STATS 235: Modern Data Analysis
Generalized Additive Models

Babak Shahbaba

Department of Statistics, UCI

Background

- Previously, we discussed splines as a class of models to handle nonlinear relationships between the response variables and predictors.
- In this lecture, we discuss "generalized additive models" (GAM) as an alternative approach to build nonlinear regression models.
- These models have the following form:

$$E(y|x) = \mu(x) \quad = \quad \alpha + f_1(x_1) + \ldots + f_p(x_p)$$

or in general,

$$g(\mu(x)) \quad = \quad \alpha + f_1(x_1) + \ldots + f_p(x_p)$$

where f_j are *smooth* (nonparametric) functions (we can make some of f_j simple linear functions)

- We could find each f_j using basis expansion, and then fit the overall model using the least squares method
- Instead, we are interested in fitting these functions simultaneously

Fitting additive models

- The functions f_j are typically estimated using a scatterplot smoother such as the cubic smoothing spline discussed in the previous lecture.

- We then minimize the following penalized residual sum of squares:

$$
\begin{aligned}
PRSS(\alpha, f_1, \ldots, f_p) &= \sum_{i=1}^{N} \Big(y_i - \alpha - \sum_{j=1}^{p} f_j(x_{ij}) \Big)^2 + \\
&\quad \sum_{j=1}^{p} \lambda_j \int f_j^{''}(t_j)^2 dt_j
\end{aligned}
$$

- Each function f_j is a cubic spline depending on x_j only and with knots at each unique values of x_{ij}.

- For identifiability, we set $\sum_{i=1}^{N} f_j(x_{ij}) = 0$ for all j.

Backfitting

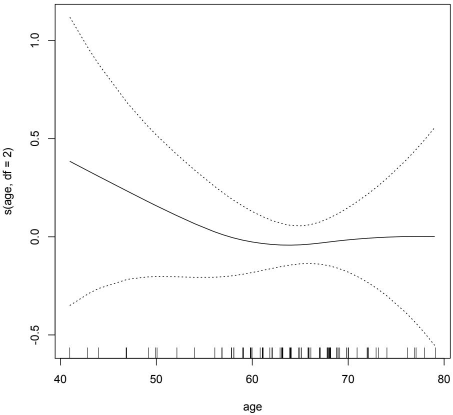
- We can use the following iterative procedure, called *backfitting*, to estimate the functions:
  1. Initialize $\hat{\alpha} = \mathrm{avg}(y_i)$, $\hat{f}_j \equiv 0, \quad \forall i, j$
  2. Iteratively update the functions f_j as follows until they stabilize (i.e., they don't change substantially from one iteration to another):
     1. Apply a cubic smoothing spline S_j to model $\{y_i - \hat{\alpha} - \sum_{k \neq j} \hat{f}_k(x_{ik})\}_1^N$ as a function of x_{ij} to obtain a new stimate $\hat{f}_j$
     2. Center $\hat{f}_j$ so its mean becomes zero (i.e., subtract the mean).

Example

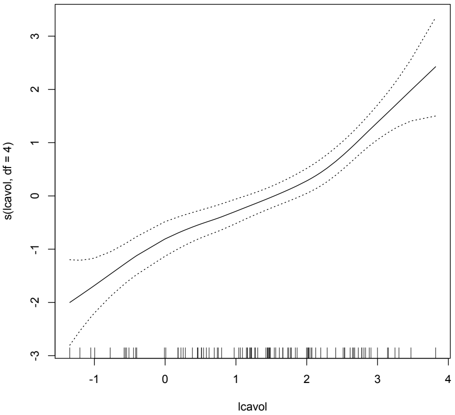
- For the prostate datasets, we use the `gam` library to build a generalized additive model for `lpsa` as a nonlinear function of `age`, `lcavol`, and `bph`.

```
gam1 <- gam(lpsa ~ s(age, df=2) + s(lcavol, df=4) +
s(lbph, df=3), data = Prostate)
```

Example



Example



Example