Chapter 7: Moving Beyond Linarity

So far we have mainly focused on linear models.

Linear models are relatively simple to describe and implement.

Advantages: interpretation & inference.

Disadvantages: can have limited predictive performance because linearity is always an approximation.

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 model ⟹ lowering 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. → extensions of linear model.

We've seen this one already.

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

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

   +: non-linear fit

   –: with large powers polynomial can take strange shapes (especially near the boundary).

2. Step functions: cut the range of a variable into K distinct regions to produce a categorical variable. Fit a piecewise constant function to X.

3. Regression splines: more flexible than polynomial and step functions (extends both)

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

   Polynomials are constrained so that they are smoothly joined.

4. Generalized additive Models extend above to deal w/ multiple predictors.

We are going to start w/ predicting Y on X (single predictor) and extend to multiple.

Note: We can talk regression or classification w/ above. e.g. Logistic regression $P(1|X) = \dfrac{\exp(\beta_0 + \beta_1 X + \beta_2 X^2 + \dots + \beta_d X^d)}{1 + \exp(\beta_0 + \beta_1 X + \beta_2 X^2 + \dots + \beta_d X^d)}$

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 constant in each bin.

details: (1) Create cutpoints $c_1, \dots, c_K$ in the range of X

(2) Construct $K+1$ new variables

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

indicator variables "dummy variables"

Note for any X, $C_0(X) + C_1(X) + \dots + C_K(X) = 1$ because X must lie in exactly 1 interval.

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

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

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

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

When $X < c_1$ all $C_1(X), \dots, C_K(X) = 0$.

$\Rightarrow$ β_0 interpreted as the mean value of Y when $X < c_1$

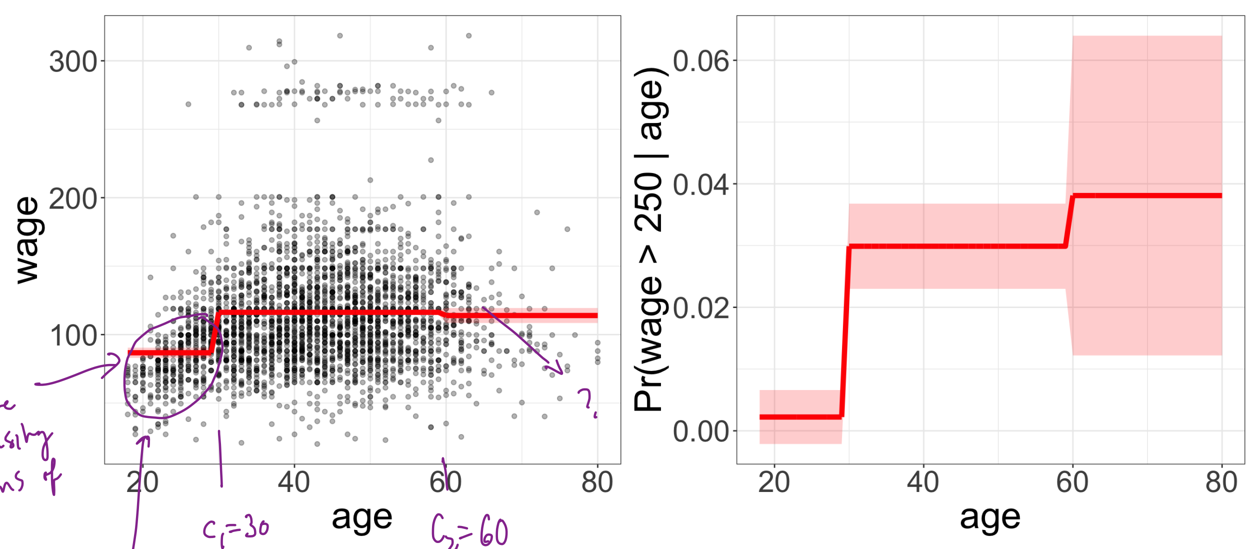
β_j represent average increase in response for $X \in [c_j, c_{j+1})$ relative to $X < c_1$

We can also fit logistic regression for classification

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

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



fitted value of wage using step functions of age.

missing increasing trend.

logistic regression modeling prob of being high earner given age (wage > 250k)

using step function w/ knots at $x = 30, 60$.

Unless there are natural breakpoints in the predictor piecewise constant functions 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

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

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 these ahead of time).

e.g. polynomial regression $b_j(x_i) = x_i^j \quad j = 1, \dots, K$.

e.g. step functions $\quad b_j(x_i) = \mathbb{I}(c_j \le x_i < c_{j+1})$ for $j = 1, \dots, K$.

