

Lecture 8: Beyond linearity

STATS 202: Data Mining and Analysis

Linh Tran

tranlm@stanford.edu



Department of Statistics
Stanford University

July 24, 2023



- ▶ HW2 and Midterm are being graded.
- ▶ Survey is up until this Friday.
- ▶ No session this Friday.
- ▶ 3 weeks left!
 - ▶ Kaggle submissions due in two weeks



- ▶ Extending linear models
- ▶ Basis functions
- ▶ Piecewise models
- ▶ Smoothing splines
- ▶ Local models
- ▶ Generalized additive models



We can use linear models to for regression / classification,
e.g.

$$\mathbb{E}[Y|\mathbf{X}] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots \quad (1)$$

- ▶ When relationships are non-linear, could e.g. include polynomial terms (e.g. X^2, X^3, \dots) or interactions (e.g. $X_1 * X_2, \dots$)
- ▶ Could then use e.g. stepwise regression to do model fitting
 - ▶ n.b. Recall that using too many parameters can result in overfitting



We can use linear models to for regression / classification,
e.g.

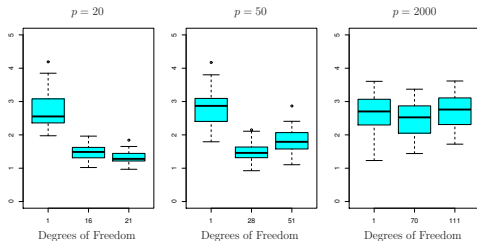
$$\mathbb{E}[Y|\mathbf{X}] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots \quad (1)$$

- ▶ When relationships are non-linear, could e.g. include polynomial terms (e.g. X^2, X^3, \dots) or interactions (e.g. $X_1 * X_2, \dots$)
- ▶ Could then use e.g. stepwise regression to do model fitting
 - ▶ n.b. Recall that using too many parameters can result in overfitting

Question: Assuming $n \gg p$, why not just do every transformation we can think of and throw all of them into e.g. Lasso?

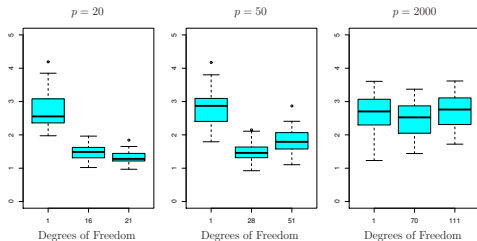


Consider a distribution where only 20 predictors are truly associated to the outcome (out of 20, 50, or 2000 total predictors)





Consider a distribution where only 20 predictors are truly associated to the outcome (out of 20, 50, or 2000 total predictors)



- ▶ Test error increases with more predictors and parameters
- ▶ *Lesson:* Using too many predictors/parameters can hurt performance
 - ▶ Selecting predictors carefully can/will lead to better performance



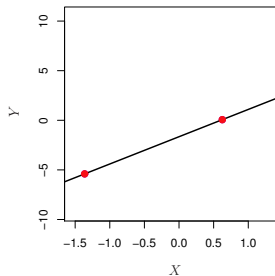
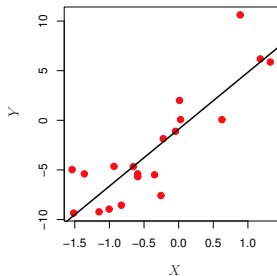
Question: What about if we have $p \geq n$? (pretty common now, since collecting data has become cheap and easy.)

Examples:

- ▶ **Medicine:** Instead of regressing heart disease onto just a few clinical observations (blood pressure, salt consumption, age), we use in addition 500,000 single nucleotide polymorphisms.
- ▶ **Marketing:** Using search terms to understand online shopping patterns. A *bag of words* model defines one feature for every possible search term, which counts the number of times the term appears in a person's search. There can be as many features as words in the dictionary.

