Advanced Machine Learning

Week 4

Moving Beyond Linearity

– Splines

– GAMs

Instructor: Dr. Muhammad Fahim

Contents

- Moving beyond linearity
    - Linear models
    - Polynomial regression
    - Step functions
    - Splines
    - Generalized additive models (GAMs)
- Summary

Linear Models

- Linear models are relatively simple to describe and implement

- They have advantages over other approaches in terms of interpretation and inference

Linear Models

- Limitation of standard linear regression (in terms of predictive power)
  - Because the linearity assumption is almost always an approximation and sometimes a poor one

- **Solution**
  - We can improve upon least squares using ridge regression, the lasso, principal components regression, and other techniques

The improvement is obtained by reducing the complexity of linear model

Linear Models

- We relax the linearity assumption while still attempting to maintain as much interpretability as possible.

- We do this by examining very simple extensions of linear models like
  - Polynomial regression
  - Step functions
  - Splines
  - Local regression
  - Generalized additive models.

Polynomial Regression

Standard Linear Model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

- Extend linear regression to settings in which the relationship between the predictors and the response is non linear

Polynomial Regression

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \ldots + \beta_d x_i^d + \epsilon_i$$

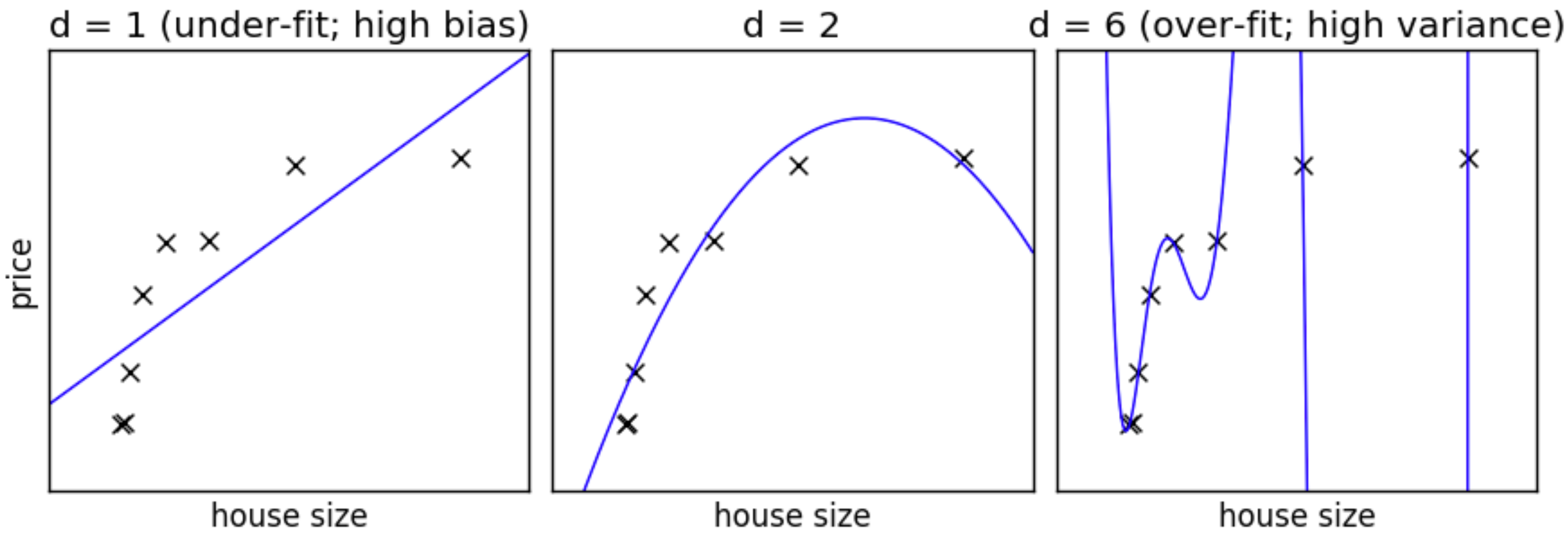
Polynomial Regression

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \ldots + \beta_d x_i^d + \epsilon_i$$

- For large enough degree d, a polynomial regression allows us to produce an extremely non-linear curve
  - As d increases, this can produce some really weird shapes

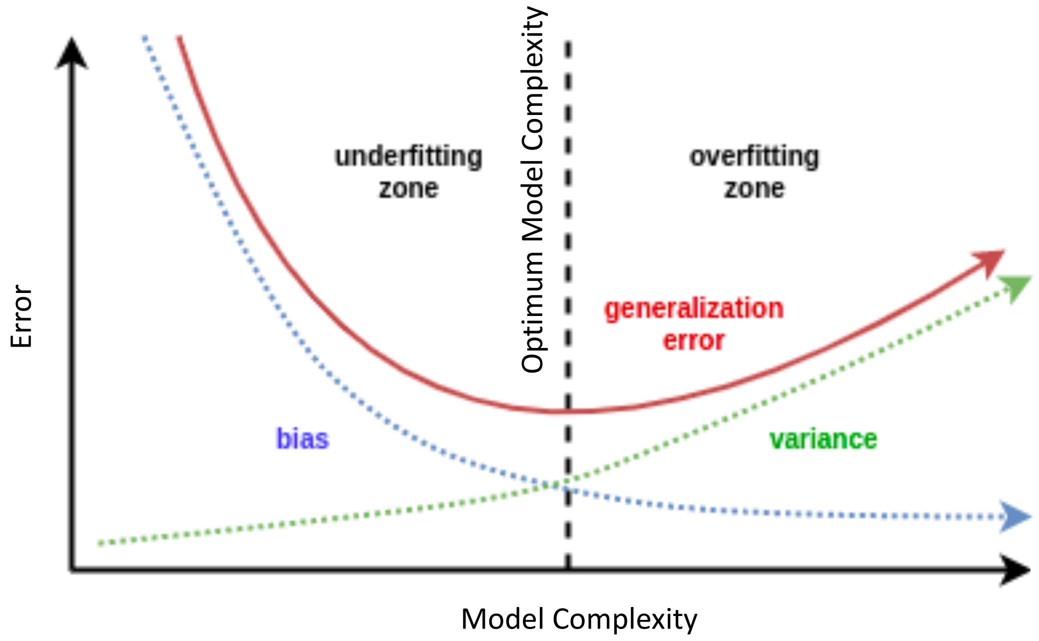
- **Question:** What's happening in terms of bias vs. variance?

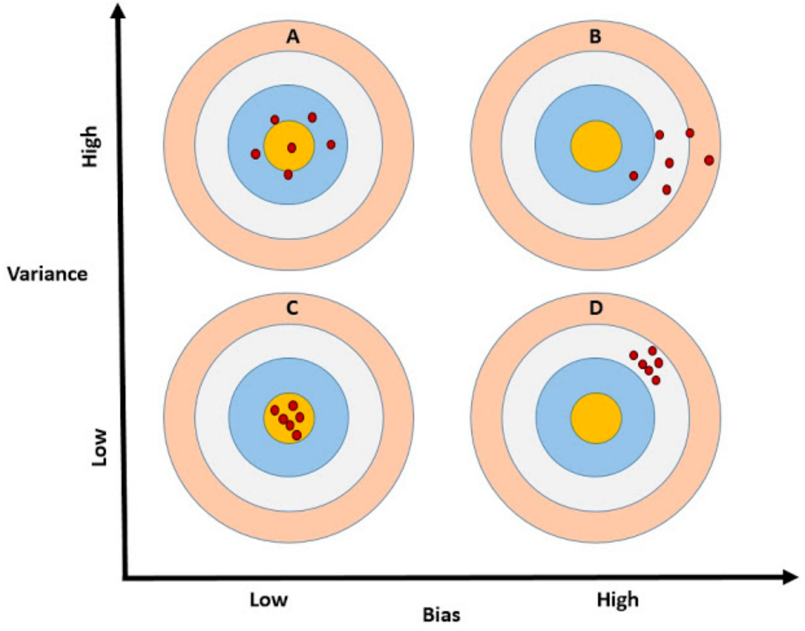
Polynomial Regression – Model Complexity



Source: http://scipy-lectures.org/packages/scikit-learn/auto_examples/plot_bias_variance.html

Polynomial Regression – Model Complexity





Source: ttps://djsaunde.wordpress.com/2017/07/17/the-bias-variance-tradeoff/
http://www.askanalytics.in/2016/11/the-concepts-of-bagging-and-boosting.html

