamda. So v is called the eigenvector , that points in the direction of the stretching done by the tranformation, and λ is called the eigenvalue.

One can determine the eigenvalues and eigenvectors of the matrix by solving this characteristic equation.

$$det(A - \lambda * I) = 0$$

- From given dataset in 2D, we separated the X and Y coordinates , and stored them in an array, which can be later used for finding the mean, standard deviation, variance and covariance.
- Variance is simply the square of standard deviation, and it indicates the spread of data. And, covariance is the measure of joint variability of any two random variables, and it indicates orientation of our data.
- Any data can be represented in form of covariance matrix, and it indicates the shape of data. In order to determine the spread and orientation of data along the axes, we need to compute eigen vectors and eigen values of the covariance matrix.
- Now, for each dataset, once we determine the the covariance matrix with the following formulae, we can apply the concept of eigenvalues and eigenvectors on this matrix. Since, it is 2*2 matrix , we will get two eigenvalues and two eigenvectors.

1.2 Geometric Interpretation of eigenvectors and eigenvalues:

From the figure, we can say that:

1. The largest eigenvector represents the direction of the largest spread of data, and its magnitude is equal to the corresponding eigenvalue.
2. And other vector, the second eigenvector will always be perpendicular(orthogonal) to the largest eigenvector, and represents the direction of second largest spread of data.

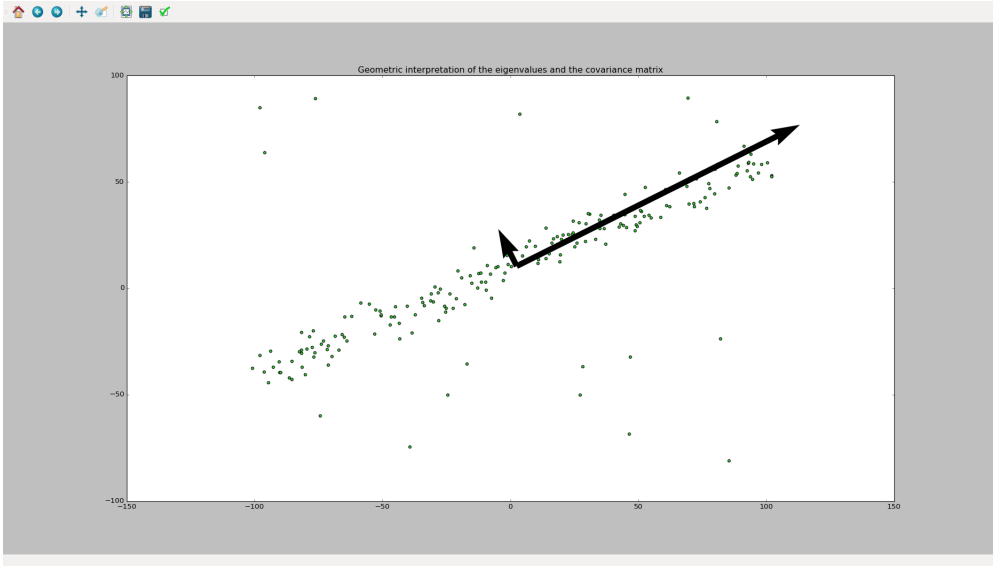

Figure 1: Visualization of covariance matrix for data-set.

2 Problem 2

2.1 Ordinary Least Squares

It is the most widely used method for least square technique to determine the model that best fits the data. In general, we have more equations than the variables, and it is not possible to find the solution by solving the system of linear equations.

In it, we determine the model(line) based on minimizing the squares of errors(vertical distance of the point to the predicted line). In other words, we minimize the sum of squares of the difference between the observed data point, and the one which is predicted by the linear model.

In general, for linear relation, we write the equation of our linear regression model as follows:

$$y_i = \alpha + \beta * x_i + \epsilon_i$$

, where, x_i and y_i are the values of the datapoints, and ϵ is the error of the observation point.