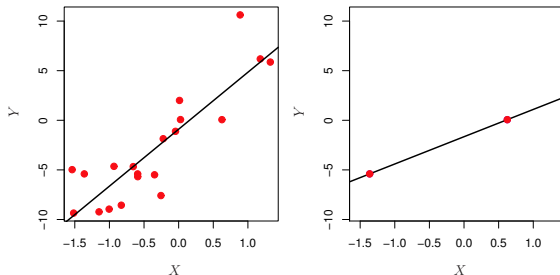


When $n = p$, we can find a perfect fit for the training data, e.g.





When $n = p$, we can find a perfect fit for the training data, e.g.

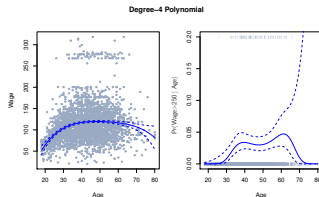


When $p > n$, OLS doesn't have a unique solution.

- ▶ Can use e.g. stepwise variable selection, ridge, lasso, etc.
- ▶ Still have to deal with estimating $\hat{\sigma}^2$.



Back to using polynomial terms



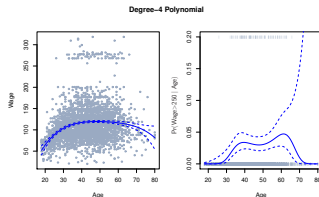
- ▶ Generalizing the notation, we have

$$\mathbb{E}[Y|\mathbf{X}] = \beta_0 + \beta_1 b_1(X) + \beta_2 b_2(X) + \dots + \beta_d b_d(X) \quad (2)$$

- ▶ where $b_i(x) = x^i$



Back to using polynomial terms



- ▶ Generalizing the notation, we have

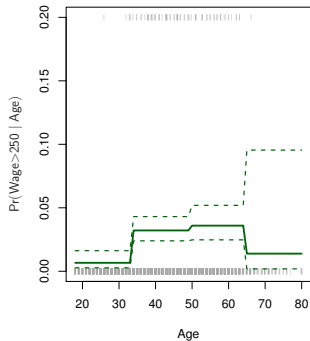
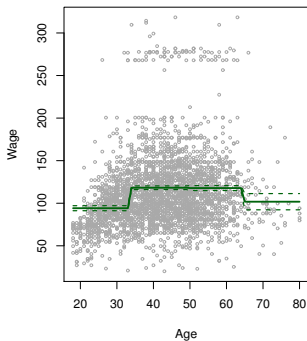
$$\mathbb{E}[Y|\mathbf{X}] = \beta_0 + \beta_1 b_1(X) + \beta_2 b_2(X) + \dots + \beta_d b_d(X) \quad (2)$$

- ▶ where $b_i(x) = x^i$
- ▶ We don't have to just use polynomials (e.g. could use cutpoints instead)
 - ▶ i.e. $b_i(x) = \mathbb{I}(c_i \leq x < c_{i+1})$



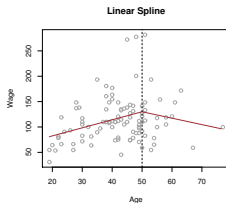
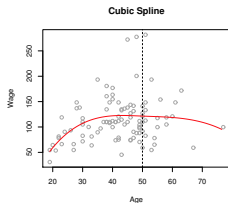
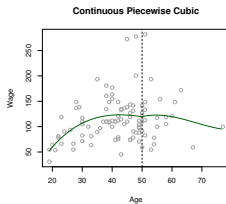
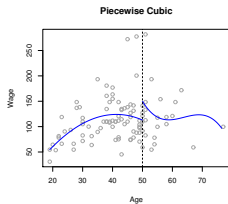
Example using *Wage* data.

Piecewise Constant





We can also use other *basis functions*,
e.g. piecewise polynomials



Note: points where coefficients change are called a *knots*.



