Question 1 – linear regression

I have chosen to use Maximum Likelihood solution much like as I did in the second assignment of the course because I already have a good implementation of it and I understood the algorithm very well so it was easy to make the implementation. I have chosen to make the implantation in Matlab and I used mostly the functions that we have written for that assignment.

I solved in question by going through the following steps:

* At first, I read the data from both files and did not normalize because it has no effect on this linear model.
* I created the design matrix corresponding to variable Φ (according to Eq. 3.16 from Bishop book) which is needed in the algorithm for building the model by inserting a column of 1 corresponding to the bias parameter and adding the 10 columns corresponding to the first 10 parameters in the train dataset.
* I calculated the model by applying the Eq. 3.15(from Bishop book): where is the design matrix and is the matrix with column 11 corresponding to target values of redshift from the training data. This calculation is done in the function *my\_wML* I obtained the following model:
* By using the model obtained previously, I applied the test data to the linear model by using the function *my\_lmfunction* which takes as input the first 10 columns of test data and the model to make the matrix multiplication. The result is a column matrix containing the prediction for the redshift of each galaxy in test.
* By using the function *my\_rms* which was implemented for the assignment the mean-squared-root error is calculated by computing the equation:

The resulting mean-squared-error for train data is .

The resulting mean-squared-error for test data is .

In order to also have a visual estimation on how the model behaves, I have also created a plot as in assignment 2 to visualize the difference between the redshift that was calculated for the test set and the prediction obtained by using the linear model. I have only plotted the values for galaxies 100-200 because it was hard to visualize if I was plotting for all the 2500 galaxies.

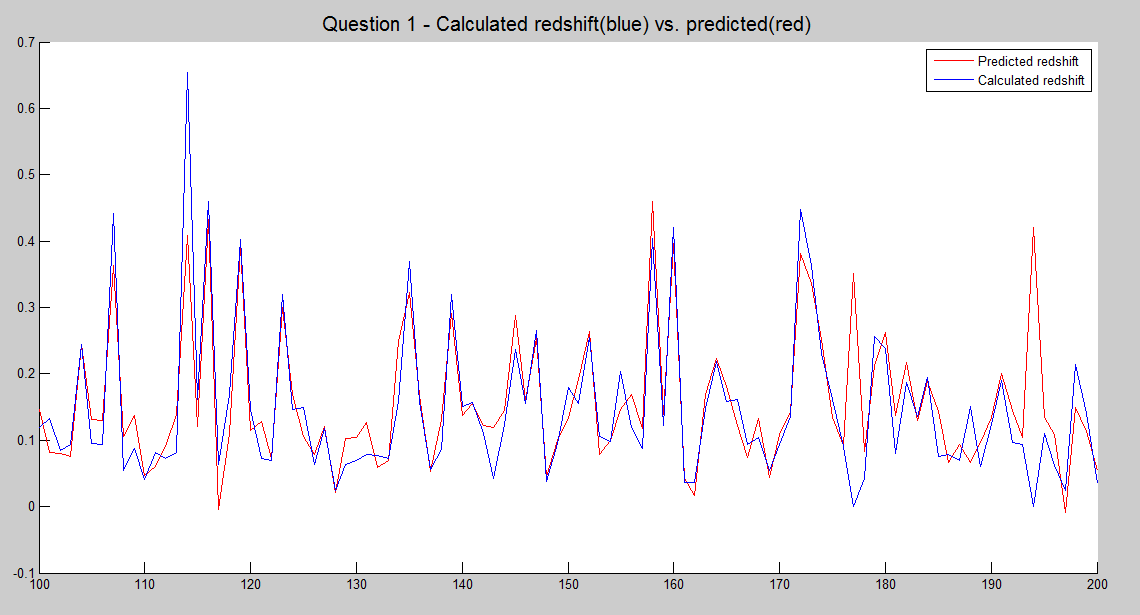


Figure 1. The plotting of the redshift from the test set and the redshift predicted by the linear model that I have obtained for the test data

Both the RMS and the plot show that the model obtained using the linear regression are quite good and there is no overfitting. Because the train data offers good generalization and the large dataset for training helped to create a better model and obtain better precision than in assignment where we only had 100 patterns.