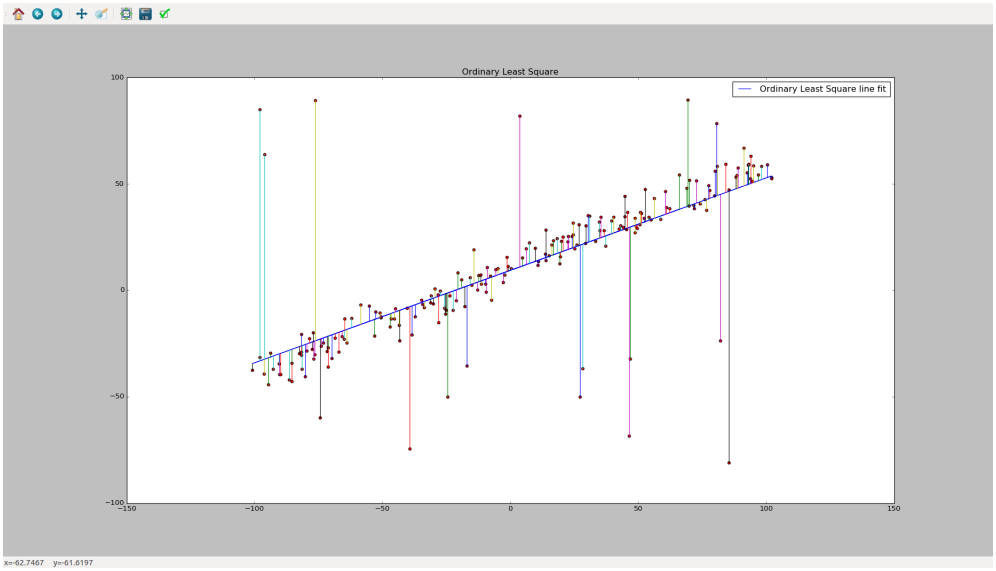

Figure 2: Ordinary Least Squares linear fit

2.2 Total Least Squares:

It is another method for least square estimation. The only difference is that, we minimize the sum of squares of the perpendicular(orthogonal) distances to the predicted line.

In other words, it accounts for the errors in both the x-axis and y-axis directions of the given data-points.

Implementation:

Based on the above theory, the equations for the predicted model parameters are written and code implements the same. The graph depicting the predicted line is plotted for each of the given datasets. We observe that the outliers do affect the prediction of our model.