Example of a piecewise cubic model with 1 knot & $d = 3$:

$$y_i = \begin{cases} \beta_{01} + \beta_{11}x_i + \beta_{21}x_i^2 + \beta_{31}x_i^3 + \epsilon_i & \text{if } x_i < c \\ \beta_{02} + \beta_{12}x_i + \beta_{22}x_i^2 + \beta_{32}x_i^3 + \epsilon_i & \text{if } x_i \geq c \end{cases} \quad (3)$$

i.e. We're fitting two regression models (for $x_i < c$ and $x_i \geq c$)

- ▶ This can be generalized to any number of knots $\leq n$
- ▶ Problem: resulting function is discontinuous



Very popular, since they give very smooth predictions over X .

- ▶ Define a set of knots $\xi_1 < \xi_2 < \dots < \xi_K$.



Very popular, since they give very smooth predictions over X .

- ▶ Define a set of knots $\xi_1 < \xi_2 < \dots < \xi_K$.
- ▶ We want the function $Y = f(X)$ to:
 1. Be a cubic polynomial between every pair of knots ξ_i, ξ_{i+1} .
 2. Be continuous at each knot.
 3. Have continuous first and second derivatives at each knot.



Very popular, since they give very smooth predictions over X .

- ▶ Define a set of knots $\xi_1 < \xi_2 < \dots < \xi_K$.
- ▶ We want the function $Y = f(X)$ to:
 1. Be a cubic polynomial between every pair of knots ξ_i, ξ_{i+1} .
 2. Be continuous at each knot.
 3. Have continuous first and second derivatives at each knot.

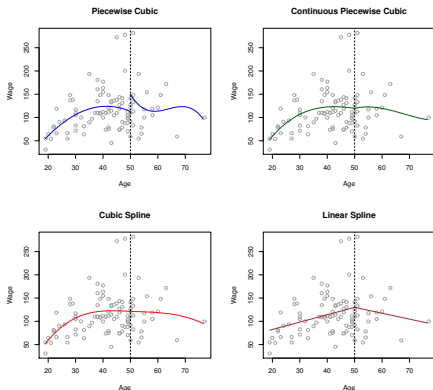
Fact: Given constraints, we need $K + 3$ basis functions:

$$f(X) = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \beta_4 h(X, \xi_1) + \dots + \beta_{K+3} h(X, \xi_K) \quad (4)$$

where,

$$h(x, \xi) = \begin{cases} (x - \xi)^3 & \text{if } x > \xi \\ 0 & \text{otherwise} \end{cases}$$

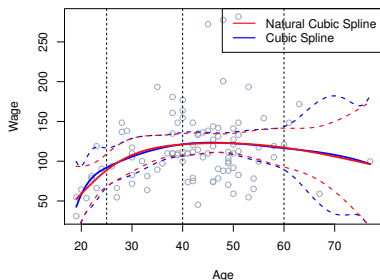
Piecewise polynomials vs cubic splines



- ▶ Piecewise model has $df = 8$.
- ▶ Cubic spline has $df = 5$.
- ▶ In general, k knots result in $df = 4 + K$



Spline which is linear instead of cubic for $X < \xi_1$ & $X > \xi_K$

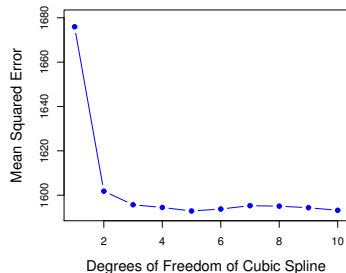
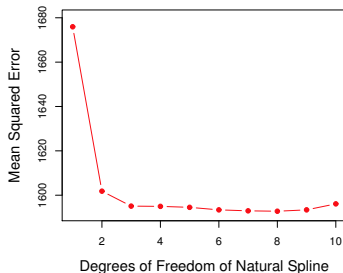