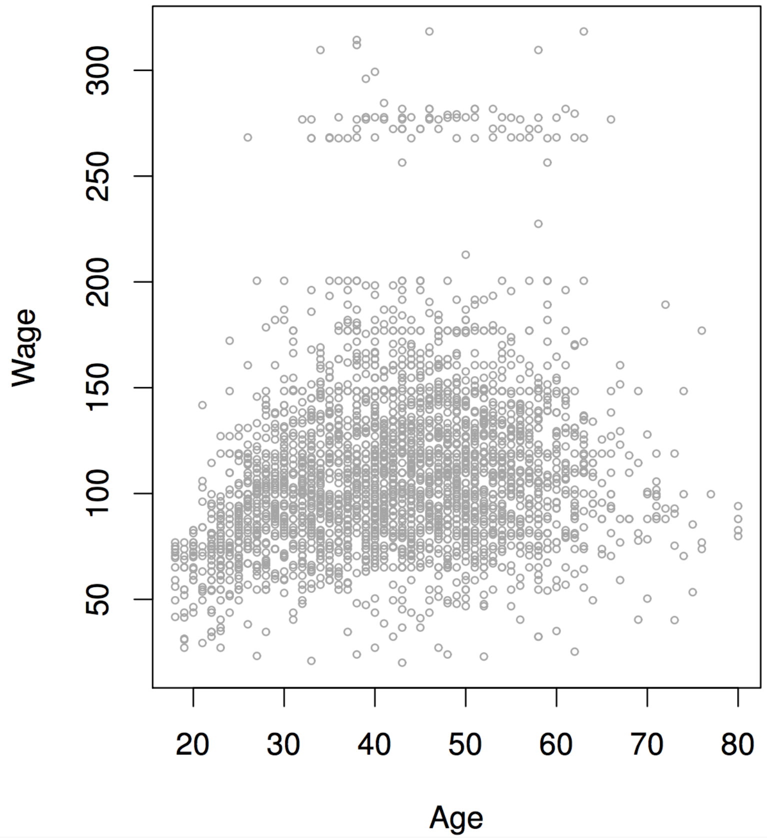
Global Structure in Polynomial Regression

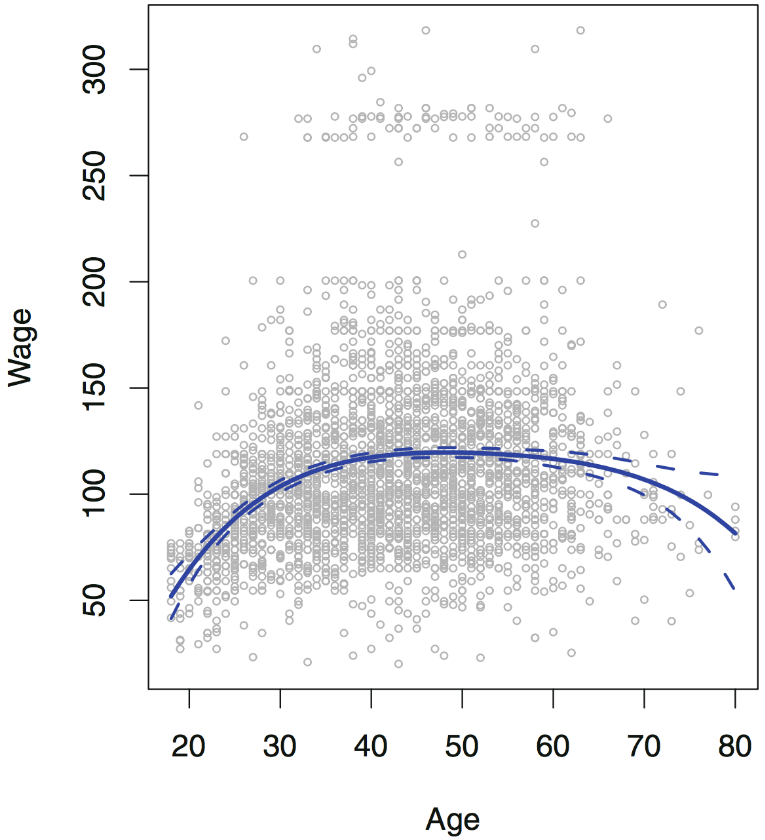
- Polynomial regression gives us added flexibility, but imposes global structure on the non-linear function of X

- **Question:** What's the problem with this?
- **Answer:** When data behave differently in different parts of the domain, function can to get really complicated

Polynomial Regression

- Wage Dataset: It contains income and demographic information for males who reside in the central Atlantic region of the United States





Results of fitting a degree-4 polynomial using least squares (solid blue curve).

Step Functions

Step Functions

Idea: if our data exhibits different behavior in different parts, we can fit a separate "mini-model" on each piece and then glue them together to describe the whole

Step Functions

- **Process**
  1. Create k cutpoints c_1, c_2,…,c_K in the range of X, and
  2. then construct K + 1 new variables (a.k.a dummy variables)

$$\begin{aligned}
C_0(X) &= I(X < c_1), \\
C_1(X) &= I(c_1 \leq X < c_2), \\
C_2(X) &= I(c_2 \leq X < c_3), \\
&\vdots \\
C_{K-1}(X) &= I(c_{K-1} \leq X < c_K), \\
C_K(X) &= I(c_K \leq X),
\end{aligned}$$

- Where I(·) is an *indicator function* that returns a 1 if the condition is true, indicator and returns a 0 otherwise

Notice that for any value of X:

$C_0(X) + C_1(X) + \ldots + C_K(X) = 1$

Since X must be in exactly one of the K + 1 intervals.

Step Functions

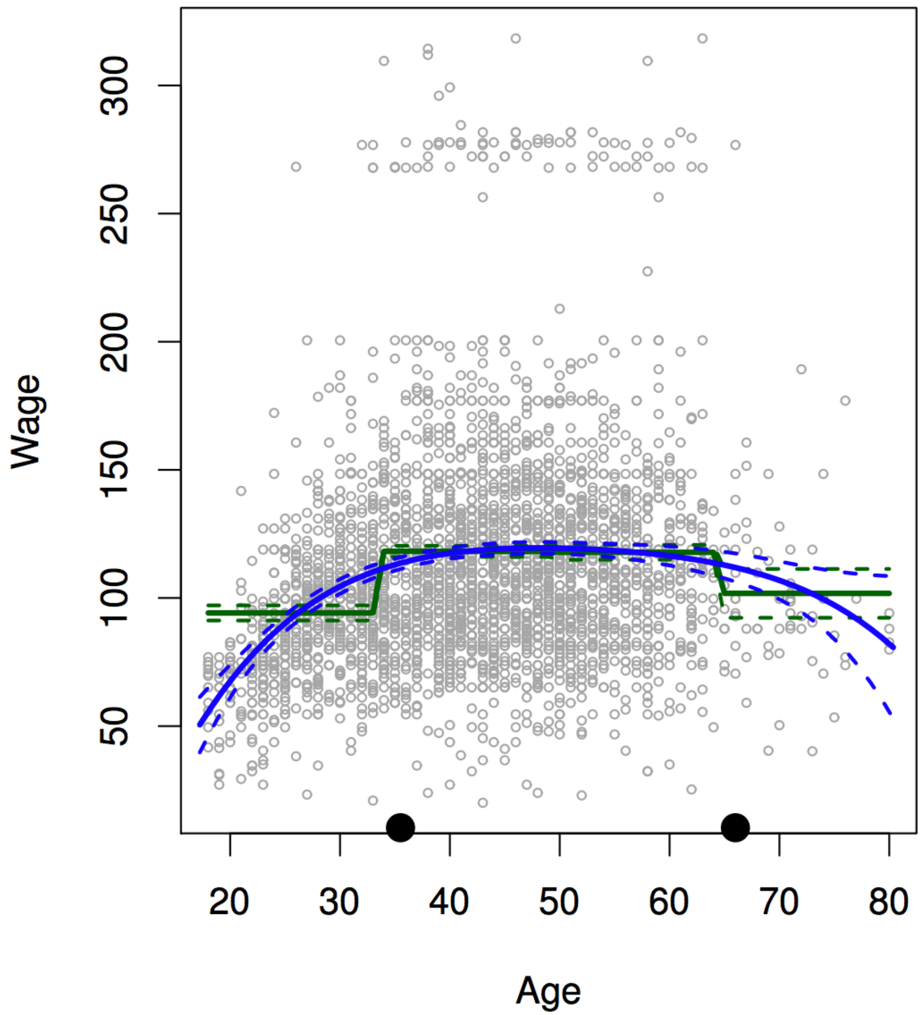
- We then use least squares to fit a linear model using $C_1(X)$, $C_2(X)$,…,$C_K(X)$ as predictors

$$y_i = \beta_0 + \beta_1 C_1(x_i) + \beta_2 C_2(x_i) + \ldots + \beta_K C_K(x_i) + \epsilon_i$$

$$\begin{array}{lcl} C_0(X) & = & I(X < c_1), \\ C_1(X) & = & I(c_1 \leq X < c_2), \\ C_2(X) & = & I(c_2 \leq X < c_3), \\ & \vdots & \\ C_{K-1}(X) & = & I(c_{K-1} \leq X < c_K), \\ C_K(X) & = & I(c_K \leq X), \end{array}$$

