Problem Set: Regression Splines

Problem 1:

In this problem, you will input the mall dataset to fit regression splines.

Recall that we are still in a one-predictor case here.

Part 1: Load packages and dataset

a) Install required packages: **splines**

b) Load the dataset Position Salary into R (file **mall.csv**)

c) Use attach() to make the database be attached to the R search path.

d) Use summary() and str() to take a brief observation of the dataset

Part 2: Fit a regression spline to predict spending score using age

e) Use bs() to fit regression splines to predict spending score using age

Hint: (1) Use cubic spline to fit each region

(2) Use only 1 knot here and choose the location on your own

f) Explain how you decide the location of the knot

g) Calculate the MSE of the regression splines

h) Plot the resulting fit and also the vertical line of where the knot is

Part 3: Fit a regression spline again to predict spending score using age

i) Use bs() to fit regression splines again and choose the number of knots and also the locations on your own

j) Explain how you decide the number and locations of these knots

k) Plot the resulting fit and also the vertical lines of where these knots are for exercise i

l) Use bs() to fit regression splines again and use attr(bs(…),’knots’) to return the position of the knots

Hint: (1) this time, let R place the corresponding number of knots at uniform quantiles of the data automatically

(2) make sure the number of knots are the same as exercise im) Calculate the MSE of the regression splines for both the manual-decided one and the R-decided one and compare the MSEs

Problem 2:

In this problem, you will input the mall dataset to fit regression splines and natural splines.

Part 1: Fit a regression spline and a natural spline to predict spending score using age

a) Use bs() and ns() to fit a regression spline for a range of degrees of freedom from 3 to 30

b) Report the associated MSE in a table

c) Plot the results

Part 2: Do cross-validation to determine the optimal number of knots

e) Set seed to 5072

f) Use cv.glm() to do cross validation here to decide how many knots to locate in order to generate the lowest MSE

g) Report the associated knots and MSEs in a table

h) Plot the results