Learning Systems

DA8008 (HT12) Halmstad University

Regression Model for cooling in a H2O2 process

Xu Fei, Qiu Yinan

1. Introduction

This report is focus on constructing a model for cooling a H2O2 process. This is a non-linear regression task. The process that generated the data is EKA chemicals Hydrogen Peroxide production.

The output is the valve opening of a valve connected to one (out of two) heat exchangers. This opening is part of a feed-back control loop based on the temperature of the fluid that passes thru the heat exchanger. The goal with the control is to keep the temperature constant and the fluid in liquid format. In cases where a large amount of cooling is required (large opening of the valve) its indicating that the fluid is in transition to gas form and this information is of particular significance. The goal of the modeling is to construct a model with all or part of the variables available to model the valve opening.

In this project, I used linear regression, stepwise regression, PCA, PLS and MLP to construct the model, and at last, I make a committee machine to improve the result. Among all the models, the linear regression combine with forward selection and backward elimination shows the best performance.

2. Methodology

In this section, method implemented for model construct is presented, including data covariance and correlation estimation, linear regression, stepwise regression, cross-validation ,PCA (Principal Component Analysis), PLS(Partial Least-Squares) and MLP (Multi-Layer Perceptron) learning using move the outliers

2.1 Covariance and Correlation Estimation

Covariance between input and output is calculated to show how much the input feature effect the output. Correlation coefficient is the normalization of covariance matrix, the most importance input feature is the feature has biggest correlation coefficient with the corresponding output.

2.2 Linear Regression and Stepwise regression

Linear regression model is constructed by the least-square method, and use all the features, in Matlab, I add the robust function for the linear regression, and I use Cook's Distance to find the outliers and move them out. Stepwise regression is a regression that include the forward selection and backward elimination, this method combine them together and do the bidirectional estimation, but this function don't have the robust function.

In this model, I combine the stepwise regression and linear regression together, use the stepwise regression to choose the features, and use the linear regression to get the result.

Robust regression works by assigning a weight to each data point. Weighting is done automatically and iteratively using a process called iteratively reweighted least squares. In the first iteration, each point is assigned equal weight and model coefficients are estimated using ordinary least squares. At subsequent iterations, weights are recomputed so that points farther from model predictions in the previous iteration are given lower weight. Model coefficients are then recomputed using weighted least squares. The process continues until the values of the coefficient estimates converge within a specified tolerance.

2.3 Cross-Validation

This means using a test data set, which is a subset of the available data (typically 25-35%) that is removed before any training is done, and which is not used again until all training is done. The performance on this test data will be an unbiased estimate of the generalization error, provided that the data has not been used in any way during the modeling process. If it has been used, e.g. for model validation when selecting hyper parameter values, then it will be a biased estimate.

2.4 PCA (Principal Component Analysis)

Principal component analysis (PCA) is a mathematical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. The number of principal components is less than or equal to the number of original variables. This transformation is defined in such a way that the first principal component has the largest possible variance (that is, accounts for as much of the variability in the data as possible), and each succeeding component in

turn has the highest variance possible under the constraint that it be orthogonal to (i.e., uncorrelated with) the preceding components. Principal components are guaranteed to be independent only if the data set is jointly normally distributed. PCA is sensitive to the relative scaling of the original variables.

2.5 PLS (Partial Least-Squares)

Partial least squares regression (PLS regression) is a statistical method that bears some relation to principal components regression; instead of finding hyper planes of minimum variance between the response and independent variables, it finds a linear regression model by projecting the predicted variables and the observable variables to a new space. Because both the X and Y data are projected to new spaces, the PLS family of methods are known as bilinear factor models. Partial least squares Discriminant Analysis (PLS-DA) is a variant used when the Y is binary.

2.6 MLP (Multi-Layer Perceptron)

A multilayer perceptron (MLP) is a feed forward artificial neural network model that maps sets of input data onto a set of appropriate outputs. An MLP consists of multiple layers of nodes in a directed graph, with each layer fully connected to the next one. Except for the input nodes, each node is a neuron (or processing element) with a nonlinear activation function. MLP utilizes a supervised learning technique called back propagation for training the network. MLP is a modification of the standard linear perceptron and can distinguish data that are not linearly separable.