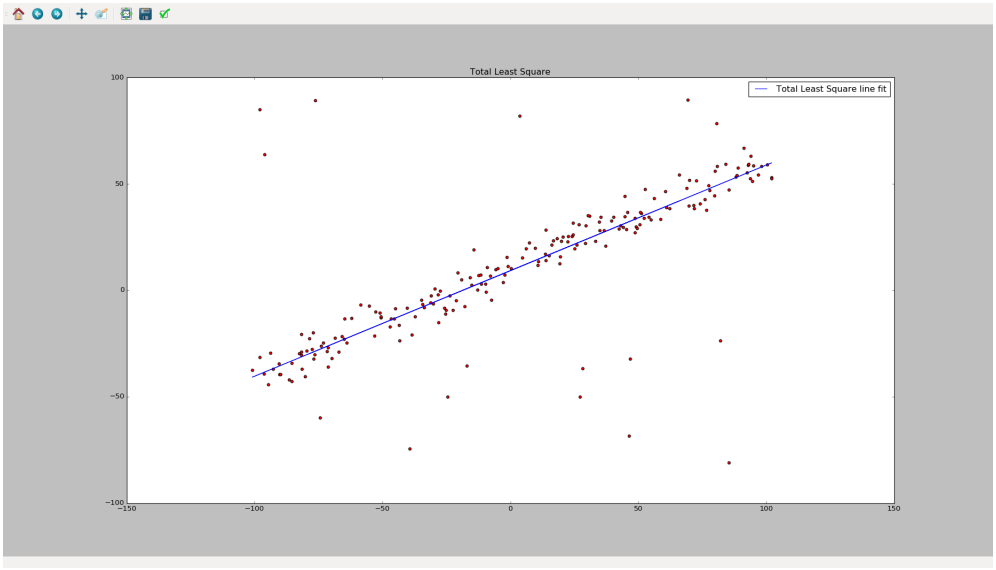

Figure 3: Total Least Square linear fit

2.3 Ordinary Least Square vs Total Least Square

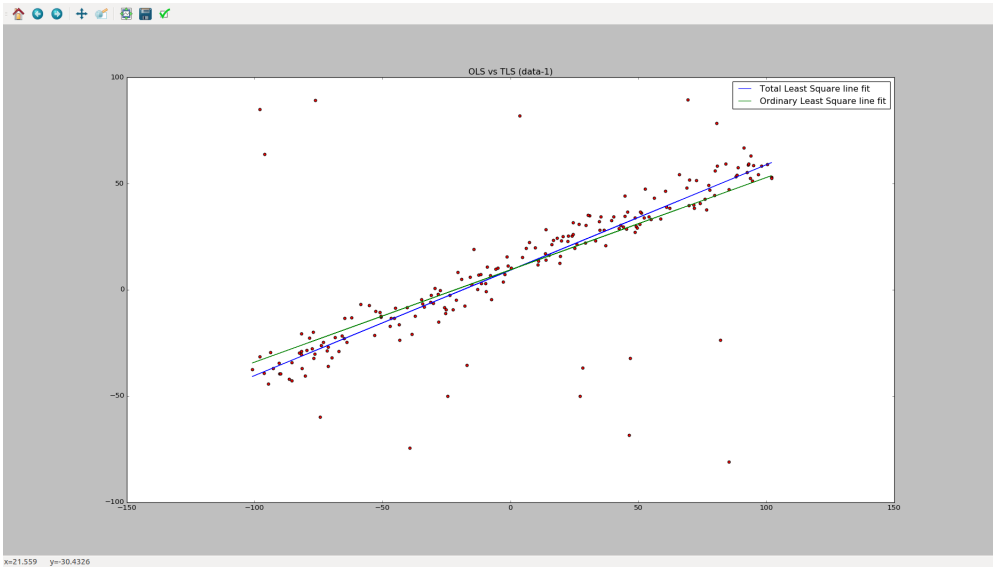

Figure 4: OLS vs TLS for data-1

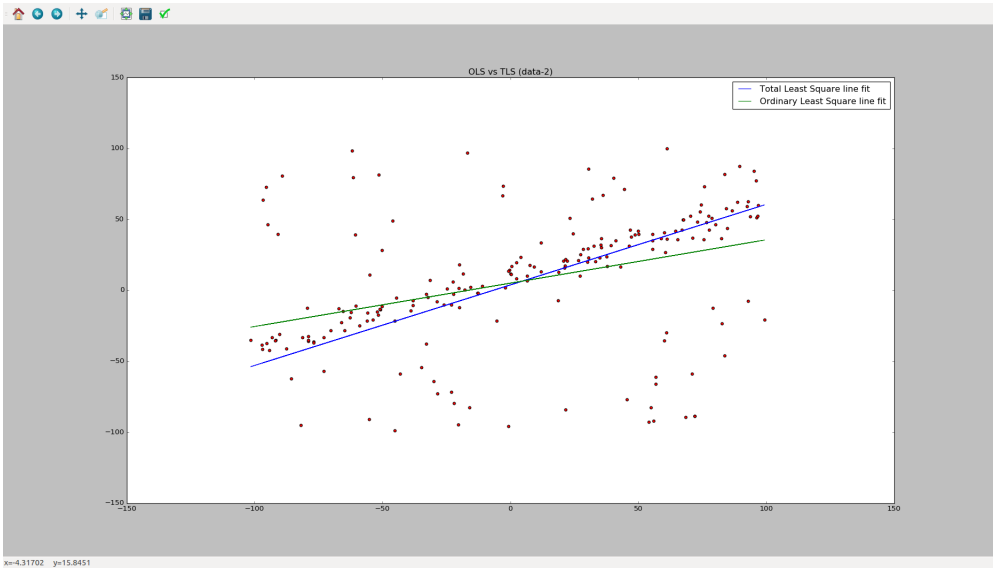

Figure 5: OLS vs TLS for data-2

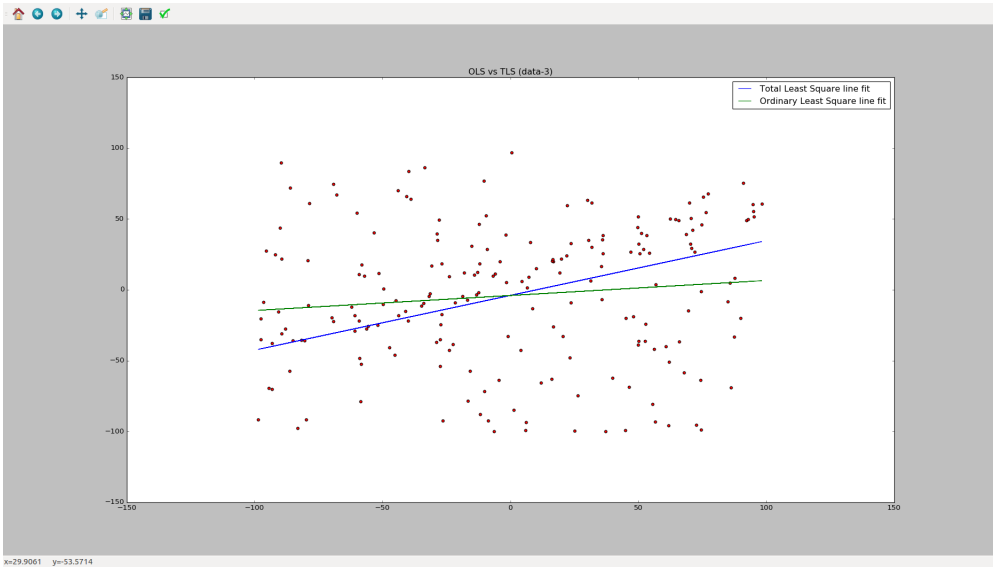

Figure 6: OLS vs TLS for data-3

3 Problem 3

In any given datasets, there is always some noise present, which can effect the estimation of the best linear model that describes the data. When working with least squares, there are some datapoints called outliers which influence the estimation of the model. These outliers are generally far away from the predicted line, and hence they increase the error term, with adversely effect on our minimization of least squares.

Implementation for RANSAC:

The most widely used technique for outlier rejection technique is RANSAC(Random Sample Consensus), the working of which is explained below.

The algorithm is implemented as follows:

1. Firstly, a sample dataset containing minimum random points(2 in our case) are selected and a fitting model(line) parameters are calculated.
2. Then the test data points (points other than randomly selected) are used to calculate their distance from the line, and decide whether they are inliers based on the predefined threshold.
3. Update the model-parameters if the number of inliers are greater than the previous one.
4. Return the best model parameters after running steps 1 to 3 for some N iterations.

