Chapter 7: Moving Beyond Linarity

So far we have mainly focused on linear models.

Linear models are relatively simple to describe and implement.

+: interpret & inference

−: can have limited predictive performance because linearity assumption is alway an approximation (may not be a good one).

Previously, we have seen we can improve upon least squares using ridge regression, the lasso, principal components regression, and more.

improvement obtained by reducing complexity of linear models ⟹ lower variance of estimates.

still a linear model! Can only be improved so much.

Through simple and more sophisticated extensions of the linear model, we can relax the linearity assumption while still maintiaining as much interpretability as possible.

We've already seen this.

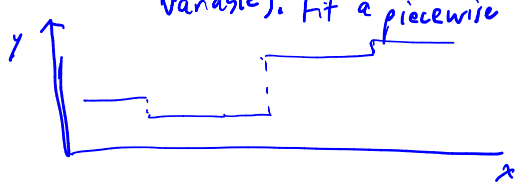
1. Polynomial regression: adding extra predictors that are original variables raised to a power

   e.g. cubic regression use X, X^2, X^3 as predictors, e.g. $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \varepsilon$

   +: non-linear fit

   −: with large powers, polynomial can take very strange shapes (especially at boundary).

2. Step functions: cut the range of predictor into K distinct regions (to producing categorical variable). Fit a piecewise constant function to (binned) X.



3. Regression Splines: more flexible than polynomials & step functions (extends both)

   idea: cut range of X into K distinct regions & polynomial is fit within each region.

   polynomials constrained so they smoothly joined.

4. Generalized additive models: extend above ideas to deal w/ multiple predictors.

Note: We can talk about regression or classification, e.g. logistic regression (polynomial): $P(Y=1|X) = \dfrac{\exp(\beta_0+\beta_1 X + \dots + \beta_d X^d)}{1+\exp(\beta_0+\beta_1 X+\dots+\beta_d X^d)}$

1 Step Functions

Using polynomial functions of the features as predictors imposes a *global* structure on the non-linear function of X.

We can instead use *step-functions* to avoid imposing a global structure.

idea: Break range of X into bins and fit different constant to each bin.

detail: ① create cut points $c_1, \dots, c_K$ in the range of X

② Construct $K+1$ new variables.

$$
\left.
\begin{aligned}
C_0(X) &= \mathbb{I}(X < c_1) \\
C_1(X) &= \mathbb{I}(c_1 \le X < c_2) \\
&\vdots \\
C_{K-1}(X) &= \mathbb{I}(c_{K-1} \le X < c_K) \\
C_K(X) &= \mathbb{I}(c_K \le X)
\end{aligned}
\right\}
$$

indicator variable $C_0(X) = \mathbb{I}(X < c_1) = \begin{cases} 1 & X < c_1 \\ 0 & \text{o.w.} \end{cases}$

"dummy variables"

③ Use least squares to fit a linear model using $C_1(X), C_2(X), \dots, C_K(X)$

$$Y = \beta_0 + \beta_1 C_1(X) + \dots + \beta_K C_K(X) + \varepsilon$$

leave out $C_0(X)$ because it is equivalent to including an intercept.

For a given value of X, at most one of $C_1, \dots, C_K$ can be non-zero.

$C_0(X) + C_1(X) + \dots + C_K(X) = 1$ since X must be in exactly one interval.

when $X < c_1 \Rightarrow$ all of predictors $C_1, \dots, C_K = 0$

$\Rightarrow \beta_0$ interpreted as the mean value of Y when $X < c_1$

β_j represent the average increase in the response for $c_j \le X < c_{j+1}$ relative to $X < c_1$.