3. Data

In this regression problem, data consists 3 matrix: XtrainDS (4466x65), YtrainDS (4466x1) and XtestDS (2971x65). All input variable of the process is contained in XtrainDS and output matrix YtrainDS present the output value, which is the valve opening.

First, I use the covariance between input and output to rank the affection from input to output, here I plot out top4 input features that effect the output. For top4 features: 10, 9, 65, 37, Fig 3.1 shows the result:

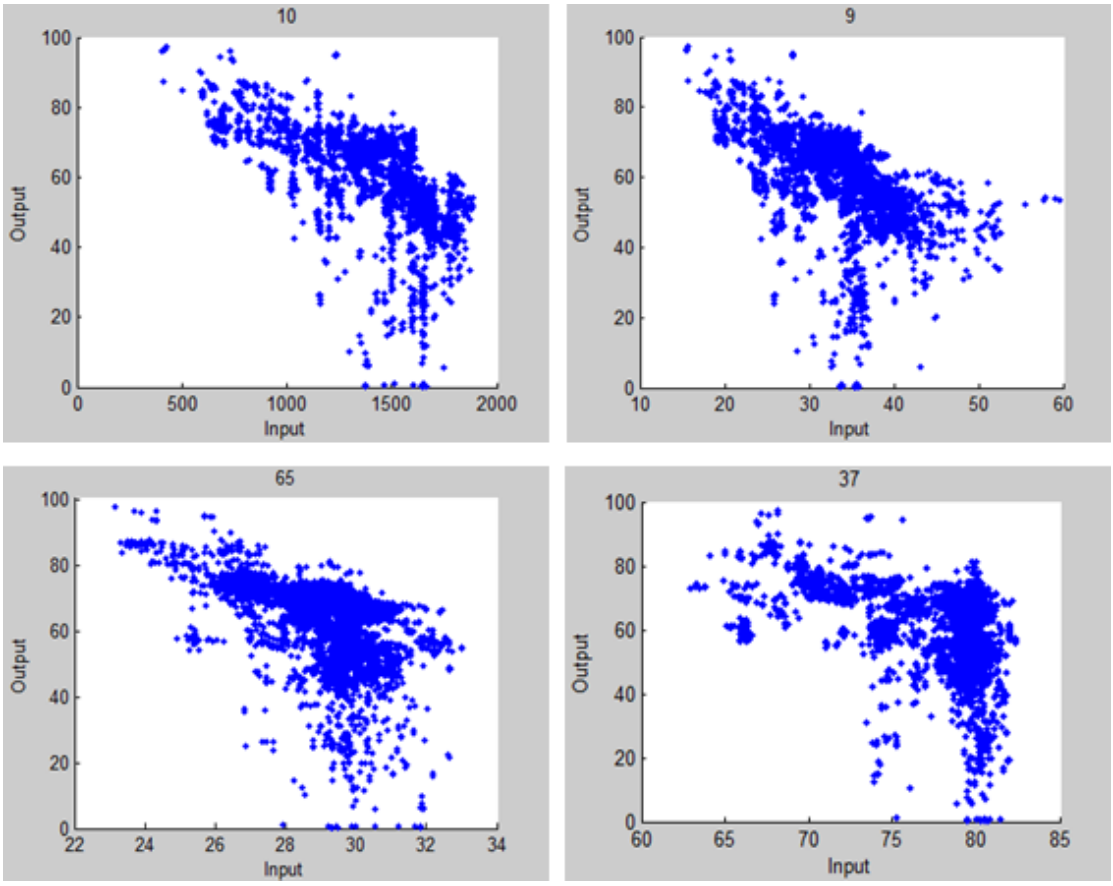

Fig 3.1 Input Features 10, 9, 65, 37 with respect of Y

From the result, we can see the relation between input and output have similar distribution, but there are a lot outliers in the data, we need to move the outliers and get the bet set of the features to construct the model for regression.

Use stepwise regression to get the appropriate set of the input feature, and use these features to construct the model, the features that we use for linear regression is:1,2,3,5,6,7,9,10,11,12,14,19,20,27,28,29,33,34,36,37,40,41,42,44,46,47,50,51,54,56,57,58,61,62.