- **NOTE:** We exclude $C_0(X)$ as a predictor in above equation because it is redundant with the intercept

Step Functions



Granularity in Step Functions

- Step functions give us added flexibility by letting us model different parts of X independently

- **Question:** what’s the problem with this?
- **Answer:** if our data doesn’t have natural breaks, choosing the wrong step size might mean that we “miss the action”

Splines

Basis Functions

- In order to understand we need to learn about the basis functions.

- Polynomial and piecewise-constant regression models are in fact special cases of a basis function approach.

- The idea is to have at hand a family of functions or transformations that can be applied to a variable X:

$$b_1(X), b_2(X), \ldots, b_K(X)$$

Basis Functions

- Instead of fitting a linear model in X, we fit the model

$$y_i = \beta_0 + \beta_1 b_1(x_i) + \beta_2 b_2(x_i) + \beta_3 b_3(x_i) + \ldots + \beta_K b_K(x_i) + \epsilon_i$$

- **Note:** The basis functions $b_1(\cdot)$, $b_2(\cdot)$,…,$b_K(\cdot)$ are fixed and known

Basis Functions

- A standard linear model with predictors

$$y_i = \beta_0 + \beta_1 b_1(x_i) + \beta_2 b_2(x_i) + \beta_3 b_3(x_i) + \ldots + \beta_K b_K(x_i) + \epsilon_i$$

  - The basis functions for polynomial regression

$$b_j(x_i) = x_i^j$$

  - The basis functions for piecewise constant functions

$$b_j(x_i) = I(c_j \leq x_i < c_{j+1})$$

Hence, we can use least squares to estimate the unknown regression coefficients

Regression Splines

- A flexible class of basis functions that extends upon the polynomial regression and piecewise constant regression approaches

- To understand regression splines, we need to understand Piecewise Polynomials

Piecewise Polynomials Regression

Piecewise Polynomials Regression

Idea: It involves fitting separate low-degree polynomials over different regions of X

Piecewise Polynomials Regression

- **For Example:** A piecewise cubic polynomial works by fitting a cubic regression model of the form:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \epsilon_i$$

- A piecewise cubic polynomial with a single *knot at a point *c* takes the form

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

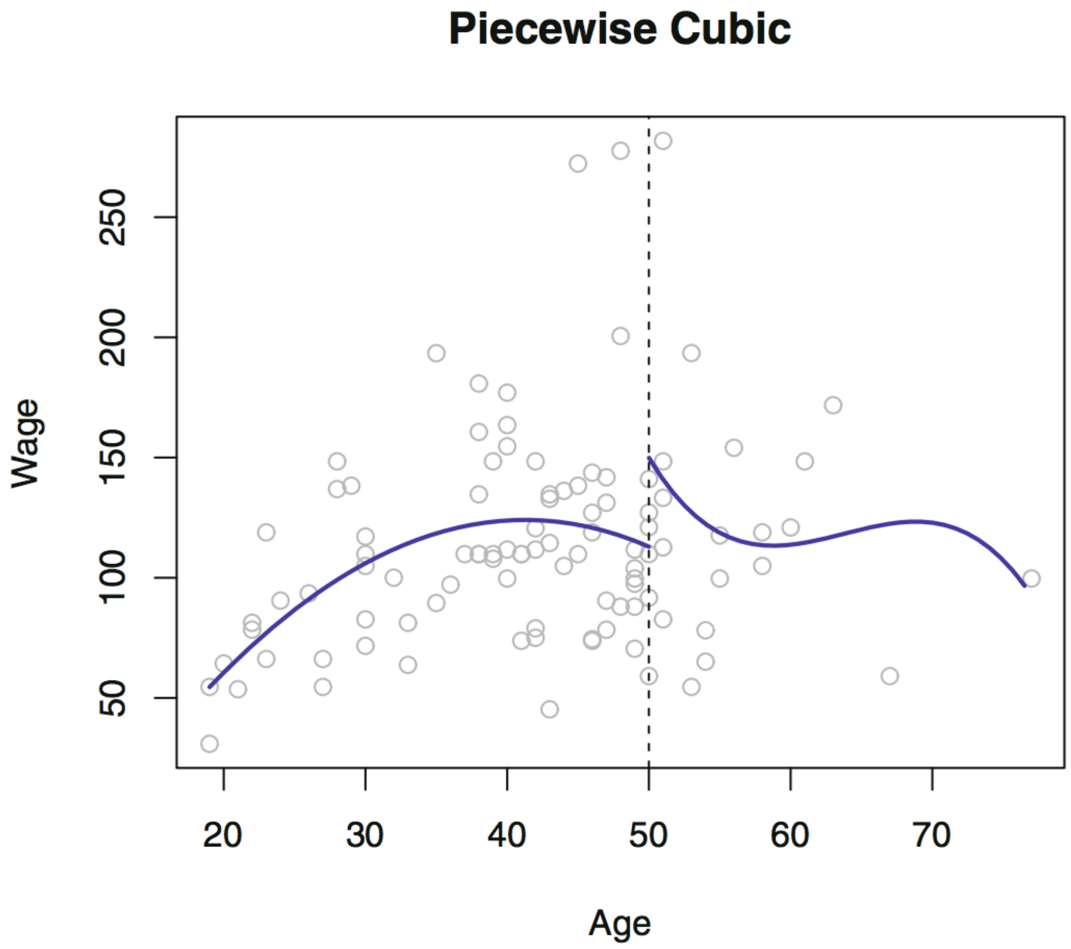
Question: If we have k different knots over the range of X then how many cubic polynomial will be fitted:

Answer: K + 1 different cubic polynomials

*Points where coefficients change = “knots”