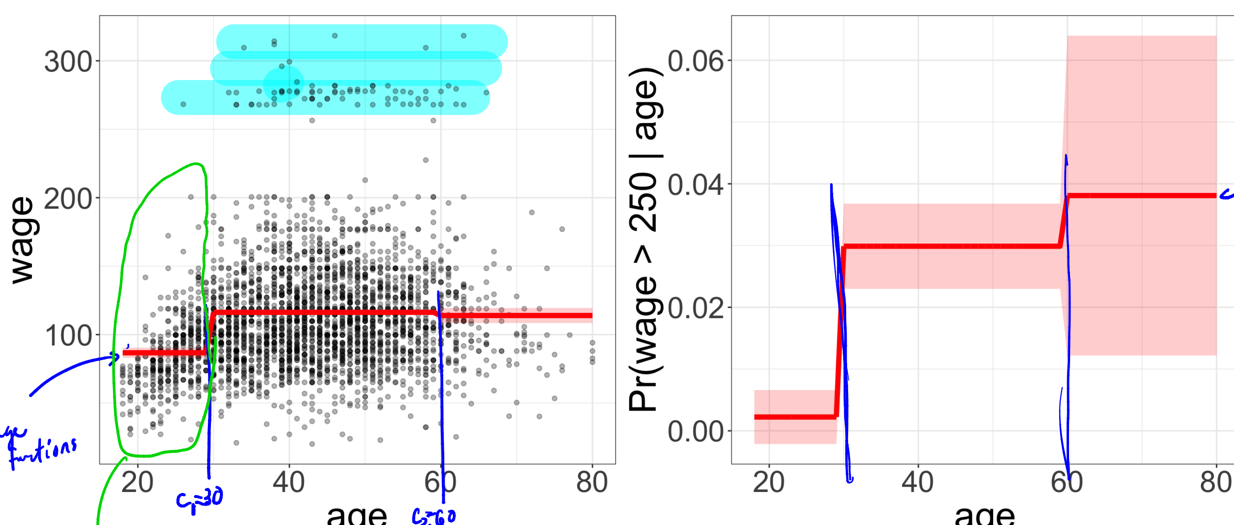
We can also fit a logistic regression model for classification:

$$P(Y=1|X) = \frac{\exp(\beta_0 + \beta_1 C_1(X) + \dots + \beta_K C_K(X))}{1 + \exp(\beta_0 + \beta_1 C_1(X) + \dots + \beta_K C_K(X))}$$

Example: Wage data. for a group of 3000 male workers in mid-atlantic region

year	age	maritl	race	education	region	jobclass	health	health_ins	logwage	wage
2006	18	1. Never Married	1. White	1. < HS Grad	2. Middle Atlantic	1. Industrial	1. <=Good	2. No	4.318063	75.04315
2004	24	1. Never Married	1. White	4. College Grad	2. Middle Atlantic	2. Information	2. >=Very Good	2. No	4.255273	70.47602
2003	45	2. Married	1. White	3. Some College	2. Middle Atlantic	1. Industrial	1. <=Good	1. Yes	4.875061	130.98218
2003	43	2. Married	3. Asian	4. College Grad	2. Middle Atlantic	2. Information	2. >=Very Good	1. Yes	5.041393	154.68529

$c_1 = 30$
$c_2 = 60$



fitted value of wage using step functions of age.

missing clear upward trend.

fitted model.

logistic regression modeling probability of being high wage earner given age. (wage ≥ 250k). stepwise linear model. w/ knots at $x = 30, 60$.

Unless there are natural breakpoints in the predictor, piecewise constant can miss trends.

2 Basis Functions

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

Idea:

have a family of functions or transformations that can be applied to a variable X

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

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

$$y_i = \beta_0 + \beta_1 b_1(x_i) + \dots + \beta_K b_K(x_i) + \varepsilon_i$$

Note that the basis functions are fixed and known. we choose them ahead of time.

ex: polynomial regression $\quad b_j(x_i) = x_i^j, \quad j = 1, \dots, d.$

ex: step function (piecewise constant)

$$b_j(x_i) = \mathbb{I}(c_j \le x_i < c_{j+1}) = \begin{cases} 1 & c_j \le x_i < c_{j+1} \\ 0 & \text{o.w.} \end{cases}$$

We can think of this model as a standard linear model with predictors defined by the basis functions and use least squares to estimate the unknown regression coefficients.

$\Rightarrow$ we can use all our inference tools for linear models.

e.g. $se(\hat{\beta}_j)$ and F-statistic for model significance.

Many alternatives exist for basis functions.

e.g. wavelets, fourier series, regression splines (next).