After we have constructed the linear regression model, we need to move the outliers that got by Cook's Distance, Fig 3.2 show the Cook's Distance that we calculate out, the left fig is before we move the outliers, the right fig is after we move the outliers.

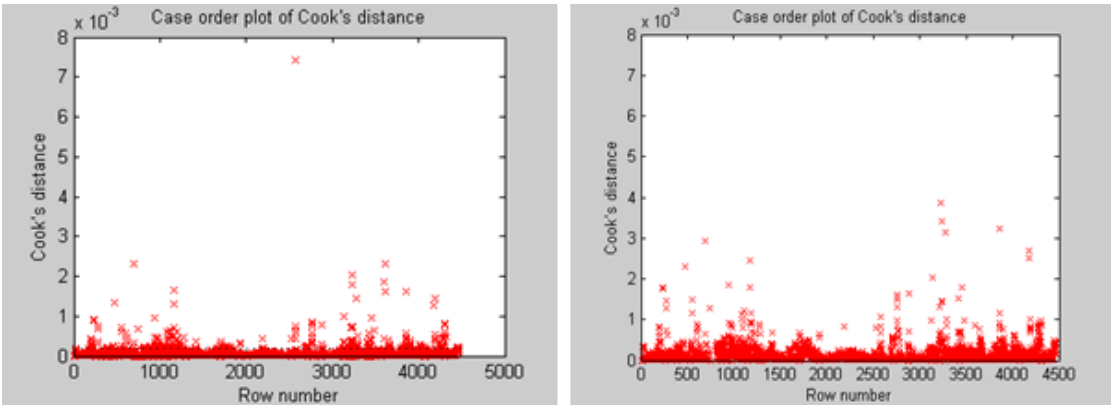

Fig 3.2 Case order plot of Cook's distance

The next step is too process the residual, we regard the residual less than -10 and bigger than 10 as the outliers too, Fig 3.3 shows the residual case.

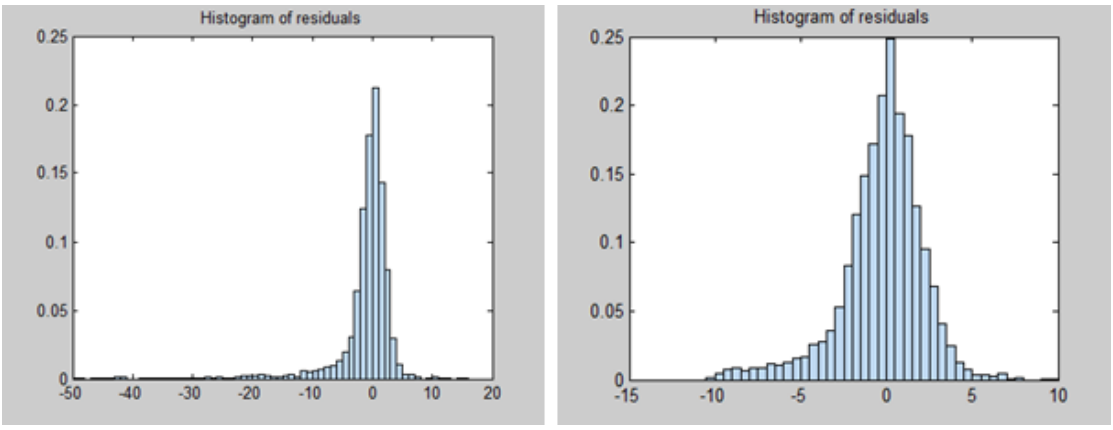

Fig 3.3 Histogram of residuals

We use PCA to analysis the contribution of each input feature, and use PLS to do the regression with the top10 contribution features. Fig 3.4 show the PCA result.