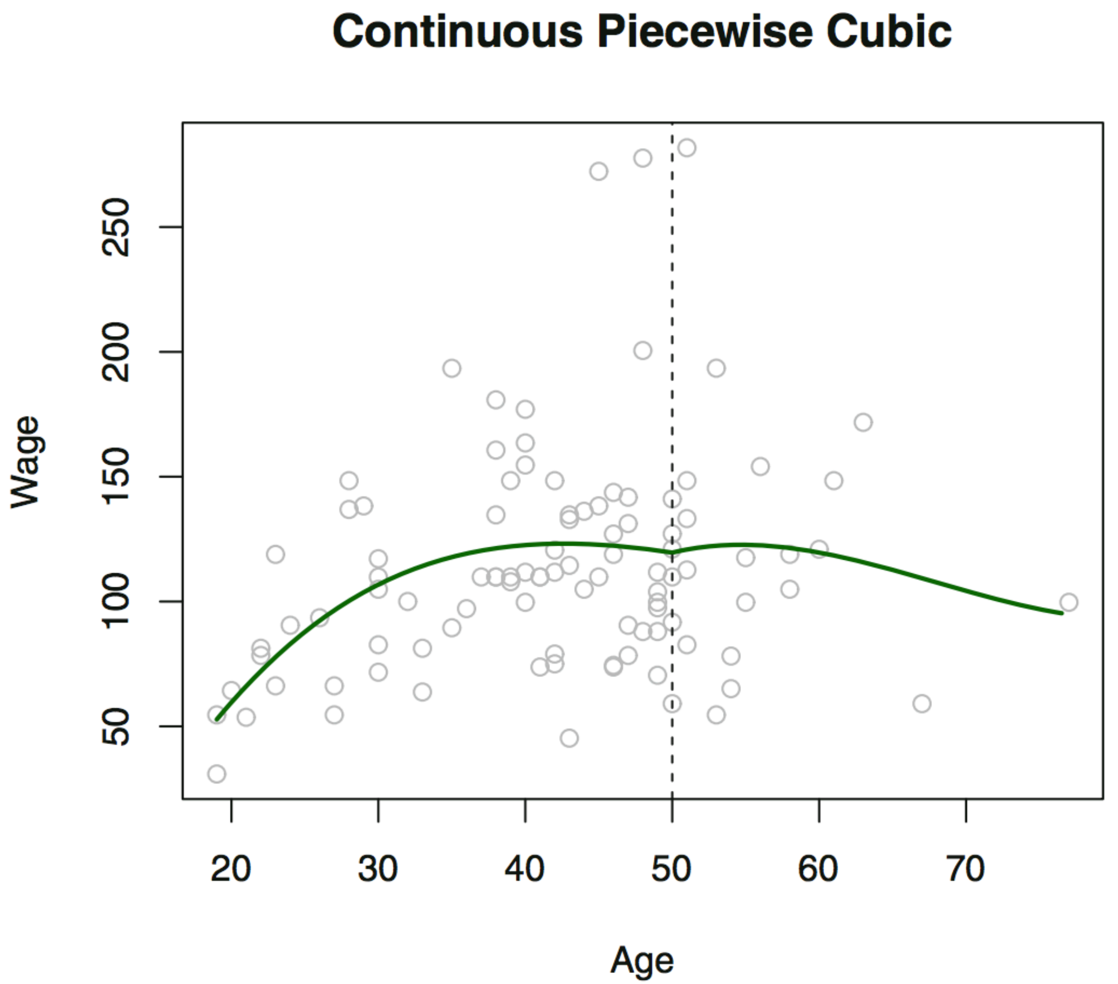
Piecewise Polynomials Regression

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



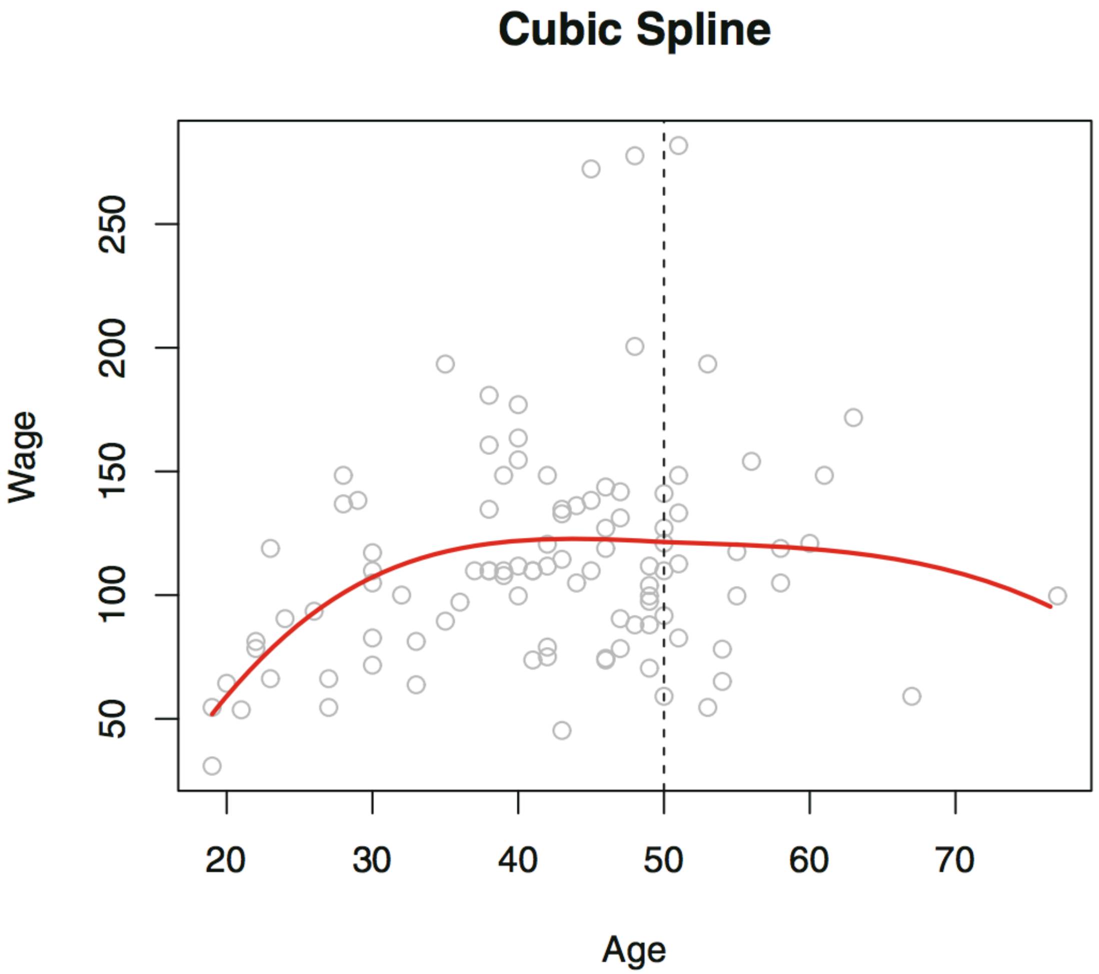
The cubic polynomials are unconstrained (knot at age 50).

Piecewise Polynomials Regression



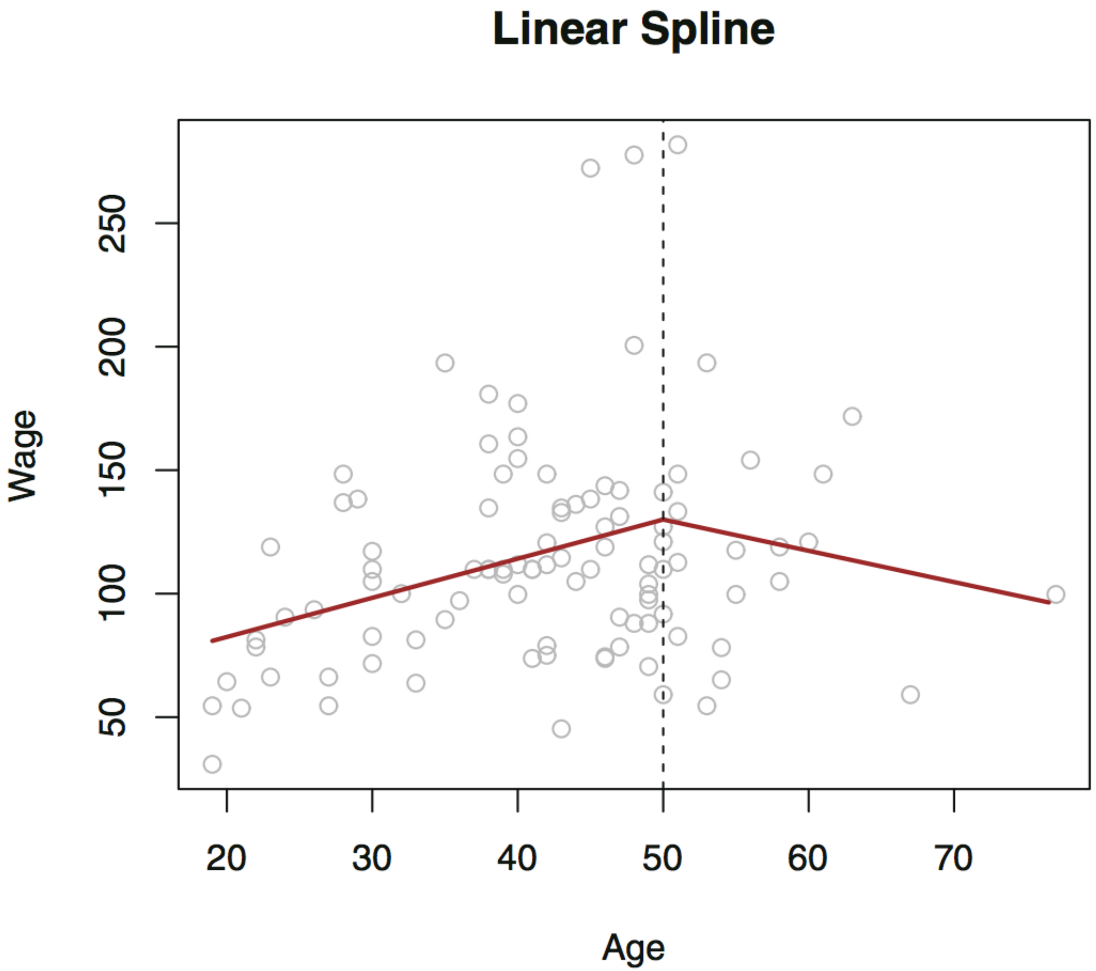
The cubic polynomials are constrained to be continuous at age=50.