3 Regression Splines

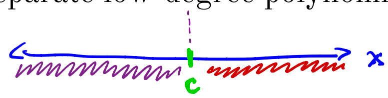
Regression splines are a very common choice for basis function because they are quite flexible, but still interpretable. Regression splines extend upon polynomial regression and piecewise constant approaches seen previously.

start

3.1 Piecewise Polynomials

Instead of fitting a high degree polynomial over the entire range of X, piecewise polynomial regression involves fitting separate low-degree polynomials over different regions of X.

e.g. one knot at c



fit two polynomials to the data

one on subset for $x < c$

one on subset for $x \geq c$

each polynomial can be fit using least squares.

For example, a piecewise cubic with no knots is just a standard cubic polynomial.

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

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

Using more knots leads to a more flexible piecewise polynomial.

If we place K knots $\Rightarrow$ fit $K+1$ polynomials

In general, we place K knots throughout the range of X and fit $K + 1$ polynomial regression models.

This leads to $(d+1)(K+1)$ degrees of freedom in model

(# of parameters to fit $\approx$ complexity / flexibility).

3.2 Constraints and Splines

To avoid having too much flexibility, we can *constrain* the piecewise polynomial so that the fitted curve must be continuous.

i.e. there cannot be a jump at knots.

To go further, we could add two more constraints

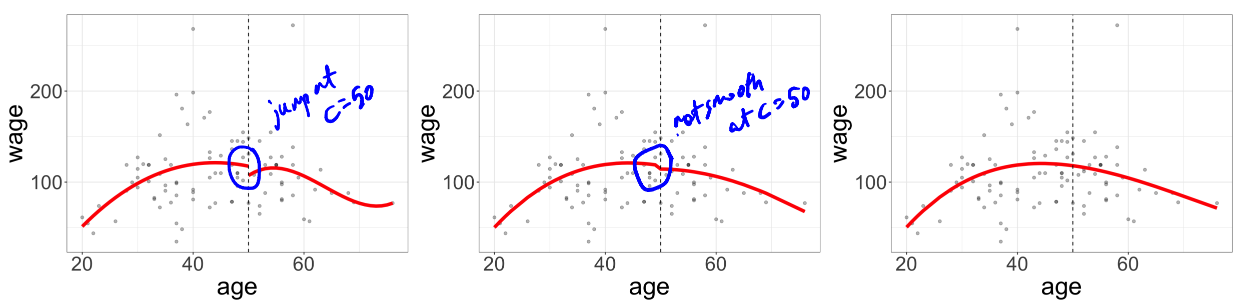
1. 1st derivative of piecewise polynomial must be continuous
2. 2nd derivative of piecewise polynomial must be cts.

In other words, we are requiring the piecewise polynomials to be *smooth*.

Each constraint that we impose on the piecewise cubic polynomials effectively frees up one degree of freedom, by reducing the complexity of the resulting fit.

The cubic fit with continuity and 2 smoothness contraints is called a cubic *spline*.

A degree-d spline is a piecewise degree-d polynomial w/ continuity in derivatives up to degree $d-1$ at each knot.



piecewise cubic polynomial

piecewise cubic polynomial w/ continuity

cubic spline cts & smooth

3.3 Spline Basis Representation

Fitting the spline regression model is more complex than the piecewise polynomial regression. We need to fit a degree d piecewise polynomial and also constrain it and it's upto $d-1$ derivatives to be continuous at the knots.

We can use the basis model to represent a regression spline.

eg. cubic spline w/ K knots

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

for appropriate basis functions $b_1, b_2, \dots, b_{K+3}$ $\qquad x, x^2, x^3$

● = d-dimensional spline.

The most direct way to represent a cubic spline is to start with the basis for a cubic polynomial and add one *truncated power basis* function per knot. → basis for d degree polynomial

$$h(x, \xi) = (x-\xi)_+^{3\,d} = \begin{cases} (x-\xi)^{3\,d} & \text{if } x > \xi \\ 0 & \text{o.w.} \end{cases} \qquad \text{where } \xi \text{ is a knot.}$$

$$\Rightarrow \quad y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \sum_{j=1}^{K} \beta_{3+j}\, h(x, \xi_j)$$

this will lead to discontinuity in only the 3rd derivative at each ξ_j with continuous first and second derivatives and continuity at each ξ_j