Makes predictions more stable for extreme values of X

- ▶ Will typically have less degrees of freedom than cubic splines



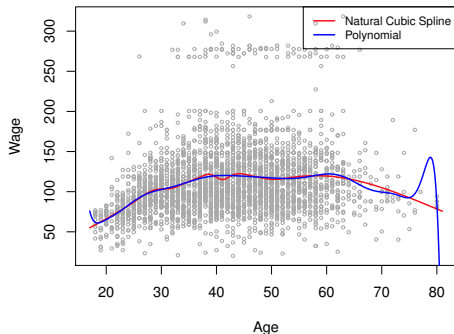
- ▶ The locations of the knots are typically quantiles of X .
- ▶ The number of knots, K , is chosen by cross validation



Splines vs polynomial regression



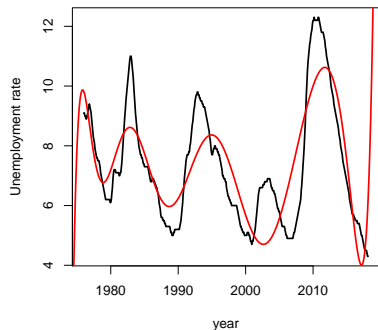
- ▶ Splines can fit complex functions with few parameters
- ▶ Polynomials require high degree terms to be flexible
 - ▶ May also result in non-convergence
- ▶ High-degree polynomials (like vanilla cubic splines) can be unstable at the edges





Applied example: California unemployment rates

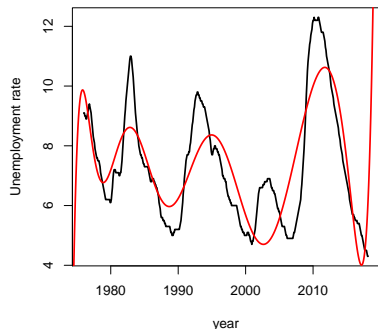
Polynomial regression (df=12)



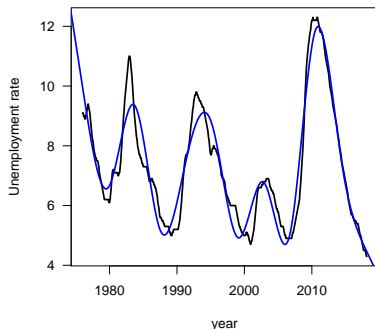


Applied example: California unemployment rates

Polynomial regression (df=12)



Natural cubic spline (df=12)





Our goal is to find the function f which minimizes

$$\sum_{i=1}^n (y_i - f(x_i))^2 + \lambda \int f''(x)^2 dx$$

- ▶ The RSS of the model.
- ▶ A penalty for the roughness of the function.



Our goal is to find the function f which minimizes

$$\sum_{i=1}^n (y_i - f(x_i))^2 + \lambda \int f''(x)^2 dx$$

- ▶ The RSS of the model.
- ▶ A penalty for the roughness of the function.

For regularization, we have that $\lambda \in (0, \infty)$

- ▶ When $\lambda = 0$, f can be any function that interpolates the data.
- ▶ When $\lambda = \infty$, f will be the simple least squares fit



Notes:

- ▶ \hat{f} becomes more flexible as $\lambda \rightarrow 0$.
- ▶ The minimizer \hat{f} is a natural cubic spline, with knots at each sample point x_1, \dots, x_n .
 - ▶ Though usually not the same as directly applying natural cubic splines due to λ
 - ▶ Our function can therefore be represented as

$$\hat{f}(x) = \sum_{j=1}^n N_j(x) \hat{\theta}_j \quad (5)$$

- ▶ Obtaining \hat{f} is similar to a Ridge regression, i.e.

$$\hat{\theta} = (\mathbf{N}^T \mathbf{N} + \lambda \Sigma_N)^{-1} \mathbf{N}^T \mathbf{y} \quad (6)$$



Natural cubic splines

- ▶ Fix $K \leq n$ and choose the locations of K knots at quantiles of X .
- ▶ Find the natural cubic spline \hat{f} which minimizes the RSS:

$$\sum_{i=1}^n (y_i - f(x_i))^2$$

- ▶ Choose K by cross validation.