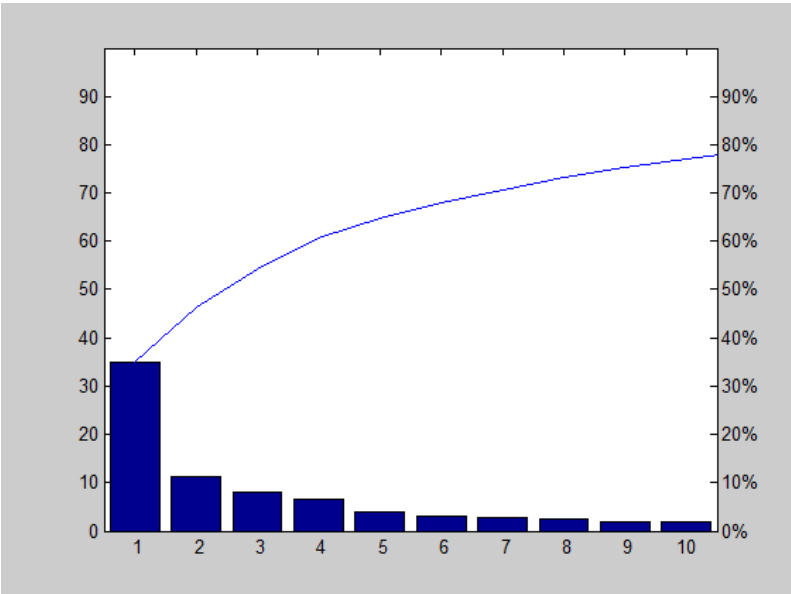

Fig 3.4 PCA

In the PLS regression, we use PLS to calculate the relation between number of

PLS component and the percent variance explained in Y, Fig 3.5 present the relation between all the components and the components with top10 contribution that we calculate by PCA between with variance explained in Y.

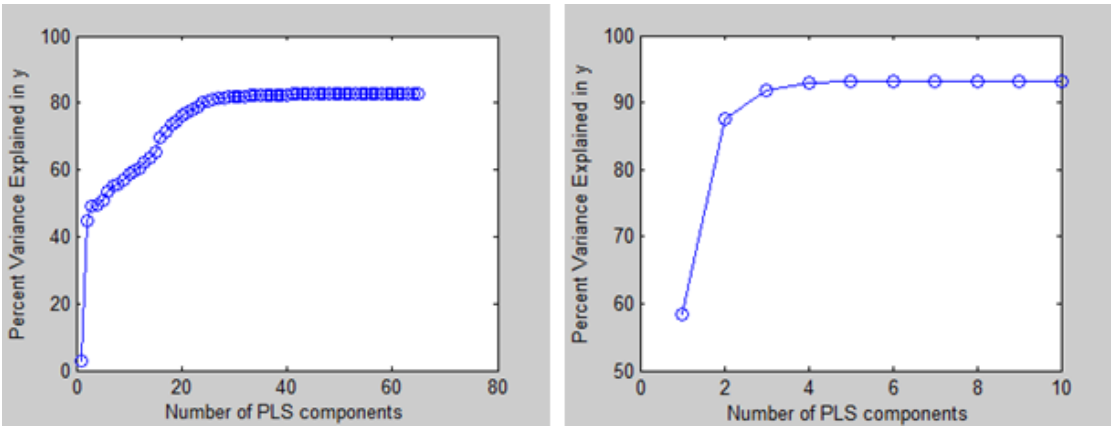

Fig 3.5 Relation between PLS components and percent variance in Y

4. Result

In this section, result of models including Linear-Stepwise Regression, PLS regression, MLP regression is presented. Cross-validation is applied on each technique, and we the K-fold method for cross-validation, and K is 10.

4.1 Linear-Stepwise Regression Model

In this model, we combine the linear regression with stepwise regression, and linear regression is based on least-square method, we use the stepwise to do the forward selection, backward elimination and bidirectional estimation, the features set that get from stepwise regression is used as input. For linear-stepwise regression, 10-fold cross-validation to estimate the generalize ability of the model. **The MSE of this model is MSE= 5.7227.**

The result that we fit the XtrainDS (input data) is presented in Fig 4.1.