Piecewise Polynomials Regression



The cubic polynomials are constrained to be continuous, and to have continuous first and second derivatives.

Piecewise Polynomials Regression



A linear spline is shown, which is constrained to be continuous

Degrees of freedom vs. constraints

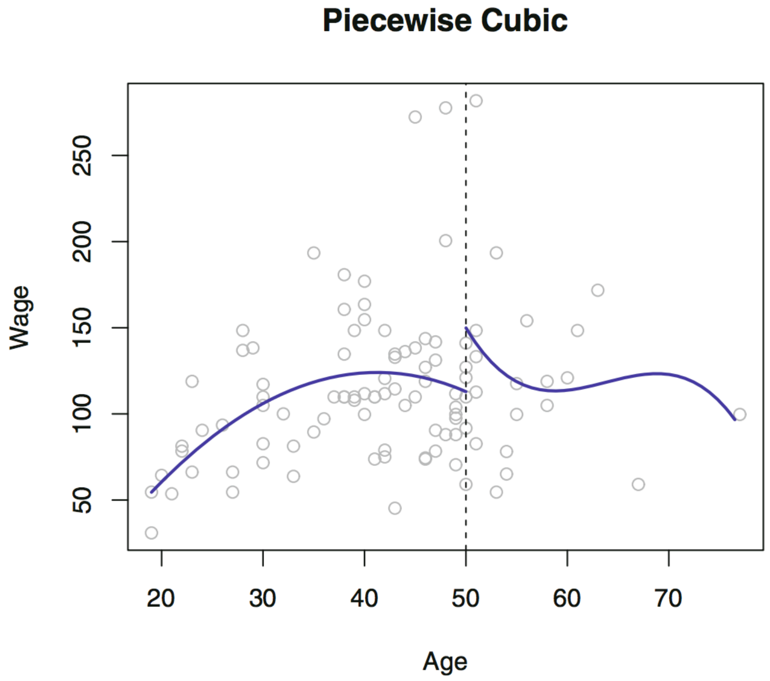
- In our piecewise cubic function with one knot, we had 8 degrees of freedom:

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

We can add constraints to remove degrees of freedom:

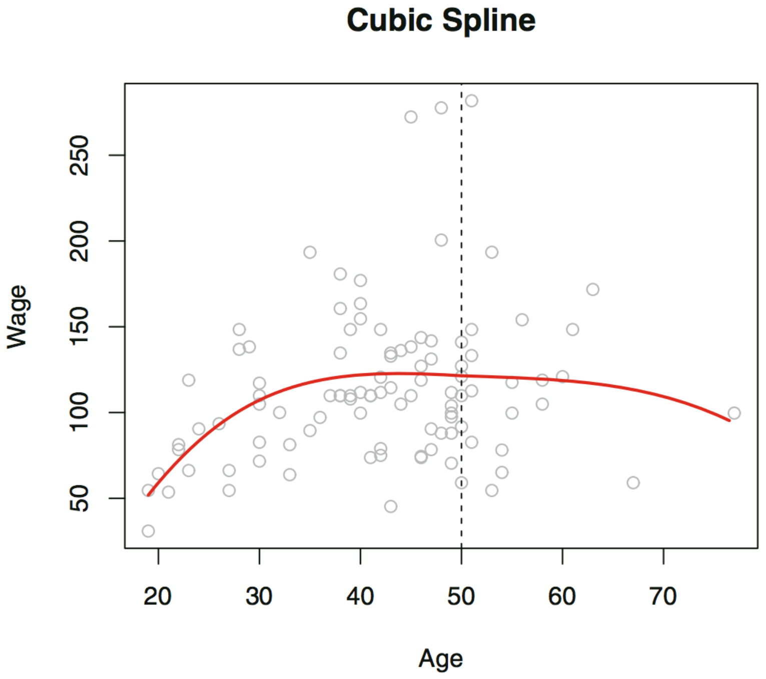
1. Function must be continuous
2. Function must have continuous 1st derivative (slope)
3. Function must have continuous 2nd derivative (curvature)

Degrees of freedom vs. constraints



- We are using eight degrees of freedom

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



- We imposed three constraints
  - (continuity, continuity of the first derivative, and continuity of the second derivative)
- So we left with five degrees of freedom
- **In general:** a cubic spline with K knots uses cubic spline a total of 4 + K degrees of freedom

The Spline Basis Representation

- How can we fit a piecewise degree-d polynomial under the constraints
- Constrains are:
  - Continuity
  - possibly its first d − 1 derivatives be continuous?

It seemed somewhat complex

- **Solution**
  - It turns out that we can use the basis model

The Spline Basis Representation

- A *cubic spline with K knots can be modeled as

$$y_i = \beta_0 + \beta_1 b_1(x_i) + \beta_2 b_2(x_i) + \cdots + \beta_{K+3} b_{K+3}(x_i) + \epsilon_i$$

- We need to choose an appropriate choice of basis functions b_1, b_2,...,b_{K+3} and model can be fit using least squares.

*Cubic splines are popular because most human eyes cannot detect the discontinuity at the knots

The Spline Basis Representation

- One common approach is to start with a standard basis for a cubic polynomial is x, x^2, x^3 and then add one *truncated power basis function* per knot ξ:

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

- Where ξ is the knot.

Explanation of the function:
https://support.sas.com/documentation/cdl/en/statug/63347/HTML/default/viewer.htm#statug_introcom_a0000003010.htm

The Spline Basis Representation

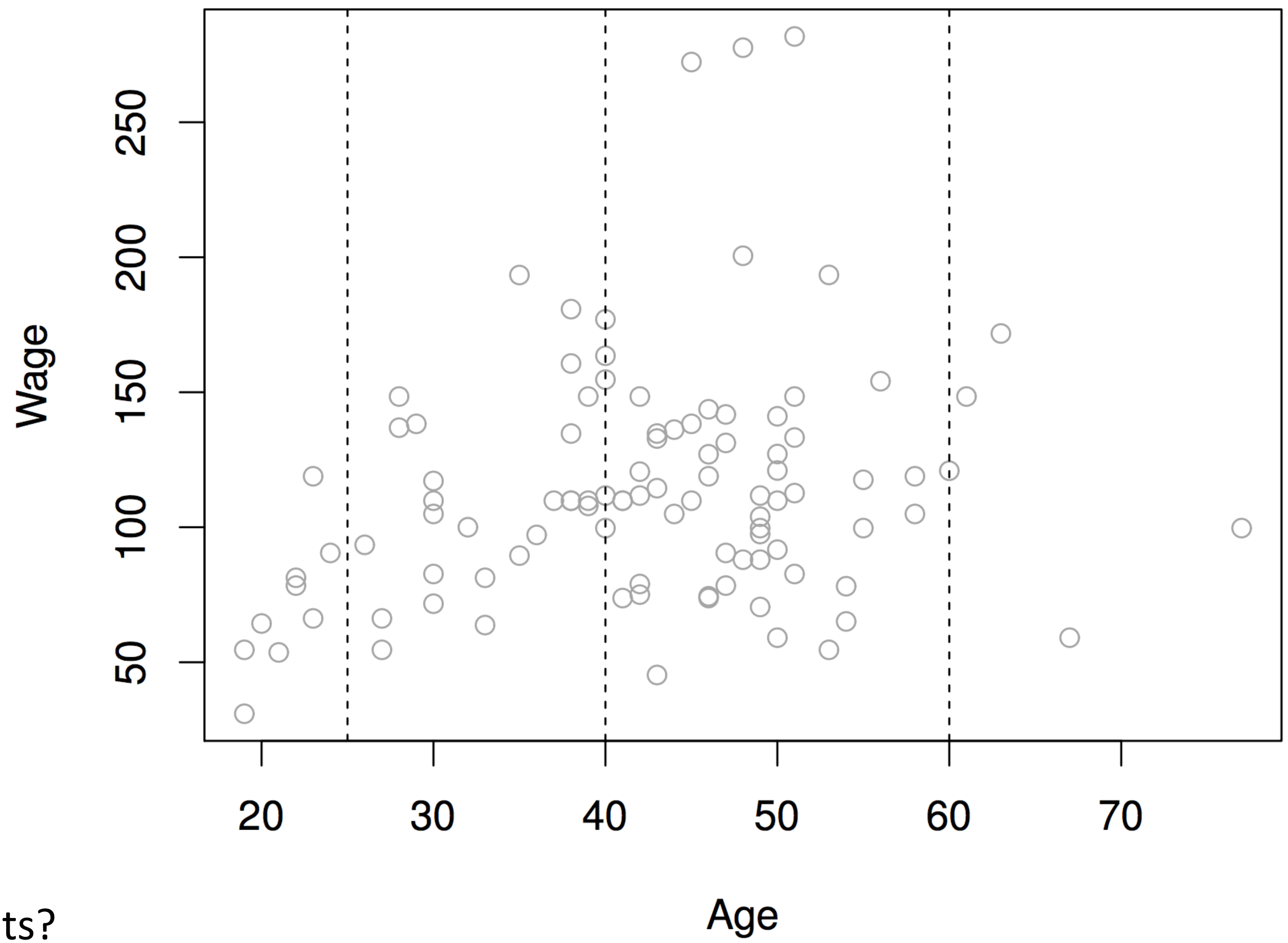
- In other words, in order to fit a cubic spline to a data set with K knots