Smoothing splines

- ▶ Put n knots at x_1, \dots, x_n .
- ▶ Obtain the fitted values which minimizes:

$$\sum_{i=1}^n (y_i - f(x_i))^2 + \lambda \int f''(x)^2 dx$$

- ▶ Choose λ by cross validation.



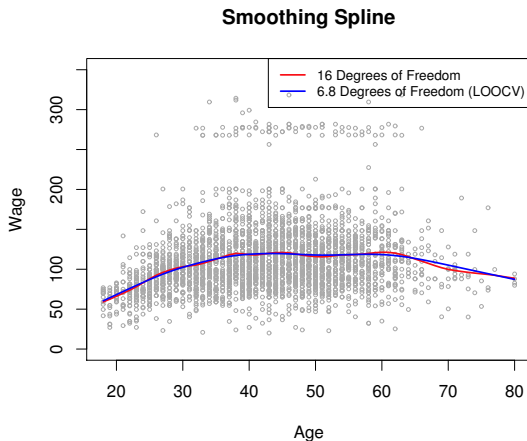
Similar to linear regression, there's a shortcut for LOOCV:

$$CV_{(n)} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{f}_{\lambda}^{(-i)}(x_i))^2 = \frac{1}{n} \sum_{i=1}^n \left[\frac{y_i - \hat{f}_{\lambda}(x_i)}{1 - \{\mathbf{S}_{\lambda}\}_{ii}} \right]^2 \quad (7)$$

for an $n \times n$ matrix \mathbf{S}_{λ}



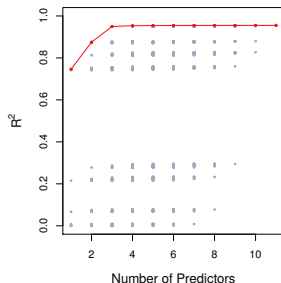
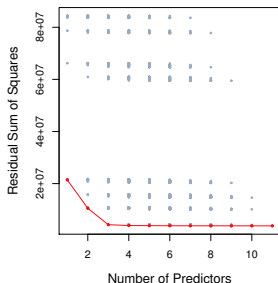
Example



n.b. We're measuring **effective** degrees of freedom



Idea: Why not just use the subset of observations *closest* to the point we're predicting at?

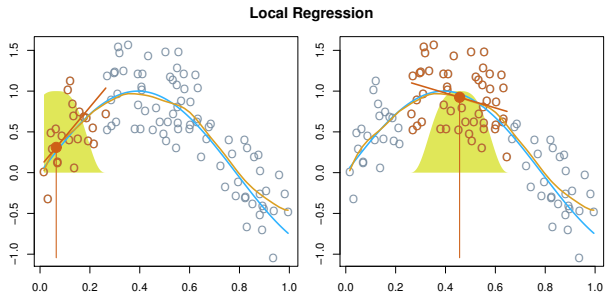


- ▶ Observations averaged *locally* for predictions.
- ▶ Can use different weighting kernels, e.g.

$$K_\lambda(x_0, x) = D \left(\frac{|x - x_0|}{h_\lambda(x_0)} \right) \quad (8)$$



Idea: Kernel smoothing with regression.



e.g. Perform regression *locally*.

