STATS 415: Generalized Additive Models

Prof. Liza Levina

Department of Statistics, University of Michigan

Generalized Additive Models

- Regression and smoothing splines are nice for one predictor x. But what if there is more than one?
- The relationship between y and each of the x's might be non-linear
- A natural way to extend the multiple linear regression model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon$$

is to replace each linear term with a non-linear function:

$$y = \beta_0 + f_1(x_1) + f_2(x_2) + \cdots + f_p(x_p) + \varepsilon$$

Generalized Additive Models (GAMs)

$$y = \beta_0 + f_1(x_1) + f_2(x_2) + \cdots + f_p(x_p) + \varepsilon$$

- Additive: we add separate functions of each of the predictors
- No interactions, but additional variables of the form $x_1 x_2$ can always be added.
- Some interpretation is retained, because we can look at individual $f_j(x_j)$ for each j
- To understand relationships between x_j's and y better, and avoid overfitting, smooth functions f_j are preferred.

Choosing functions for GAMs

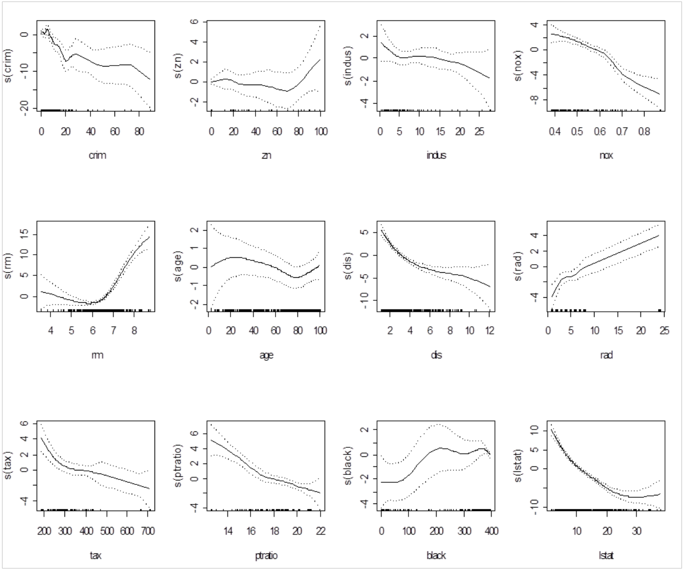
- There are many different choices for representing f_j.
- f_j could be a polynomial, as in polynomial regression.
- f_j could be represented in some other functional basis
- One of the most common approaches is to model f_j's as a spline
- Cubic/natural splines and smoothing splines are both used and often work well

Example: Boston housing data

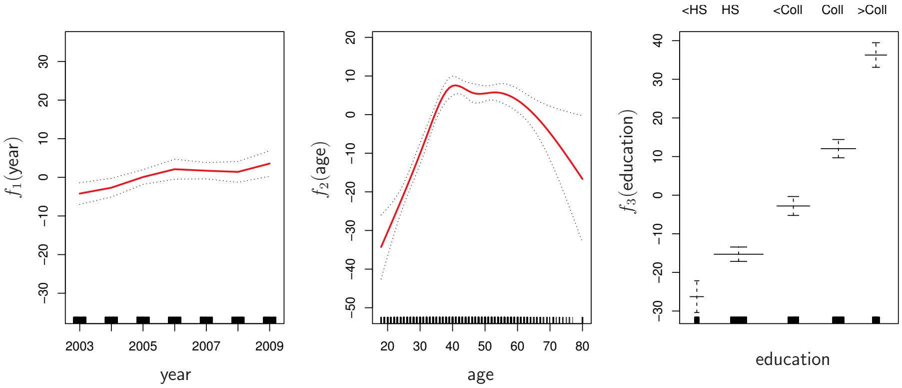
- Famous dataset on neighborhoods in Boston suburbs
- predicting median house price from variables such as crime rate, distance to the river, air quality, tax rate, student-teacher ratio, average number of rooms per dwelling, etc.
- Compare MSEs on a randomly chosen validation set of 100 observations

Method	MSE on Validation Set
Just the mean	104.93 (10.24)
Linear regression	35.36 (5.95)
GAM	22.01 (4.69)

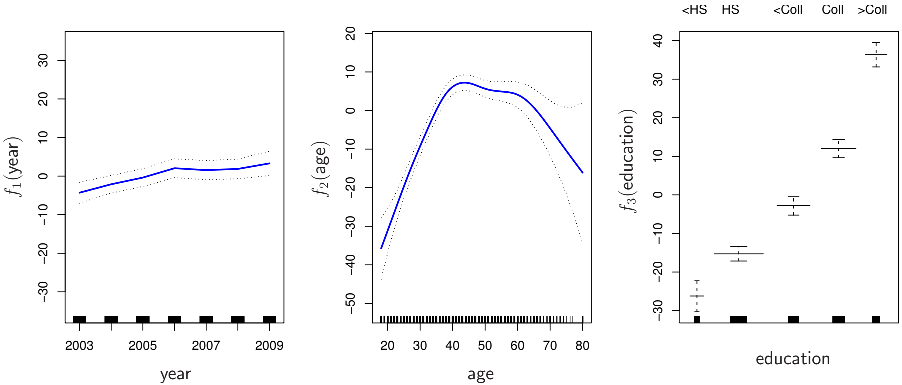
GAM on the Boston Housing Data



GAM on the Wage Data: natural splines



GAM on the Wage Data: smoothing splines



GAM for Classification

- We can also use GAM to make predictions for a categorical y, i.e. in classification.
- Recall logistic regression:

$$\log\left(\frac{P(y=c_1|X=x)}{1-P(y=c_1|X=x)}\right)=\beta_0+\sum_{j=1}^{p}\beta_j x_j$$

where $P(y=c_1|X=x)$ is the probability of class 1 given the values of predictors x.