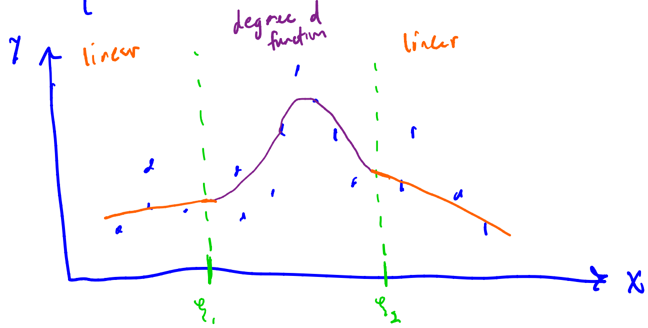
$$df = K + 4 \quad \text{(cubic spline w/ K knots)}$$

Unfortunately, splines can have high variance at the outer range of the predictors. One solution is to add *boundary contraints.*

when X is very small or very large.

⇒ "natural spline"

function required to be linear at the boundary (where X is smaller than the smallest knot and bigger than biggest knot)



additional constraint produces more stable estimates at the boundaries.

3.4 Choosing the Knots

When we fit a spline, where should we place the knots?

⇒ regression spline is most flexible in regions that contain a lot of knots (coefficients changing rapidly).

⇒ place knots where we think the function will vary rapidly and less knots where function is stable.

More common in practice: place them uniformly.

To place knots: choose desired degrees of freedom (flexibility) & use software to automatically place # knots at uniform quantiles of data

How many knots should we use?

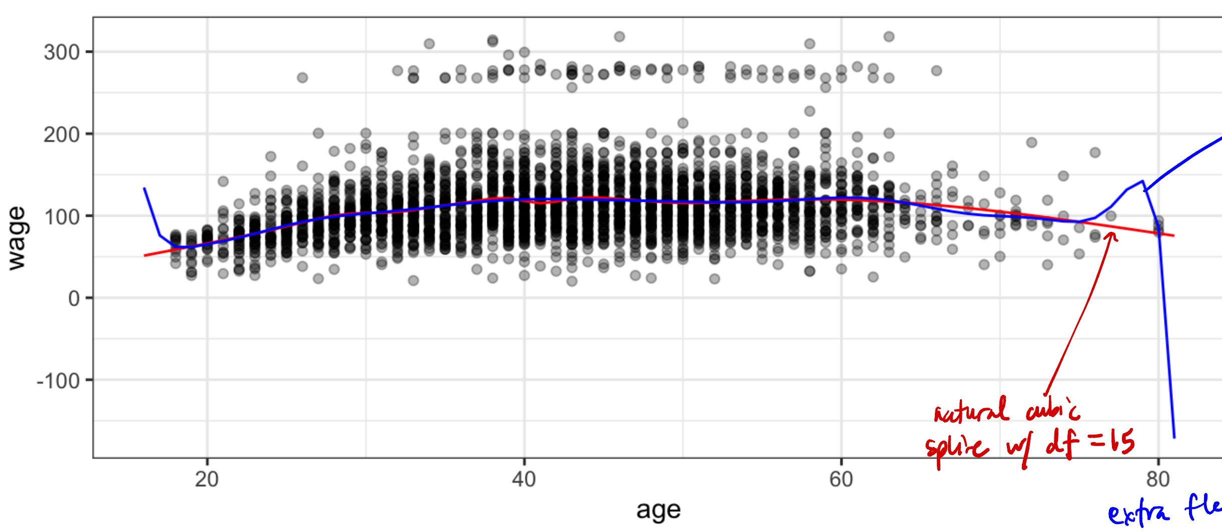
⟺ how many df should we use?

Use CV! Use K giving smallest CV MSE (CV error).

3.5 Comparison to Polynomial Regression

Regression splines often give superior results to polynomial regression.

↳ Polynomial regression must use high degrees to achieve flexibility (e.g. X^{15}), but regression splines introduce flexibility through knots (fixed degree polynomials) ⇒ more stability (esp. at boundaries)



polynomial with degree 15

extra flexibility of polynomial at boundary produces undesirable results, but NC spline w/ same flexibility still looks reasonable.

4 Generalized Additive Models

So far we have talked about flexible ways to predict Y based on a single predictor X.