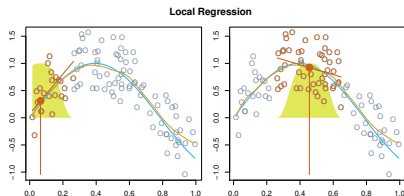


To predict the regression function f at an input x :

1. Assign a weight K_i to the training point x_i , such that:
 - ▶ $K_i = 0$ unless x_i is one of the k nearest neighbors of x .
 - ▶ K_i decreases when the distance $d(x, x_i)$ increases.
2. Perform a weighted least squares regression; i.e. find (β_0, β_1) which minimize

$$\sum_{i=1}^n K_i (y_i - \beta_0 - \beta_1 x_i)^2.$$

3. Predict $\hat{f}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$.



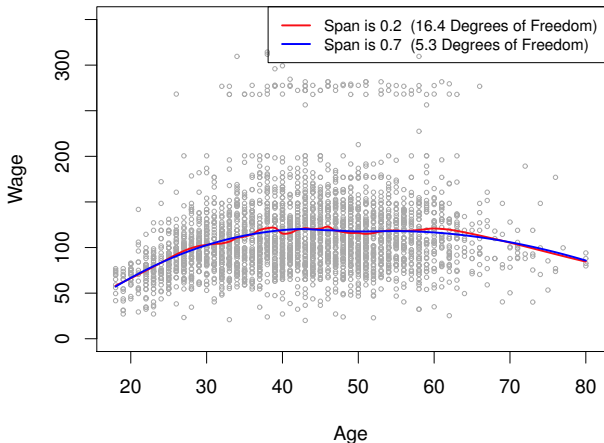
Some notes:

- ▶ The **span** k/n is the fraction of training samples used in each regression (the most important hyperparameter for the model)
- ▶ Sometimes referred to as a *memory-based* procedure
- ▶ Can mix this with linear models (e.g. global in some variables, but local in others)
- ▶ Performs poorly in large dimensional spaces

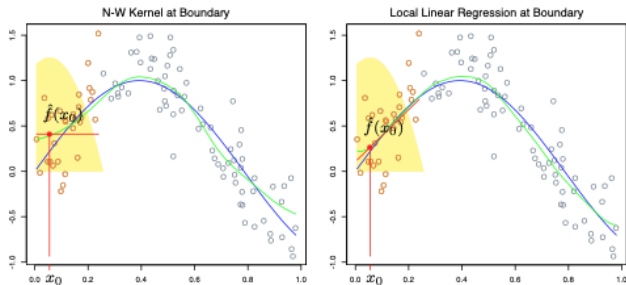


Example

Local Linear Regression



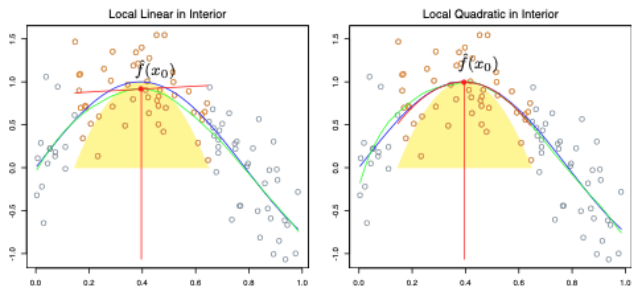
Kernel smoothing vs Local linear regression



The main differences occur in the edges.



We don't have to stick with local linear models, e.g.



Can deal with the issue of “*trimming the hills*” and “*filling the valleys*”.



The extension of basis functions to multiple predictors (while maintaining additivity) , e.g.

Linear model

$$\text{wage} = \beta_0 + \beta_1 \text{year} + \beta_2 \text{age} + \beta_3 \text{education} + \epsilon \quad (9)$$

Additive model

$$\text{wage} = \beta_0 + f_1(\text{year}) + f_2(\text{age}) + f_3(\text{education}) + \epsilon \quad (10)$$

The functions f_1, \dots, f_p can be polynomials, natural splines, smoothing splines, local regressions, etc.



If the functions f_j have basis representations, we can use least squares

- ▶ Natural cubic splines
- ▶ Polynomials
- ▶ Step functions

Otherwise, we can use *backfitting* to fit our functions

- ▶ Smoothing splines
- ▶ Local regression



- ▶ Otherwise, we can use **backfitting**:

1. Keep f_2, \dots, f_p fixed, and fit f_1 using the partial residuals:

$$y_i - \beta_0 - f_2(x_{i2}) - \dots - f_p(x_{ip}),$$

as the response.