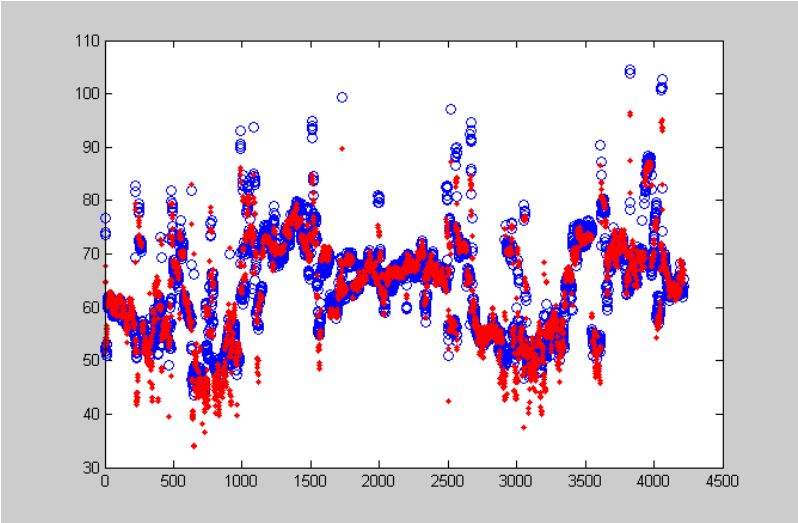

Fig 4.1 Fit plot of the Linear-Stepwise regression

4.2 PLS Regression Model

In this section, I construct two models used PLS regressions to do compare. The first one is with all the component, and the second one is with top10 contribution component that we get from PCA. **The MSE are MSE1= 28.2331 and MSE2= 85.0379 for the two models correspondingly.**

Fig 4.2 show the relation between components number with Predictors MSE and Response MSE. And from the result we can see with all the components, 30 of them are enough to do the PLS regression.

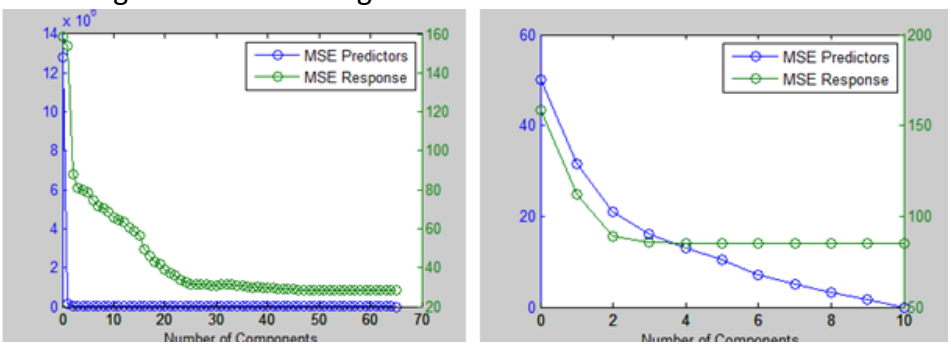

Fig 4.2 Relation between components and MSE

Fig 4.3 show the fit plot that we use PLS model to regress the XtrainDS (input data), the left one the result of model that with all components, the right one is the result of model that with 10 components.