These approaches can be seen as extensions of simple linear regression model.

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

extension: basis functions of X

Generalized Additive Models (GAMs) provide a general framework for extending a standard linear regression model by allowing non-linear functions of each of the variables while maintaining *additivity*.

flexibly predict Y on several predictors $X_1, \dots, X_p$.

4.1 GAMs for Regression

still additive models
can be used for regression or classification.

A natural way to extend the multiple linear regression model to allow for non-linear relationships between feature and response:

linear regression: $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_p x_{pi} + \varepsilon_i$

extension idea: replace each linear component $\beta_j x_{ij}$ with a smooth non-linear function.
non-linear

$$\Rightarrow \text{GAM: } y_i = \beta_0 + \sum_{j=1}^{p} f_j(x_{ij}) + \varepsilon_i$$

$$= \beta_0 + f_1(x_{i1}) + f_2(x_{i2}) + \dots + f_p(x_{ip}) + \varepsilon_i$$

"additive" because we calculate a separate f_j for each predictor X_j and add them together.

posibilities for f_j:

- linear component (leads to linear regression).
- polynomial function
- regression spline
- smoothing spline
- local linear regression.

] not covered, but see textbook Ch. 7.5 - 7.6 for details.

The beauty of GAMs is that we can use our fitting ideas in this chapter as building blocks for fitting an additive model.

Example: Consider the Wage data.

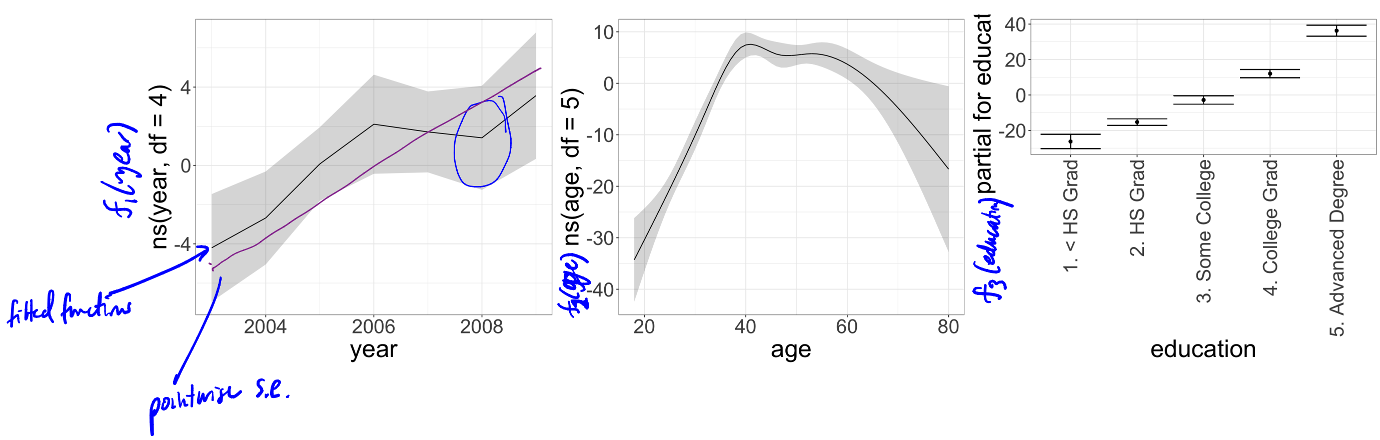
quantitative categorical.

$$Wage = \beta_0 + f_1(year) + f_2(age) + f_3(education) + \varepsilon$$

where f_1 is natural cubic spline w/ 4 df

f_2 is natural cubic spline w/ 5 df

f_3 constant functions for each value (dummy variables)



relationship between each variable and response.

- year: holding age and education fixed, wage tends to increase with year. (inflation?).
- age: holding year and education fixed, wage is low for young and old people, highest for intermediate ages.
- education: holding year + age fixed, wage tends to increase with education.

we could easily replace f_j with different smooth functions and get different fits.

just need to change basis + use least squares.

Pros and Cons of GAMs

4.2 GAMs for Classification

GAMs can also be used in situations where Y is categorical. Recall the logistic regression model:

A natural way to extend this model is for non-linear relationships to be used.

Example: Consider the Wage data.