Exercise 6

Advanced Methods for Regression and Classification

December 15, 2022

Load the data `Auto` from the package `ISLR`. The data contain different car characteristics, information about the origin, and the name of the car.

1. (a) Consider the variable `mpg` as response, and `acceleration` as explanatory variable. Regress `mpg` on a B-spline basis of `acceleration`, with a desired number of degrees of freedom (see course notes), but compute separate models for each value of `origin`. Thus, you will have different models for American, European and Japanese cars. Plot `mpg` versus `acceleration`, and show lines with the fits for `mpg` for the three different models.
   Hint: In order to generate correct spline basis functions, you first need to sort the data group-wise according to the variable `acceleration`.

   (b) Same as above, but with Natural Cubic Splines.

   (c) Same as above, but with Smoothing Splines (here, sorting is not required).

2. Consider again `mpg` as response, and all the remaining variables (except `name`) as explanatory variables. Randomly select a training set of about 2/3 of the observations, build the model, and evaluate the model for the remaining test set data.

   Linear model with natural cubic splines: Use the function `ns()` from the `library(splines)` and the function `lm()`.

   The form of the model is

   $$y = \theta_0 + \boldsymbol{h}_1(x_1)^\top \boldsymbol{\theta}_1 + \boldsymbol{h}_2(x_2)^\top \boldsymbol{\theta}_2 + \ldots + \boldsymbol{h}_p(x_p)^\top \boldsymbol{\theta}_p + \varepsilon,$$

   where each $\boldsymbol{\theta}_j$ ($j = 1, \ldots, p$) is a vector of coefficients that is multiplied by the basis function $\boldsymbol{h}_j$ (natural cubic splines) for the jth input variable.

   Every term in the model should be represented by 4 natural cubic splines. However, for some input variables (binary, categorical) this might not make sense, and they should enter the model in the usual way without splines.

   (a) Which variables (basis functions) are significant? Calculate the RMSE (root mean squared error) for the test set.

   (b) Apply stepwise variable selection using
   `step(...,direction="both")`. Which variables (basis functions) are significant? Compute the RMSE for the test set.

   (c) Plot the variables from the reduced model (b) against their estimated values, so e.g. x_j against $\hat{f}_j(x_j) = \boldsymbol{h}_j(x_j)^\top \hat{\boldsymbol{\theta}}_j$. How can you interpret these plots?