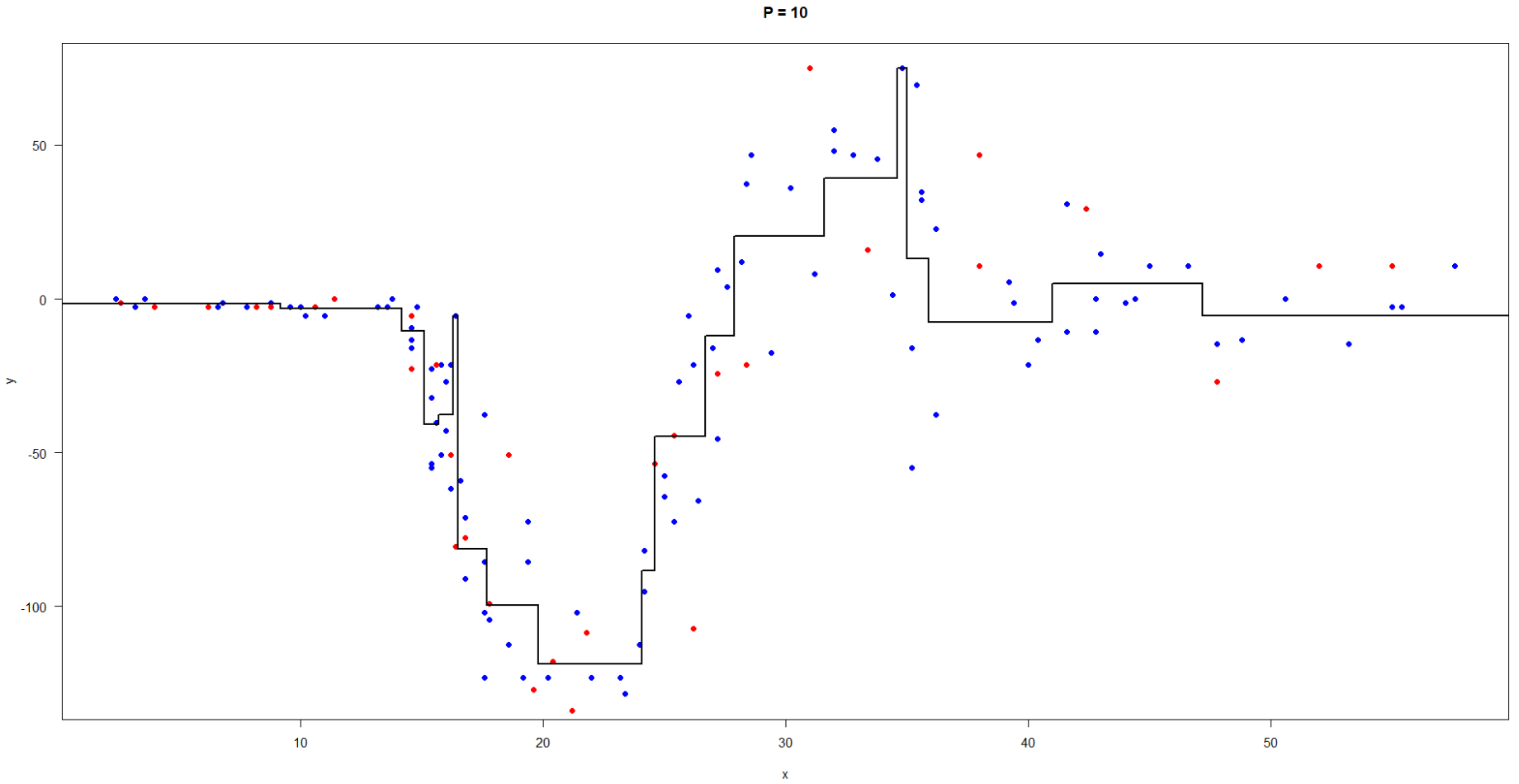
INDR421 – HW5
29-11-2018
Oğuzhan Akan (30223)

Approach

1- I first read and separated the data as instructed. Then I introduced the necessary parameters and data structures as shown in the lab session.
2- The critical points in the homework were (i) specifying what to use to determine the best split and (ii) what to use to obtain an output at each node. Best split can be found by averaging the sum-squared errors at each child-node, and the output at each node should be equal to the mean of the population that goes to the node.
3- Using a for loop, I calculated every possible split score for a parent node. Then I selected the best split which has the minimum squared error.
4- After building the tree, I found the predicted y values using test data. One obstacle here was to obtain y values when P = 1 or P = 2, because some split values were null as the data points were so close. I appended the argument is.na(node_splits[index]) for pereventing it to happen.
5- I calculated the RMSE, rules, and then plotted the data.
6- To run the code for P = 1, 2, ..., 20, I converted the code into a function named tree(x) where x = P. Then I ran the code for each P that is asked for, and calculated the RMSEs for each.
7- I shared the results in the following pages.

INDR421 – HW5
29-11-2018
Oğuzhan Akan (30223)



RMSE is 27.6841 when P is 10

INDR421 – HW5
29-11-2018
Oğuzhan Akan (30223)