- We perform least squares regression with an intercept and 3 + K predictors, of the form

$$X, X^2, X^3, h(X, \xi_1), h(X, \xi_2), \ldots, h(X, \xi_K)$$

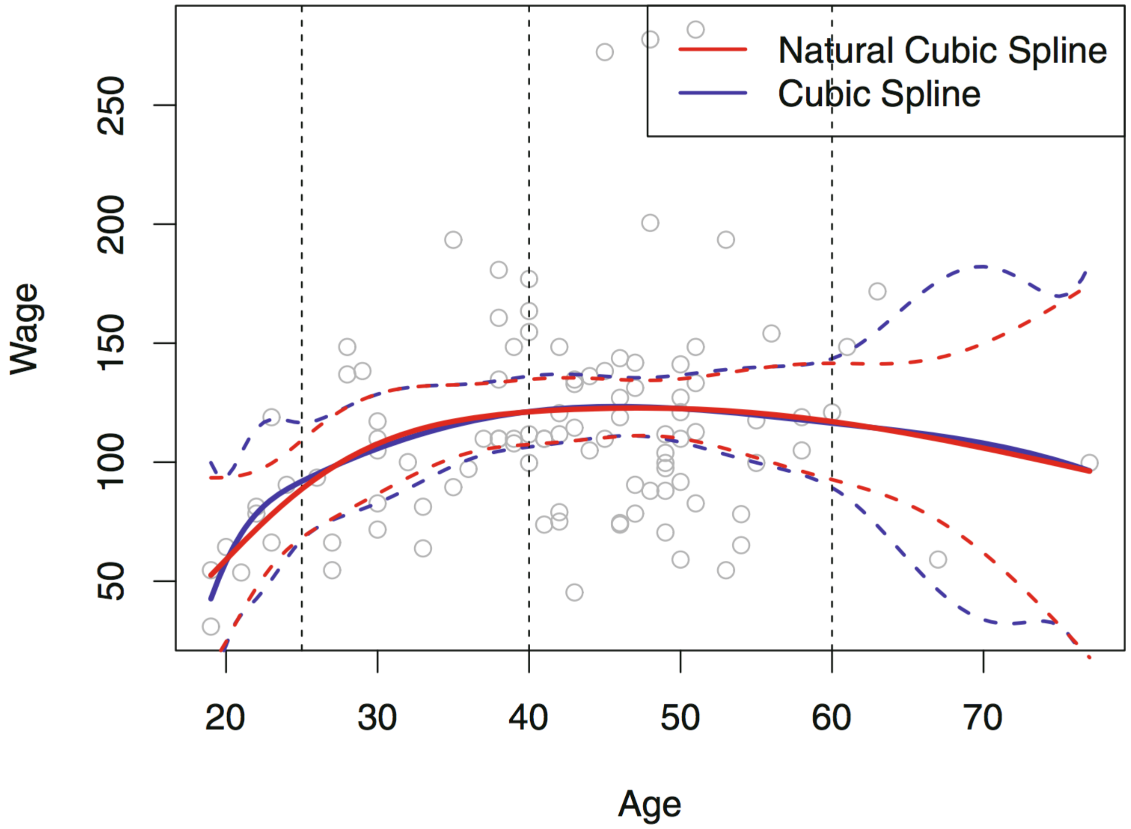
  - Where ξ_1,..., ξ_K are the knots.

Cubic Splines



- How many knots?
  - 3 knots

Cubic Splines



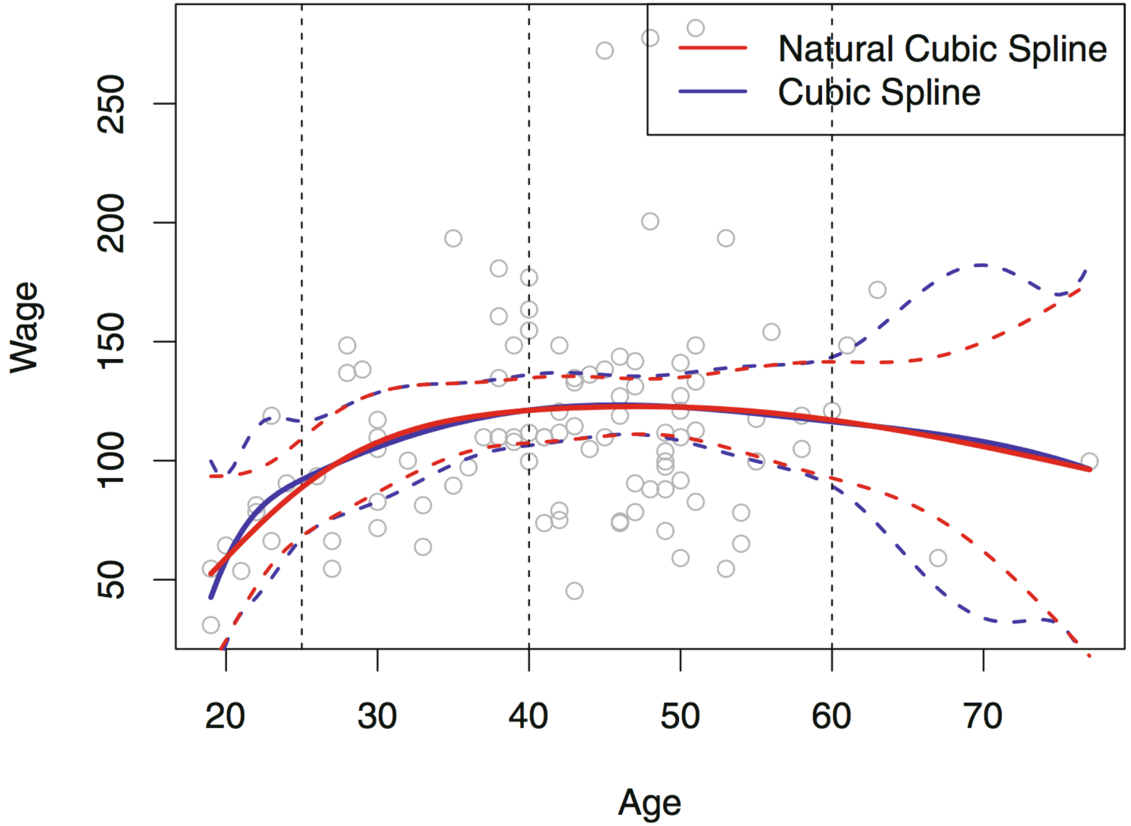
- Splines can have high variance at the outer range of the predictors
- **What to do now?**
- **Solution:** Natural Splines

A cubic spline and a natural cubic spline, with three knots, fit to a subset of the Wage data.

Natural Splines

- A natural spline is a regression spline with additional boundary constraints: the natural function is required to be linear at the boundary

- This additional constraint means that natural splines generally produce more stable estimates at the boundaries.