Generalizing Logistic Regression: Logistic GAM

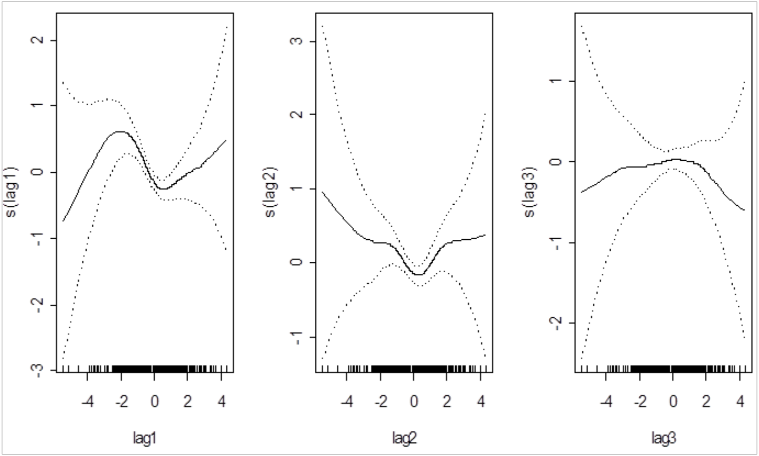
- The natural extension for logistic regression of the GAM framework is

$$\log\left(\frac{P(y=c_1|x)}{1-P(y=c_1|x)}\right)=\beta_0+\sum_{j=1}^{p}f_j(x_j)$$

- Not all terms have to be modeled the same way. It is ok to have a mix of linear, polynomial, and spline terms, etc.

GAM on S&P500 Data

- There do seem to be some patterns here.
- In particular there seems to be a non-linear relationship between lag1 and the probability that the market goes up today.



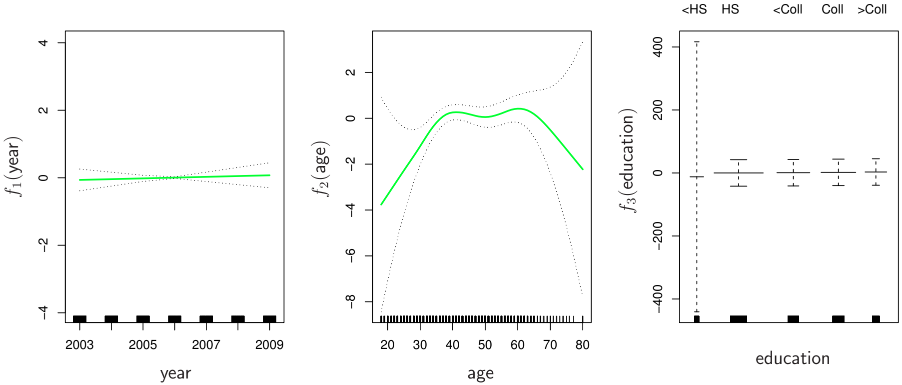
A Comparison on the S&P 500 Data

- As a comparison here is the fraction of the time we are correct, using different methods, on the last 242 days using the first 1000 days to build the models.

Method	% of Days Correct
Always Down	48.4%
Always Up	51.6%
Opposite of Yesterday	52.5%
Logistic Regression	47.9%
GAM	55.8%

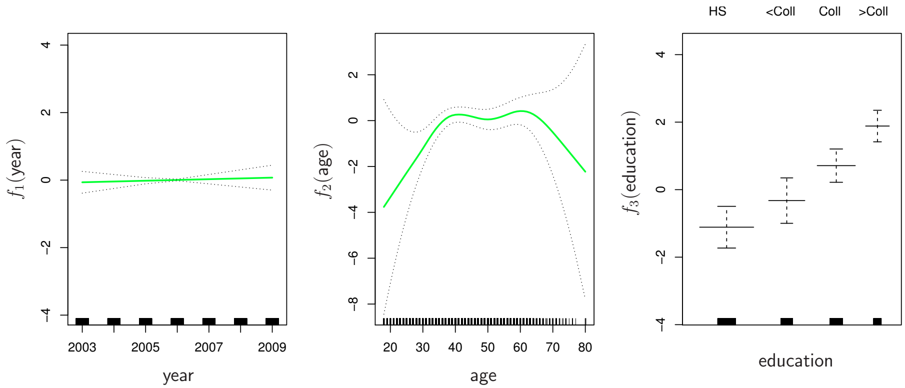
GAM on the Wage Data (classification)

- Define the outcome as 1 if wage $> \$250K$
- Model different variables in difference ways
- Year: linear; age: smoothing spline; education (categorical): changes in intercept



GAM on the Wage Data (classification)

- Removing the $<$ HS education level drastically reduces noise



Pros of GAM

- By fitting a non-linear f_j to each x_j we can automatically model non-linear relationships that standard linear or logistic regression will miss.
- Hence we can potentially make more accurate predictions.
- An additive model still allows us to examine the effect of each x_j on y individually while holding all the other x's fixed. Thus we still have some interpretation and inference.

Cons of GAM

- The model is restricted to be additive.
- Therefore cannot model interactions; for example, the simple model $y = x_1 x_2 + \varepsilon$ cannot be fitted well by $y = f_1(x_1) + f_2(x_2) + \varepsilon$ (can still manually add an interaction term)
- Sometimes interpretation is far from clear
- Tuning each function separately is computationally prohibitive (e.g. λ in a smoothing spline); choosing the same tuning parameter for all the functions does not always work well.