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.

β's.

$\Rightarrow$ We can also use all of our inferential 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

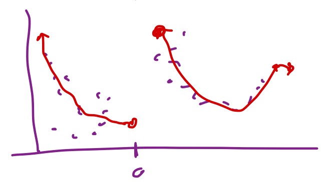
Regression splines are a very common choice for basis function because they are quite flexible, but still interpretable. Regression splines extend upon polynomial regression and piecewise constant approaches seen previously.

START w/

3.1 Piecewise Polynomials (combination of polynomial regression & piecewise constant approach).

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. piecewise cubic w/ knot at c



i.e. fit two different polynomials to data
one on subset for $x < c$
one on subset for $x \geq c$.

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

if fit polynomial of degree 0 $\Rightarrow$ piecewise constant regression.

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

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

each polynomial can be fit using least squares.

Using more knots leads to a more flexible piecewise polynomial.

If we place L knots $\rightarrow$ fitting $L+1$ polynomials
(don't have to be cubic).

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

This leads to $(d+1)(L+1)$ degrees of freedom in model
(# parameters to fit ≈ 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 a jump at the knots.

To go further, we could add two more constraints

(1) 1^{st} derivatives of the piecewise polynomials are continuous at the knots

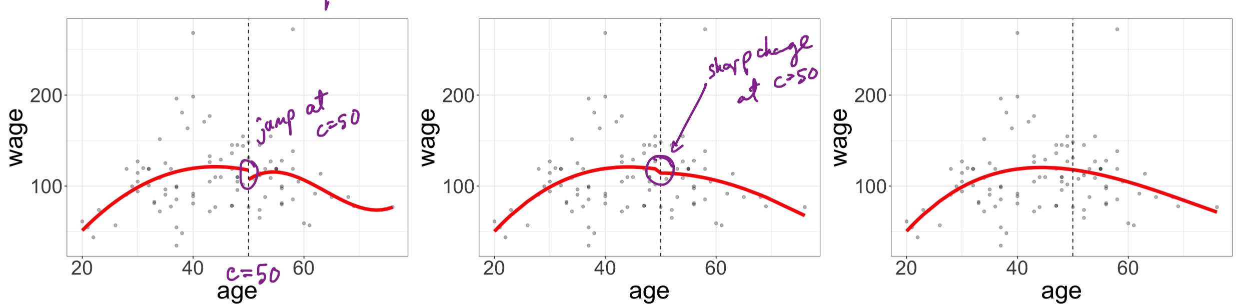
(2) 2^{nd} derivatives of the piecewise polynomials are continuous at the knots.

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, bu reducing the complexity of the resulting fit.

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

cubic spline
cts & 1^{st}, 2^{nd} derivs cts

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 up to $d-1$ derivatives to be continuous at the knots.

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

e.g. cubic spline w/ K knots

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

w/ appropriate basis functions $b_1, \dots, b_{K+3}$.

The most direct way to represent a cubic spline ($d=3$) is to start with the basis for a cubic polynomial and add one *truncated power basis* function per knot.

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

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

This will lead to discontinuity in only the 3rd derivative at each ξ_j w/ continuous first and second derivatives and continuity at ξ_j (each knot).

df : $K+4$ (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.*

i.e. when X is very small or large.

$\Rightarrow$ "natural spline"

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

additional constraint produces more stable predictions 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 change more rapidly).

=> place knots where we think relationship will vary rapidly and less where it is stable.

Most common in practice: place them uniformly

Do this: choose desired degree of freedom (flexibility) + use software to automatically place corresponding # knots at uniform quantiles of data.

funky? → How many knots should we use?