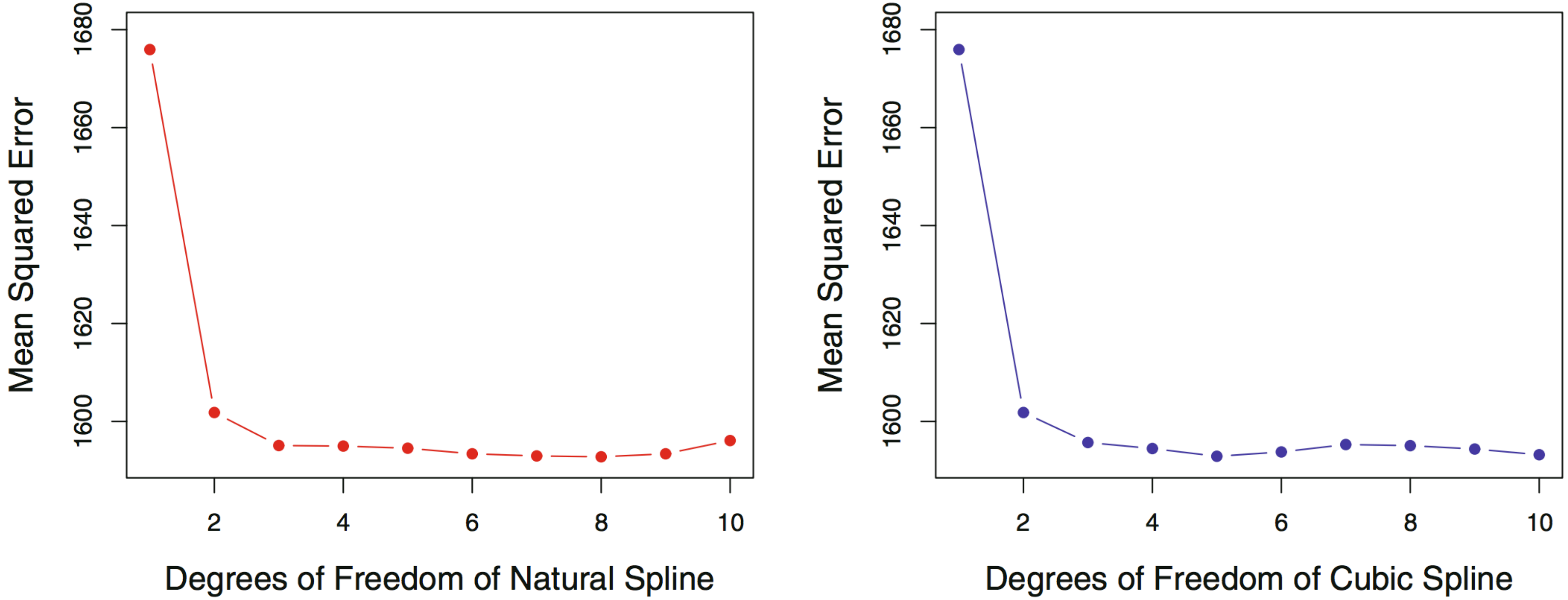


Choosing the Number and Locations of the Knots

- **How many number of Knots?**
- One option is to try out different numbers of knots and see which produces the best looking curve.
  - A somewhat more objective approach is to use cross-validation

Choosing the Number and Locations of the Knots

- **Cross validation**



Explanation: The two methods produce almost identical results, with clear evidence that a one-degree fit (a linear regression) is not adequate. Both curves flatten out quickly, and it seems that three degrees of freedom for the natural spline and four degrees of freedom for the cubic spline are quite adequate

Choosing the Number and Locations of the Knots

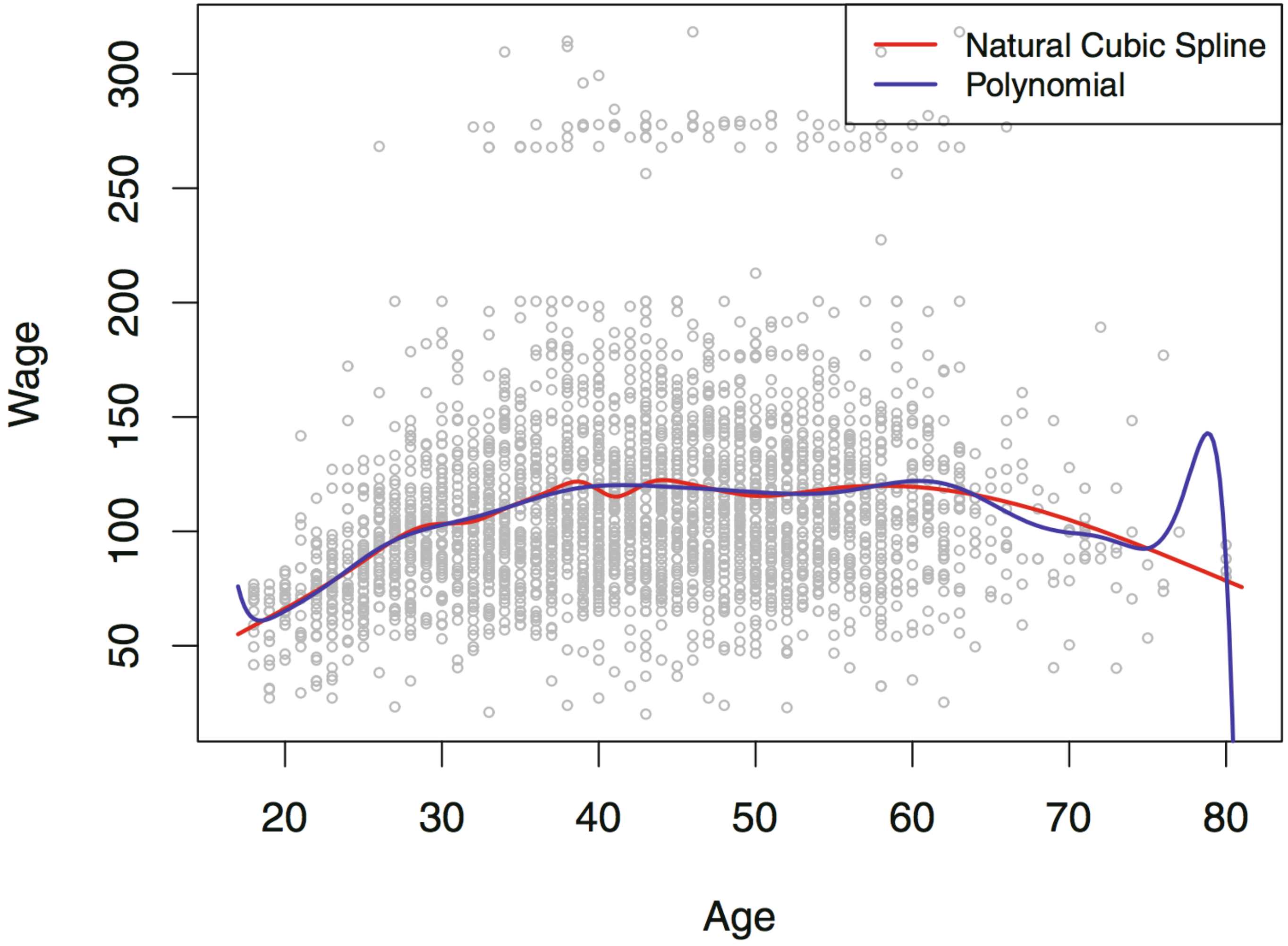
- **Where should we place the knots?**
  - Place them uniformly across the domain (common method in practice)
  - Put more knots in places where the data varies a lot
  - Place them at percentiles of interest (e.g. 25th, 50th, and 75th)

Comparison to Polynomial Regression

- Regression splines often give better results than polynomial regression

- **Reason:**
  - because they can add flexibility in places where it is needed by **adding more knots,** without having to **add more predictors**

Comparison to Polynomial Regression



Smoothing Splines

$$\sum_{i=1}^{n}(y_i - g(x_i))^2 + \lambda \int g''(t)^2 dt$$

A penalty for the roughness of the function.

Generalized Additive Models (GAMs)

- So far, we present a number of approaches for flexibly predicting a response Y on the basis of a single predictor X.
- These approaches can be seen as extensions of simple linear regression.
- In GAMs, we explore the problem of flexibly predicting Y on the basis of several predictors, $X_1,\ldots,X_p$.
- This amounts to an extension of multiple linear regression.