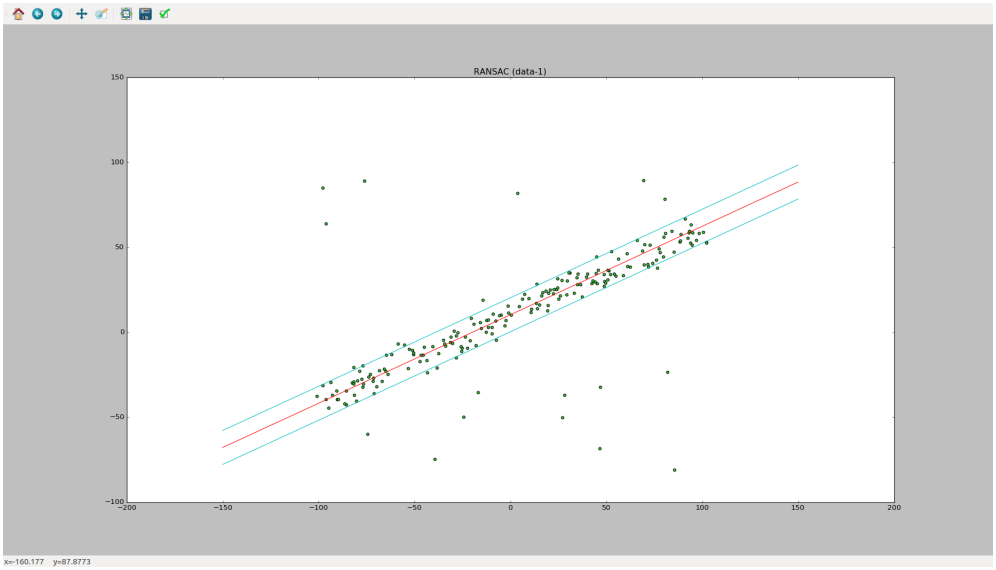

Figure 7: RANSAC for Data-1

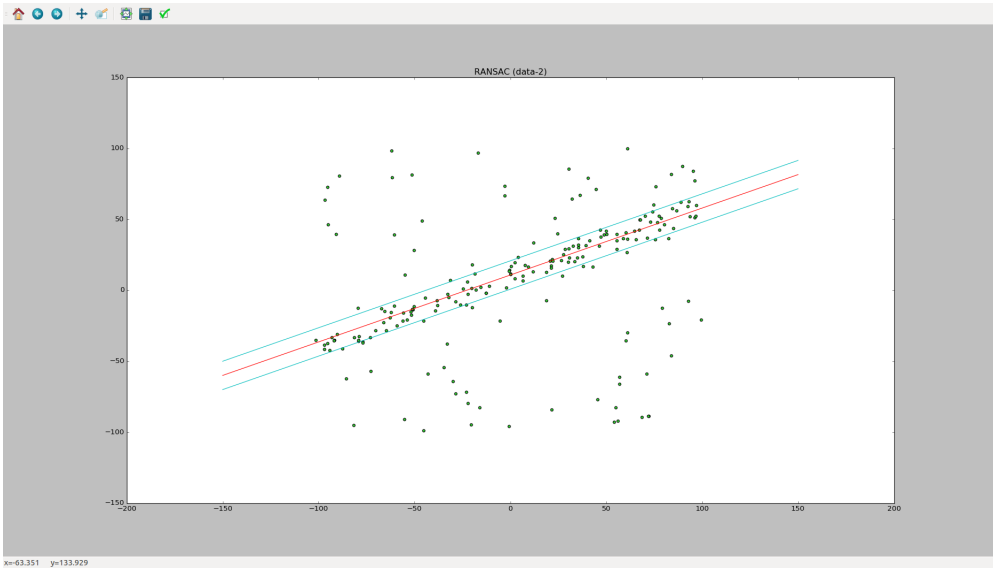

Figure 8: RANSAC for Data-2

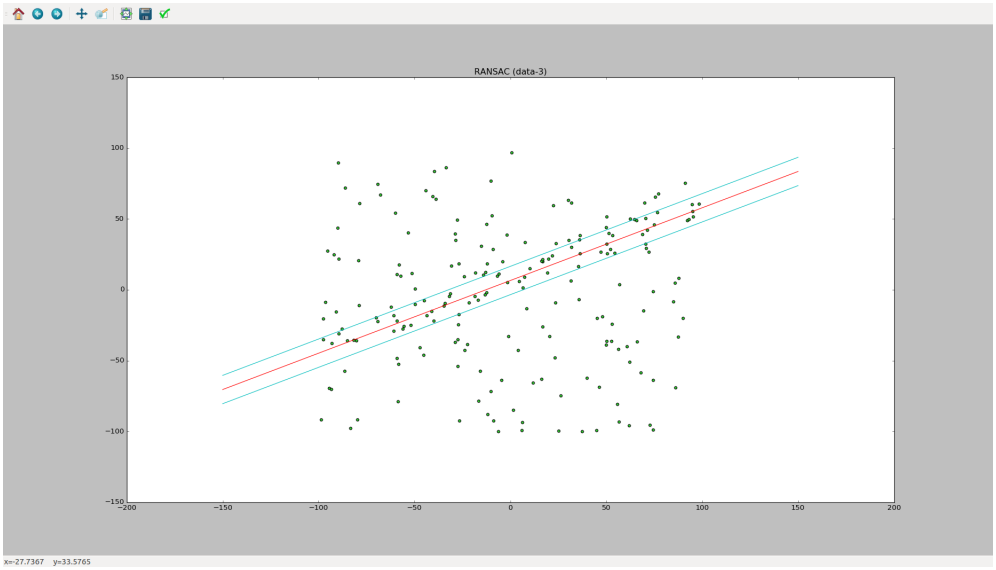

Figure 9: RANSAC for Data-3

For dataset1(data1 from pickle file):

For this data, most of the points are concentrated around the mean of the data in each direction, hence due to very few outliers, the RANSAC method will give the best results. In RANSAC, for a particular iteration, we will get most of the inliers within the threshold, and hence we can get the best fit of the line.

For dataset2(data2 from pickle file):

For this dataset, we can see that there is large deviation in the predicted model with RANSAC and other methods, of which RANSAC has more better result. Also, more than 50 percent of the inliers are included within the threshold distance for the best fit model, hence we implemented RANSAC for this dataset.

For dataset3(data3 from pickle file):

For this dataset, from the outputs of different methods, we can see that RANSAC gives the best result, hence RANSAC has been implemented for this dataset.

Implementation for Least Squares with Regularization:

In this method, extra constraint is added on the solution. With the additional term added to the closed-form solution, this method accepts little bias to reduce variance and the mean-squared error, thereby improving accuracy.