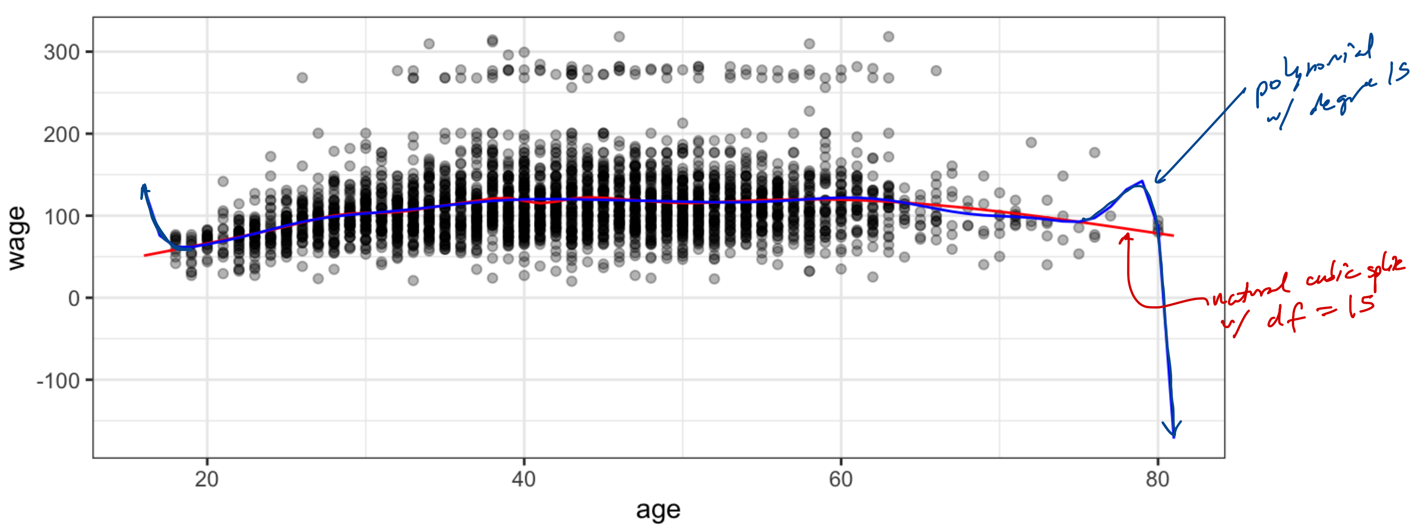
⟺ how many df should we have?

Use C.V.! use k gives smallest CV MSE (or CV error).

3.5 Comparison to Polynomial Regression

Regression splines often give superior results when compared to polynomial regression

Polynomial regression must have high degree to achieve flexible fit (e.g. X^{15}), but regression splines introduce flexibility through knots (w/ degree fixed) => more stability esp. at the boundaries.



extra flexibility of polynomial at boundary produces undesirable result but spline w/ same df looks pretty 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

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

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 based on basis of several predictors $X_1, \dots, X_p$.

4.1 GAMs for Regression

still additive models

can be used for regression or classification (more later).

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

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

idea: replace each linear component $\beta_j x_{ji}$ w/ a smooth non-linear function.

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

$$= \beta_0 + f_1(x_{1i}) + f_2(x_{2i}) + \dots + f_p(x_{pi}) + \varepsilon_i$$

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

possibilities for f_j:

- identity function (leads to linear regression)
- polynomial function
- regression splines (or natural splines).
- smoothing splines
- 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.

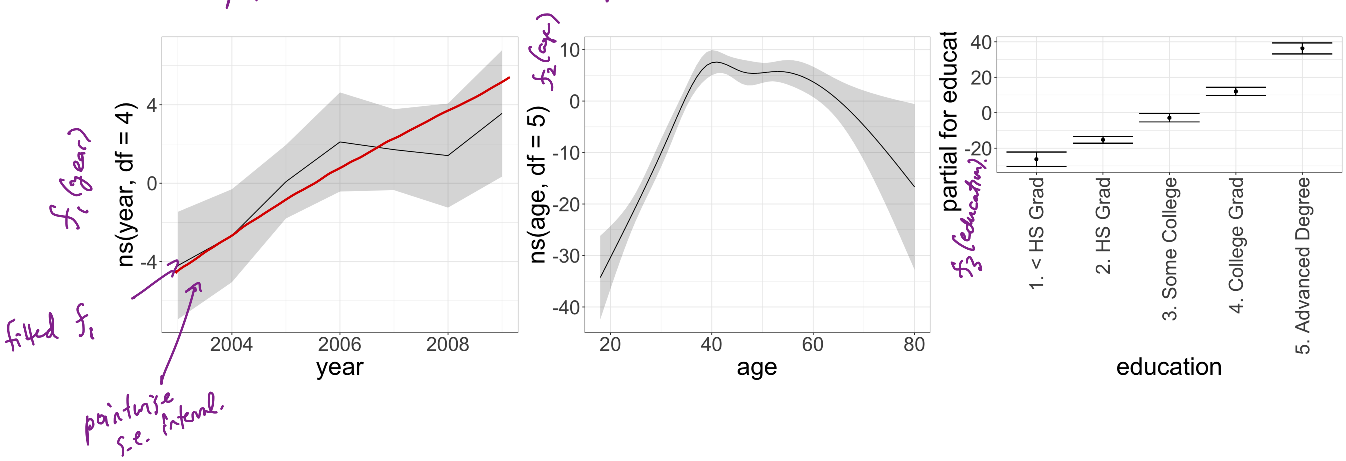
$$\text{Wage} = \beta_0 + f_1(\text{year}) + f_2(\text{age}) + f_3(\text{education}) + \varepsilon$$

where f_1 is a natural spline w/ 4 df.

