Advanced Computational Econometrics
Chapter 4: Splines and GAM

Exercise 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 is 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.

Exercise 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 oxide concentration in parts per 10 million).

We will fit a generalized additive model (GAM) using a smoothing spline transformation for each predictor. (Function gam in package gam).

1. Fit a GAM with the default degrees of freedom for each of the smoothing splines. Plot each term in the additive model with the standard errors and the residuals. Interpret the results.

2. Manually tune the degrees of freedom to obtain smoother curves for each component of the additive model.
3. Compute the cross-validation MSE for the default and manually-tuned models. Compare the obtained performances to those of linear regression.