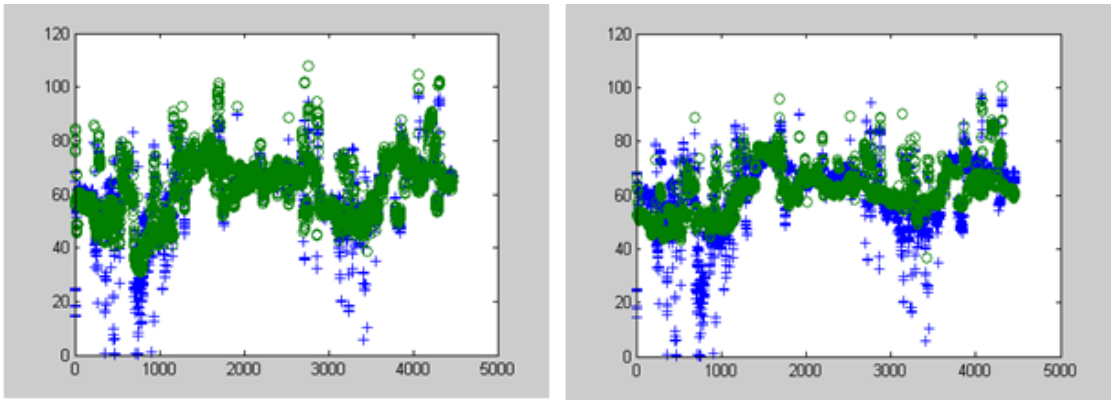

Fig 4.3 Fit plot of the PLS regression

4.3 MLP Regression Model

For multilayer perception, different construct is modeled by different input features. With variables set1 [3, 6, 16, 49, 50, 51, 52, 53, 65], 4 to 10 hidden nodes are tested, result is shown in Fig 4.4 is number of hidden layer with corresponding to validation error. **Nodes number of 4 has the lowest validation error and MSE= 19.3470.** Another set2 [4, 10, 19, 27, 28, 38, 42, 50], 4 to 10 hidden nodes are tested, result is shown in Fig 4.5 is number of hidden layer with corresponding to validation error. **Nodes number of 7 has the lowest validation error and MSE= 28.2729.** The final model that we choose is with the set we get from the stepwise regression, we use 34 features, 4 to 10 hidden nodes are tested, result is shown in Fig 4.6 is number of hidden layer with corresponding to validation error. **Nodes number of 8 has the lowest validation error and MSE= 7.6668.**