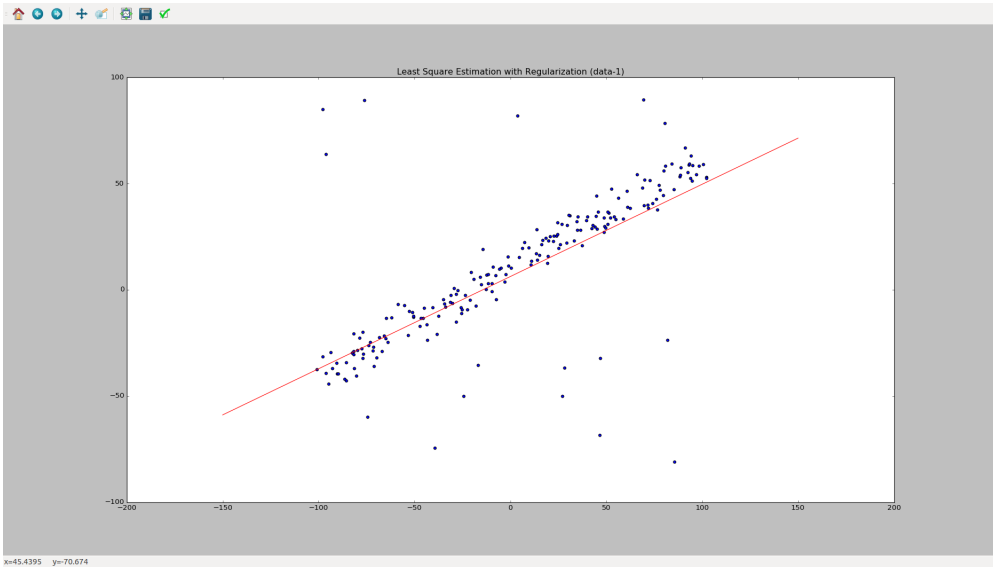

Figure 10: Least Square Regularization for Data-1

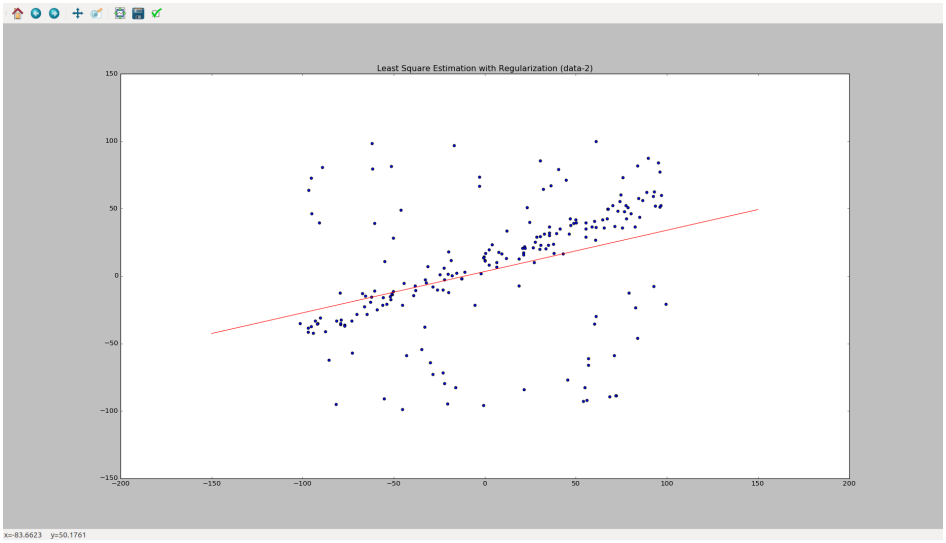

Figure 11: Least Square Regularization for Data-2

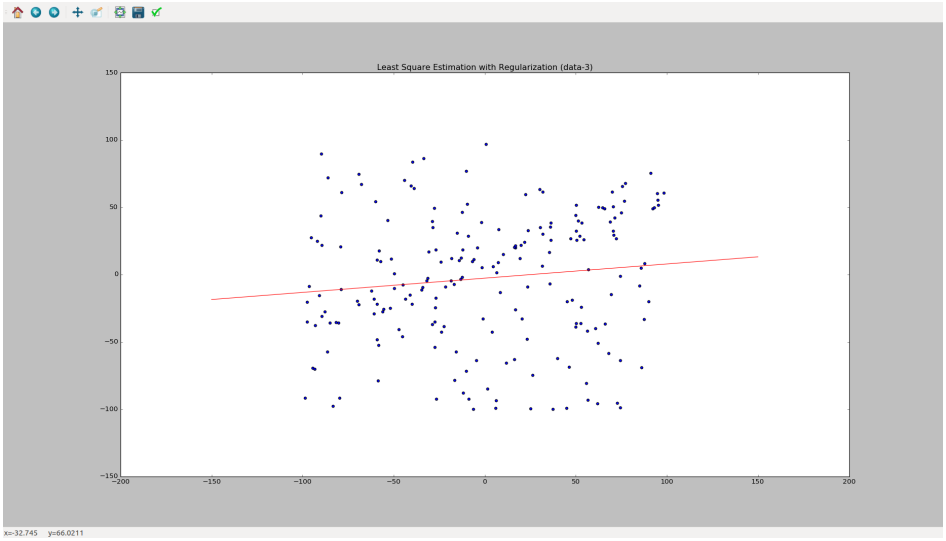

Figure 12: Least Square Regularization for Data-3

3.1 Limitations of RANSAC:

1. The number of random samples required for estimating the best fit, is generally kept high, which increase the computational efficiency of the algorithm and it requires many parameters to be tuned.

2. Also, when the inlier ratio is too low, the RANSAC is unlikely to find a good solution, since it does not test enough hypotheses data.