f_2 is a natural spline w/ 5 df

f_3 is identify function of dummy variables created for each level of education (piecewise constant).

easy to fit w/ least squares by choosing appropriate basis functions.



relationship between each variable and response (holding other variables constant):

- year: hold age and education fixed, wage tends to increase w/ year. (inflation?)
- age: holding year and education fixed, wage is low for young people and old people, highest for intermediate age.
- education: holding year and age fixed, wage tends to increase w/ education level.

> We could easily replace f_j with different smooth functions and get different fits. just need to change basis and use least squares.

Pros and Cons of GAMs

Advantages

- GAMs allow nonlinear fit f_j for each X_j to model non-linear relationships that linear regression will miss.

- nonlinear fits can potentially allow for more accurate prediction on the response (if there is a truly non-linear relationship).

- additive model $\Rightarrow$ we can still examine the effect of each X_j on Y individually while holding all other variables fixed
  $\Rightarrow$ GAMs provide a useful representation for inference / interpretation.

- smoothness of f_j for X_j can be summarised by df.

Limitations

- model is restricted to be additive
  i.e. w/ many predictors, important interactions can be missed.

  solution: as w/ linear regression we can manually add interaction terms by including additional predictors of the form $X_j * X_k$
  or add low dimensional interaction functions of the form $f_{jk}(X_j, X_k)$

  e.g. two-dimensional spline (not covered).

For fully general models, we have to look for an even more flexible approach like random forests or boosting (next).

GAMs provide a useful compromise between linear and fully nonparametric approaches.

4.2 GAMs for Classification

assume Y takes values in 0,1 (generalizations exist to more categories).

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

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p$$

logit = log-odds of $P(Y=1|X)$ vs. $P(Y=0|X)$.
as linear function of predictors.

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

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + f_1(X_1) + \dots + f_p(X_p).$$

"logistic regression GAM"

Example: Consider the Wage data.

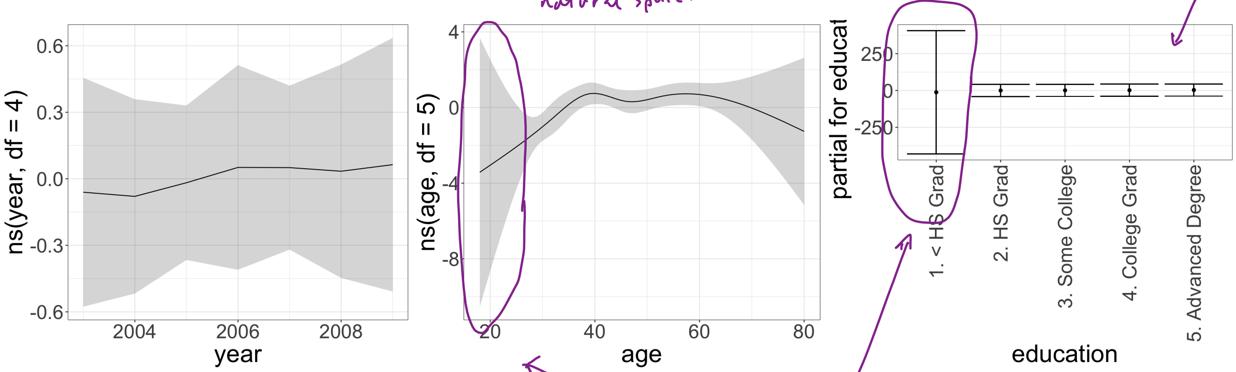
let Y = Wage > \$250k "high earners"

we could fit a logistic regression GAM

piecewise constant for each education level

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \underbrace{\overset{df=4}{f_1(\text{year})} + \overset{df=5}{f_2(\text{age})}}_{\text{natural splines}} + f_3(\text{education})$$

can't see, but increasing w/ education.



this looks quite linear
maybe replace w/ linear without much loss of information and ↑ lower variance.

Not many people in data set w/ < HS grad or low age and high earners.

looking at the scales age & education have more of effect on P(high earner | X) than year.