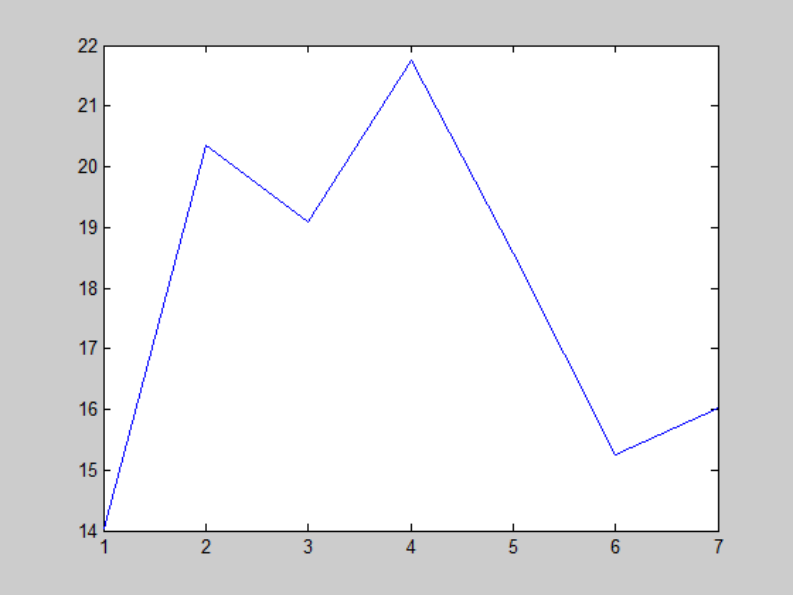

Figure 4.4 MLP node number test for variable set 1

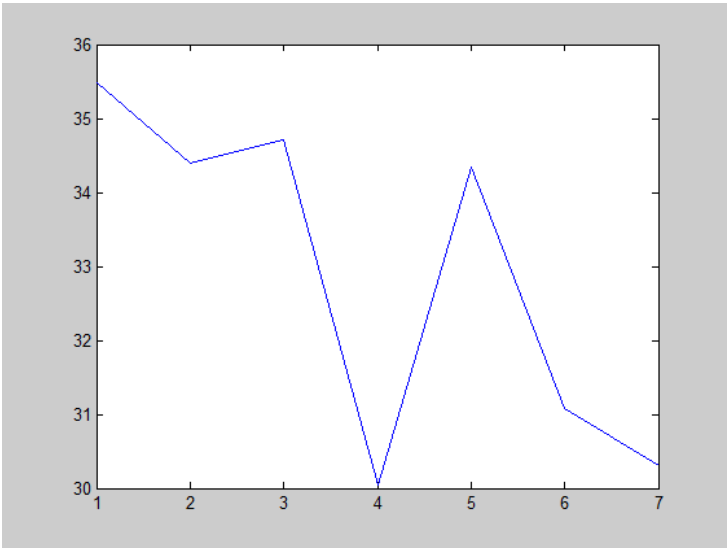

Figure 4.5 MLP node number test for variable set 2

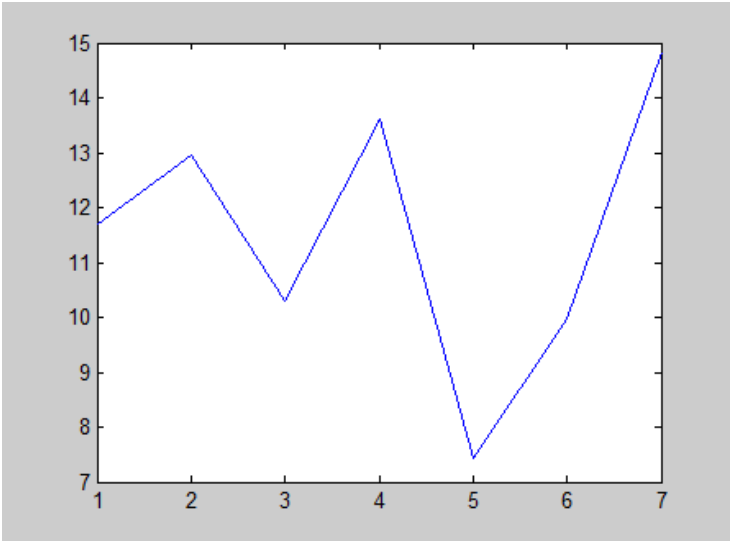

Figure 4.6 MLP node number test for variable final set

Fig 4.7 shows the performance of the MLP. Fig 4.8 shows the regression of MLP.

