Advanced Computational Econometrics
Chapter 4: Splines and GAM

Exercice 1

We consider again the `Boston` dataset from package `MASS`. We wish to predict variable `medv` (median value of owner-occupied homes in $1000s) as a function of `lstat` (lower status of the population in percent).

1. Estimate the expected value of `medv` as a function of `lstat` using order-p polynomial regression. Represent graphically the data and the estimated regression function for different values of p. Which values of p seem visually suitable ?
2. Determine the optimal value of p by cross-validation.
3. Same questions using natural splines. This time, the coefficient to be determined in the number of degrees of freedom (parameter `df` in function `ns`).
4. Same questions using smoothing splines (function `smooth.spline`). Find the optimal value of coefficient `df` using the leave-one-out, then let this coefficient vary around its optimal value and estimate the cross-validation error using the same folds as in the two previous questions.
5. Represent the regression functions estimated by the three methods on the same graph and compare their cross-validation errors.

Exercice 2

This time, we want to variable `medv` in the `Boston` dataset using as predictors variables `crim` (per capita crime rate), `lstat` (lower status of the population in percent), `dis` (weighted mean of distances to five Boston employment centers) and `nox` (nitrogen oxides concentration in parts per 10 million).

We will use an additive model using a smoothing spline transformation for each predictor. (Function `gam` in package `gam`).

1. Write a function that computes the cross-validation MSE as a function of the degrees of freedom of the smoothing splines.

2. Find the degrees of freedom that minimize the cross-validation MSE. (Use the Nelder-Mead optimization algorithm in function `optim`).
3. Plot the different additive components of the GAM together with the residuals.