GAMs for Regression Problems

- A natural way to extend the multiple linear regression model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i$$

- Replace each linear component $\beta_j x_{ij}$ with a (smooth) non-linear function $f_j(x_{ij}$).

$$\begin{aligned} y_i &= \beta_0 + \sum_{j=1}^{p} f_j(x_{ij}) + \epsilon_i \\ &= \beta_0 + f_1(x_{i1}) + f_2(x_{i2}) + \cdots + f_p(x_{ip}) + \epsilon_i \end{aligned}$$

- This is an example of a GAM. It is called an additive model because we calculate a separate f_j for each X_j , and then add together all of their contributions

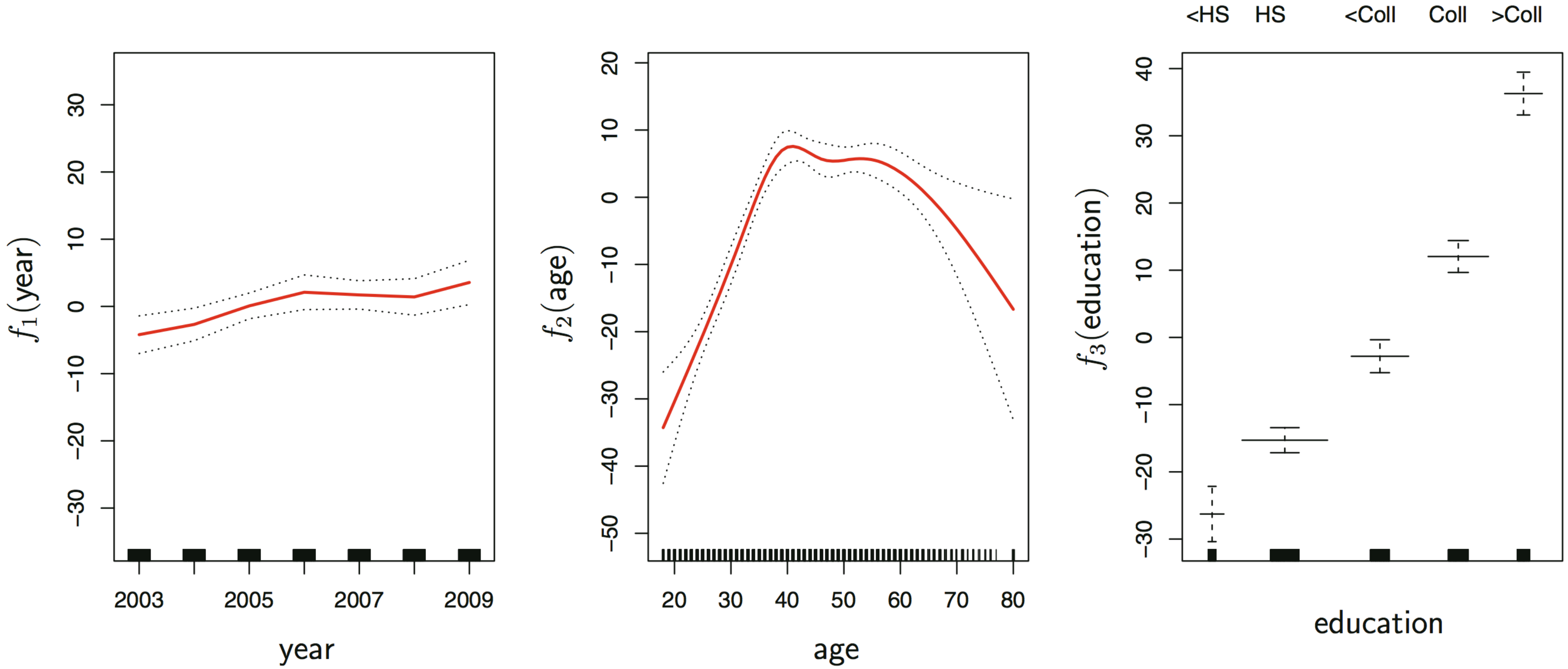
Fitting a GAM

- If the functions f have a basis representation, we can simply use least squares:
  - Natural cubic splines
  - Polynomials
  - Step functions

GAMs for Regression Problems

- For example: natural splines, and consider the task of fitting the model

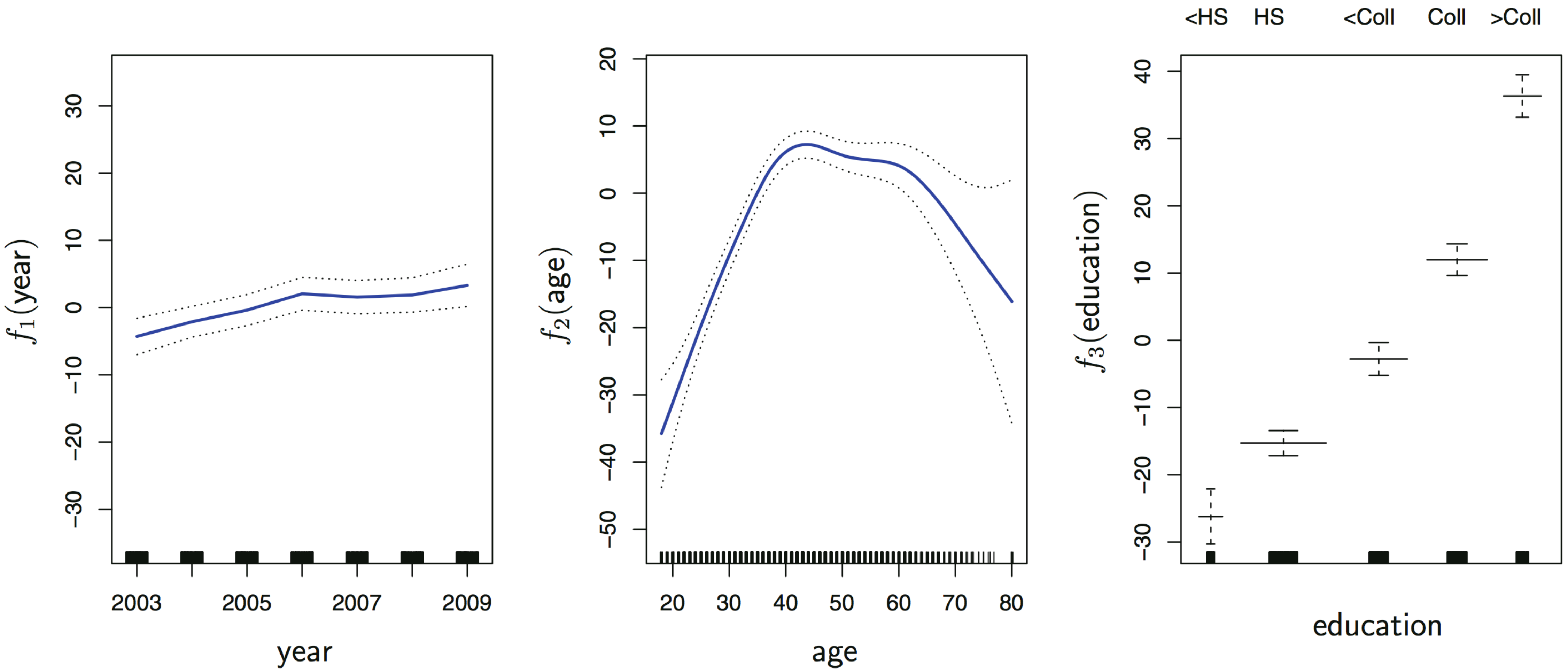
$$\texttt{wage} = \beta_0 + f_1(\texttt{year}) + f_2(\texttt{age}) + f_3(\texttt{education}) + \epsilon$$



GAM with Smoothing Spline

- In the case of smoothing splines, least squares cannot be used.
- To fit the Gam with smoothing spline is known as *backfitting.*
- *Backfitting*
  - This method fits a model involving multiple predictors by repeatedly updating the fit for each predictor in turn, holding the others fixed.
  - The beauty of this approach is that each time we update a function, we simply apply the fitting method for that variable to a partial residual

GAM with Smoothing Spline



Pros and Cons of GAMs

- **Pros**
  - GAMs allow us to fit a non-linear f_j to each X_j , so that we can automatically model non-linear relationships that standard linear regression will miss.
  - This means that we do not need to manually try out many different transformations on each variable individually
  - The non-linear fits can potentially make more accurate predictions for the response Y.
  - Because the model is additive, we can still examine the effect of each X_j on Y individually while holding all of the other variables fixed. Hence if we are interested in inference, GAMs provide a useful representation.

Pros and Cons of GAMs

- **Cons**
  - The main limitation of GAMs is that the model is restricted to be additive. With many variables, important interactions can be missed.

Reference

- This lecture is based on chapter 7 of book by Gareth James et. al. “Introduction to Statistical Learning”

Thank You ☺