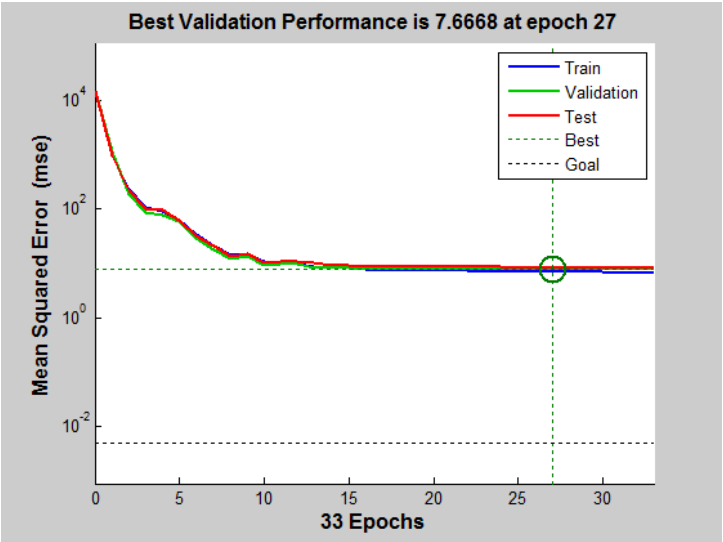

Fig 4.7 Performance of the MLP

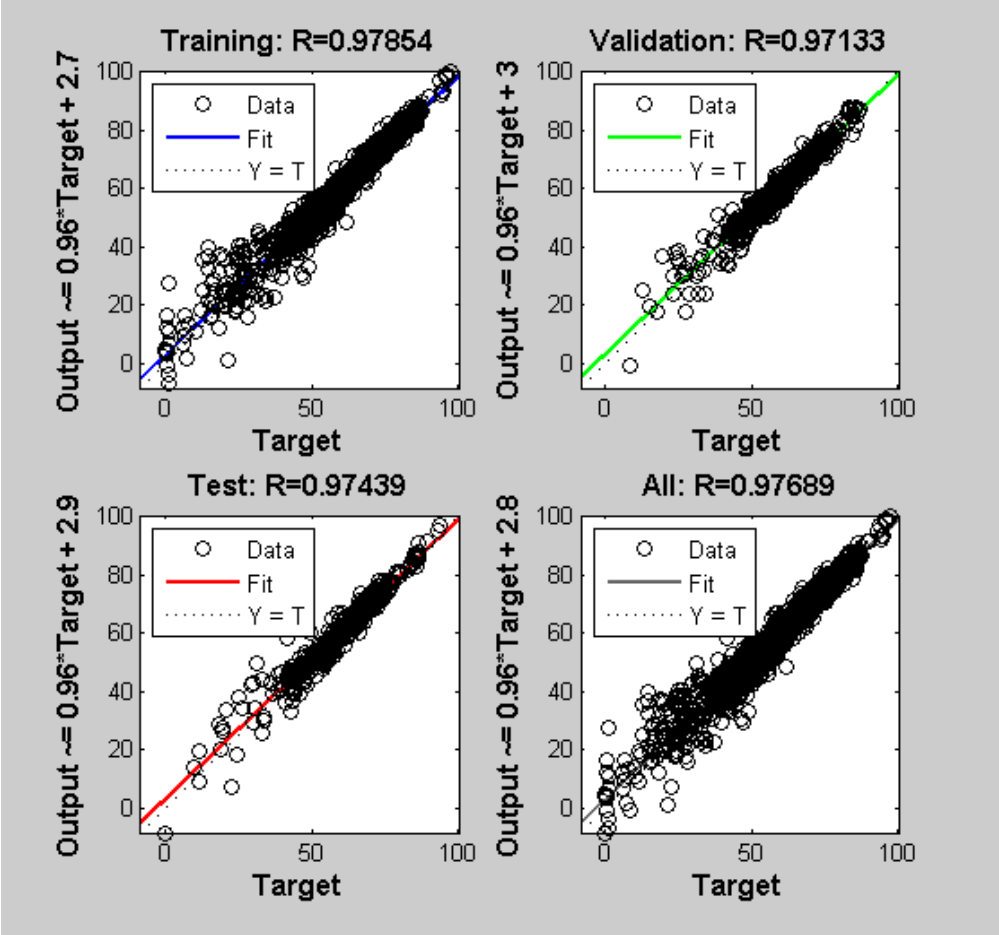

Fig 4.8 Regression of the MLP

Fig 4.9 shows the fit plot of the MLP regression.

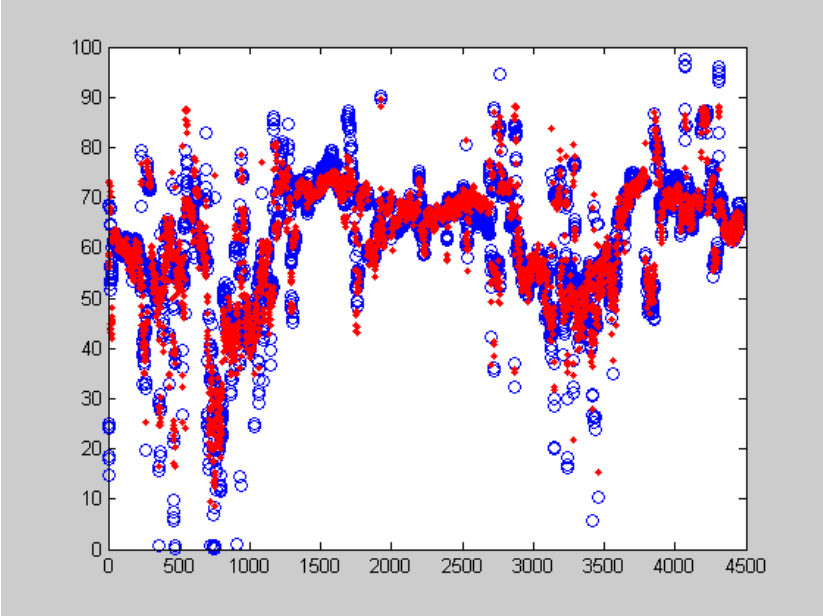

Fig 4.9 Fit plot of the MLP regression.

At the end, we use committee machine to combine all the three MLP regression models together, and use the MSE as the weight to get the new model. **The result of the new model is MSE=6.7328.**

5. Conclusion

For this project, we can get the conclusion from the result that committee machine have the best result for fit. Between the Linear Regression model, PLS Regression model and the MLP Regression model, the Linear Regression model have best result. In my opinion, I think it's because the data, I use the function to move the outlier in Linear Regression model, so I use the XtrainDS exclude the outliers, but the others two model I didn't use the XtrainDS exclude the outliers, that's one reason for the Linear Regression is the best, but the PLS Regression should be the worst model for this project, because there isn't a lot relation between the input variables.