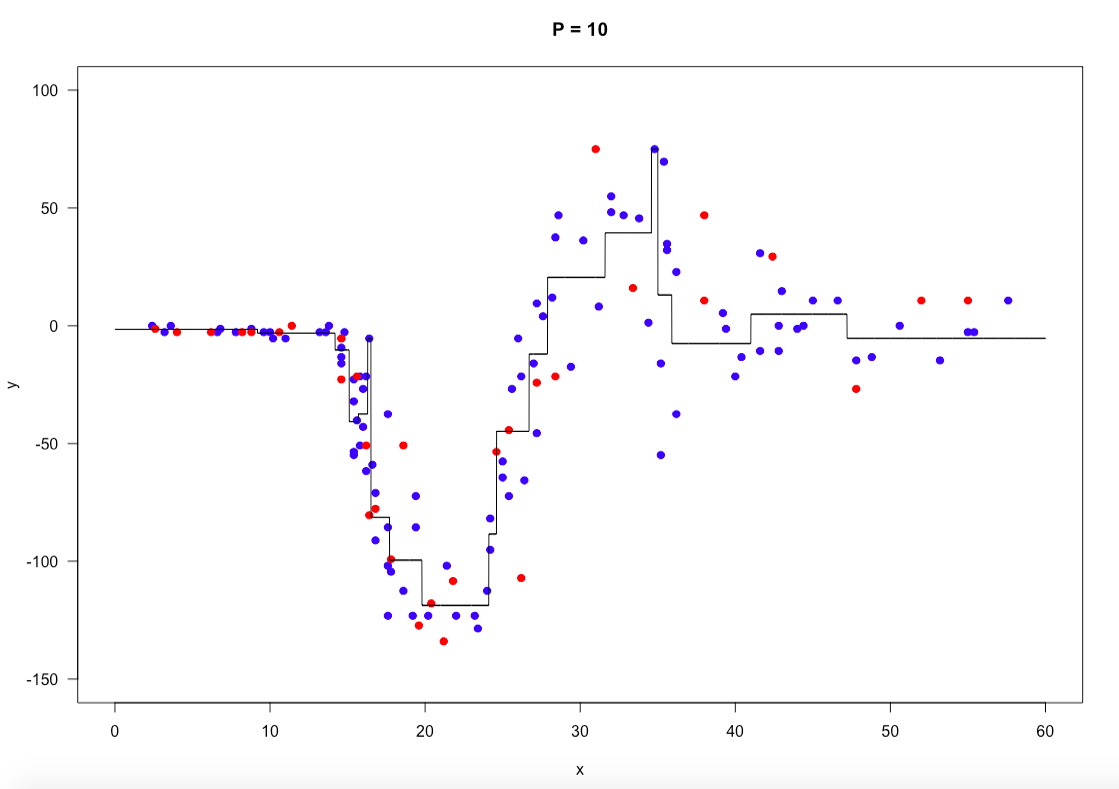
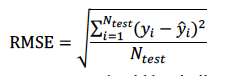
## COMP 421: Introduction to Machine Learning Homework #5

Initially, I have split my data as 100 of them being the training data and 33 of them as training data. For the regression algorithm we had to change actually two lines from our lab 7 code. 1 of them as being the node\_frequencies. Instead of this we have to calculate the means of the nodes since we are doing a regression. I have used mean function from the R to take the mean of my output training date. We have no classes in this data set unlike the lab’s data set where we had different classes, therefore I had to modify the remaining parts of the code accordingly. Additionally, in the homework we have a pre-pruning parameter which is set as 10 initially, that allows us to decide whether the node we’re currently in is a terminal node or not. For the second line of code change, I have changed how we were calculating the split scores. For the calculation of the split scores I have used the formula given on the figure and summed it for right and left indices in the code, which allows us to decide whether we should continue from the left branch of the tree or from the right branch of the tree. After training, I have calculated my y\_predicted values with using the mean function which is the same functiong that I have used through implementing my decision tree.Additionally, I have added an extra control check which is is.na to check whether my split nodes are NA or not because when we take P small enough, without reaching a terminal node it tries to check that index’s value however it is actually NA and throws an error. To prevent this error, I have assigned that index to terminal as it should be solved the problem. Afterwards, I have plotted my test and training data and for the plot line I have chosen a data interval that is between 0 to 60 (which is given in the homework’s description) and increased my interval by 0.01. For the regression line I have reused the method for my y\_predicted values but for this time I have expanded it with a for loop that extends from 1 to the length of data interval and I got the following figure which is the same as the figure given in the homework description:



After the plotting I have calculated the RMSE according to the given formula to get my error rate. The RMSE calculation formula is given in the figure. My RMSE is given on the below for when P is equal to 10.

Finally, I have iterated the RMSE’s for P is equal to 1 to P is equal to 20 and collected the RMSE values on a vector I had defined so that I would plot them and draw them. The same process for training and predicting my decision tree and my data when the P was equal to 10 as in the above case is done for when P is 1, 2, 3,….., 20. The resulting line is given in the figure below.