2. Keep f_1, f_3, \dots, f_p fixed, and fit f_2 using the partial residuals:

$$y_i - \beta_0 - f_1(x_{i1}) - f_3(x_{i3}) - \dots - f_p(x_{ip}),$$

as the response.

3. ...

4. Iterate

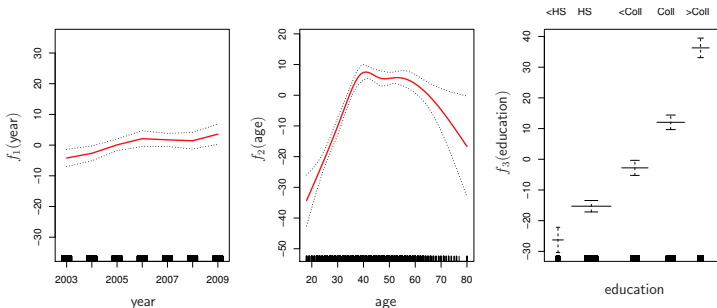
- ▶ This works for smoothing splines and local regression.



- ▶ GAMs are a step from linear regression toward a nonparametric method
 - ▶ Mitigates need to manually try out many different transformations on each variable individually
 - ▶ We can report degrees of freedom for most non-linear functions (as a way of representing model complexity)
 - ▶ The only constraint is additivity, which can be partially addressed through adding interaction variables, e.g. $X_i X_j$
- ▶ Similar to linear regression, we can examine the significance of each of the variables



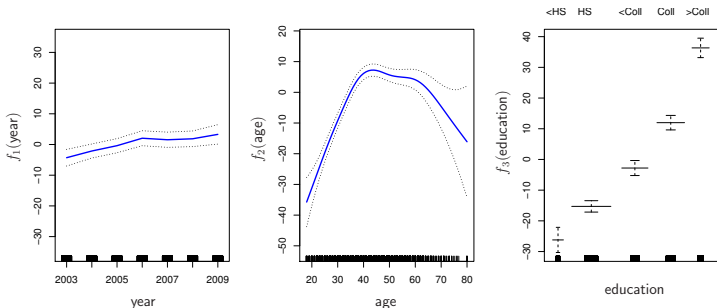
Using natural cubic splines:



- ▶ year: natural spline with $df=4$.
- ▶ age: natural spline with $df=5$.
- ▶ education: step function.



Using smoothing splines:



- ▶ year: smoothing spline with $df=4$.
- ▶ age: smoothing spline with $df=5$.
- ▶ education: step function.



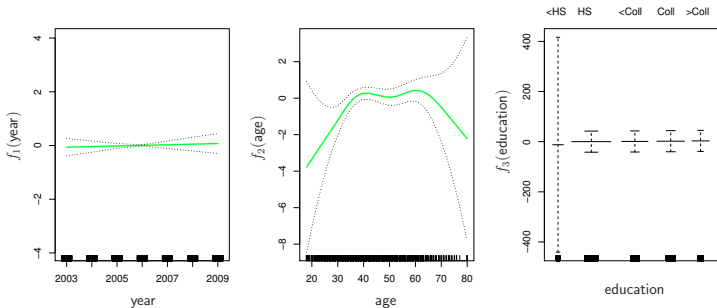
Recall: We can use logistic regression for estimating our conditional probabilities.

GAMS very naturally apply over to this:

$$\log \frac{P(Y = 1 | X)}{P(Y = 0 | X)} = \beta_0 + f_1(X_1) + \dots + f_p(X_p).$$

The fitting algorithm is a version of backfitting, but we won't discuss the details.

Example: Classification for Wage>250



- ▶ year: linear
- ▶ age: smoothing spline with $df=5$
- ▶ education: step function

n.b. The confidence interval for <HS is very large



[1] ISL. Chapter 6.4-7

[2] ESL. Chapter 5