Week 9: Exercices Splines and GAMs

Marjolein Fokkema

Exercise 1: Fit natural and cubic splines

Load the Boston Housing data:

```
library("MASS")
data(Boston)
```

Set up a cubic spline basis for variable `lstat`:

```
library("splines")
basis <- bs(Boston$lstat, df = 5)
```

a) Print and inspect the result: Where are the knots located? How many basis functions were generated? Create a plot with the value of each basis function on the y-axis and the `lstat` variable on the x-axis.

Regress response variable `medv` on the cubic spline just created:

```
library("gam")
mod_df5 <- gam(medv ~ bs(lstat, df = 5), data = Boston)
```

Also predict `medv` with a cubic spline with 8 df, and a linear spline with 2 df (to obtain a linear spline, you also need to specify argument `degree` of the `bs()` function, in addition to the `df` argument).

b) Inspect, interpret and compare the three fitted models using `summary()`, `BIC()` and `plot()` (make sure to specify `residuals = TRUE` when plotting).

c) Which spline model fits best?

Note that one could also extract the individual coefficients estimated for each of the basis functions using `coef()`, but we usually are not interested in those because they are very difficult to interpret (e.g., unclear what is a large or small value, what constitutes a negative or positive effect). We mostly care about the fitted *spline* curve, which is the sum of the basis functions, multiplied by their coefficients.

Now fit a natural cubic spline model, one with 3 and one with 5 df to predict `medv` based on `lstsat`. Hint: use function `ns()` instead of `bs()` to set up the spline basis; use function `gam()` to fit the model as before.

d) Where are the knots located for the 3 df natural spline? Which cubic spline yielded identical knot locations? What's the difference in df?

e) Inspect, interpret and compare the fitted models using `summary()`, `BIC()` and `plot()`. Which model fits best? Do you think the natural splines are an improvement over the cubic splines?

Note that the splines we have fitted up til now are parametric, so you could also have used function `lm()` or `glm()` instead of `gam()` to fit the models (but with `gam()` we obtained better plots).

Exercise 2: Fit a smoothing spline

Use function `s()` to fit a smoothing spline to the `lstat` variable and predict `medv`. Use both a high (e.g., > 100) and a low value (e.g., a value between 3 and 8) for the `df` argument of function `s()`.

Inspect, interpret and compare the fitted models using `summary()`, `plot()`, `BIC()`. Do the smoothing splines provide an improvement over the parametric natural and cubic splines fitted earlier?

Exercise 3: Fit a GAM (multiple predictor variables)

Fit a GAM using function `gam` from package `mgcv` to the Boston Housing data.

Before loading `mgcv`, unload package `gam` to avoid confusion of the `gam()` and `s()` functions:

```
detach("package:gam", unload = TRUE)
```

Use `lstat`, `rm`, `ptratio`, `crim` and `dis` as predictors of the repsonse `medv` (check `?Boston` to see the meaning of these variables) using smoothing splines (i.e., wrap each predictor in `s()` in the model formula). Specify `method = "REML"` to employ REML estimation.

Use `plot` (it is often helpful to additionally specify `residuals = TRUE`) and `summary` to evaluate the fitted model and curves.

a) Which predictor variables seem to be most important (inspect F-values in the output of `summary`; inspect the range of the y-axes in the output of `plot`)?

b) Which variables' effects deviate from a linear effect? Hint: Inspect the `edf` column in the output of `summary`. How many df would be taken up by a purely linear effect?

c) Which predictor variables have a significant effect?

Exercise 4: Prove continuity of cubic spline

We have seen that a cubic regression spline with one knot ξ can be obtained using a basis of the form 1 (or X^0), X, X^2, X^3 and $(X-\xi)^3_+$.

Show that a function of the form

$$f(X) = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \beta_4 (X-\xi)^3_+$$

is indeed a cubic *spline*.

First, establish that $f(X)$ is a piecewise polynomial:

a) Find a cubic polynomial $f_1(X) = a_1 + b_1 X + c_1 X^2 + d_1 X^3$ such that $f(X) = f_1(X)$ for all $X \leq \xi$. That is, express a_1, b_1, c_1, d_1 in terms of β_0, β_1, β_2, β_3, β_4.

b) Find a cubic polynomial $f_2(X) = a_2 + b_2 X + c_2 X^2 + d_2 X^3$ such that $f(X) = f_2(X)$ for all $X > \xi$. That is, express a_2, b_2, c_2, d_2 in terms of β_0, β_1, β_2, β_3, β_4.

Now show that:

- $f_1(\xi) = f_2(\xi)$. That is, that $f(X)$ is continuous at ξ.

- $f_1'(\xi) = f_2'(\xi)$. That is, that $f'(X)$ is continuous at ξ.
- $f_1''(\xi) = f_2''(\xi)$. That is, that $f''(X)$ is continuous at ξ.

Hint: Given a cubic polynomial $f_1(X) = a_1 + b_1 X + c_1 X^2 + d_1 X^3$, the first derivative is given by $b_1 + 2c_1 X + 3d_1 X^2$.