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 the variance of estimates

still a linear model! Can only improve 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 to linear model.

we've seen this already.

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

   e.g. cubic regression 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 very 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 polynomials & step functions (extends both)

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

   Polynomials are constrained so that they are smoothly joined.

4. Generalized additive models extends above to deal w/ multiple predictors.

We will start w/ predicting Y on X (one predictor) and extend to multiple.

Note: we can talk about regression or classification w/ above, e.g. logistic regression

$$P(Y=1|X) = \frac{\exp(\beta_0 + \beta_1 X + \beta_2 X^2 + \dots + \beta_d X^d)}{1 + \exp(\beta_0 + \beta_1 X + \beta_2 X^2 + \dots + \beta_d X^d)}$$

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 a different constant in each bin.

details: (1) create cut points $c_1, c_2, \dots, c_K$ in the range of X.

(2) Construct $K+1$ new variables

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

Note: for any X,

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

since X must be in exactly 1 interval.

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

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

note: 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.

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

$\Rightarrow$ β_0 interpreted as mean value for Y when $X < c_1$.

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

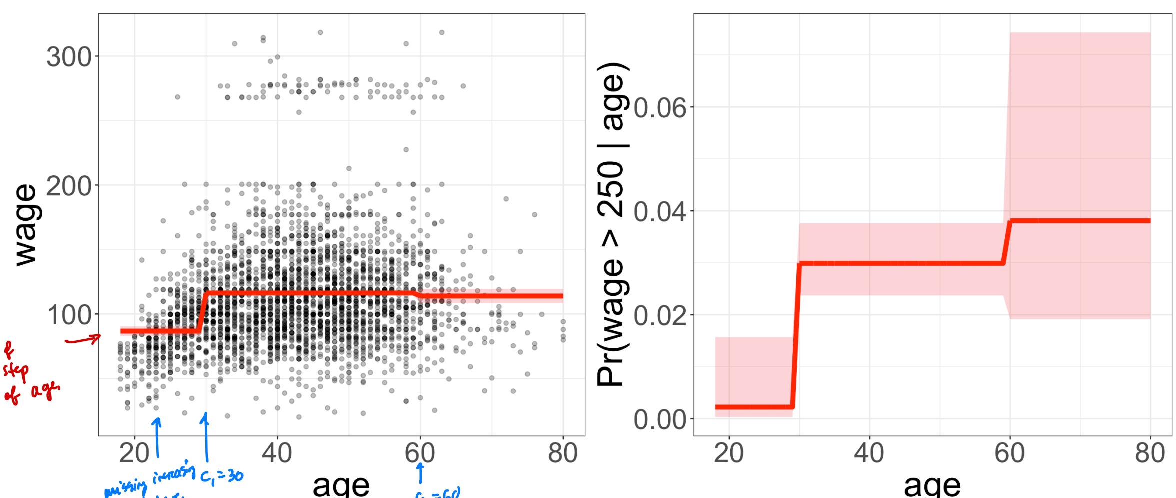
We can also fit the logistic regression model for classification:

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

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

X y: wage

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 here. $c_1=30$ $c_2=60$

Unless there are natural break points in the predictor, piecewise constants can miss trends.

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

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

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

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

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

$$y_i = \beta_0 + \beta_1 b_1(x_i) + \beta_2 b_2(x_i) + \cdots + \beta_K b_K(x_i) + \varepsilon_i$$

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

e.g. polynomial regression: $b_j(x_i) = x_i^j \quad j = 1, \ldots, d.$

e.g. step function: $b_j(x_i) = \mathbb{I}(c_j \le x_i < c_{j+1})$.

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.

$\Rightarrow$ can use all our inference tools for linear model: e.g. $se(\hat{\beta}_j)$ and F-statistics for model significance.

Many choices exist for basis functions:

e.g. wavelets, fourier series, regression splines

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 with.

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.

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

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

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

i.e. fitting different polynomials to the data, one on subset $x < c$ and one on subset $x \geq c$.

each polynomial can be fit using least squares.

Using more knots leads to a more flexible piecewise polynomial.

if we place L knots $\Rightarrow$ fit $L+1$ polynomials.

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

This leads to $(d+1)(L+1)$ parameters to fit $\approx$ complexity/flexibility "degrees of freedom" in the model.

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 the knots.

To go further, we could add two more constraints

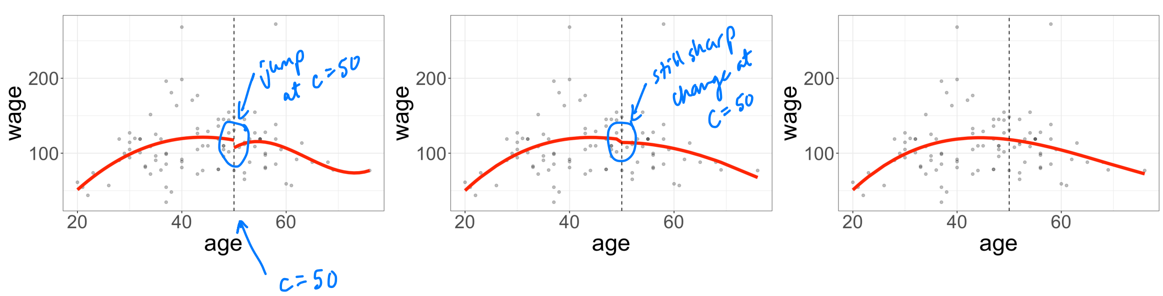
1. first derivatives of piecewise polynomials are continuous at knots
2. 2^{nd} derivatives of piecewise polynomials are continuous at knots.

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, by 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 with continuity in derivatives up to degree $d-1$ at each knot.



piecewise cubic polynomial

piecewise cubic polynomial w/ continuity.

cubic spline
cts + cts 1^{st} & 2^{nd} derivative.

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

We can use the basis model to represent a regression spline

cubic spline w/ L knots.

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

for appropriate basis functions $b_1, b_2, \dots, b_{L+3}$

x, x^2, x^3

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.

ξ = 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 a knot.

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

Homework 5 ☆ 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: $L+4$ (cubic spline w/ L knots).

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

when x is small or large.

$\Rightarrow$ "natural spline"

function required to be linear at boundary (where x is smaller than 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 have a lot of knots (coefficients change more rapidly).

$\Rightarrow$ place knots where we think function will vary rapidly and less where its more stable.

more common in practice: place them uniformly

to do this, choose desired degrees of freedom (flexibility) + use software to automatically place corresponding # of knots at uniform quantiles of the data.

How many knots should we use?

$\Longleftrightarrow$ how many degrees of freedom should we use?

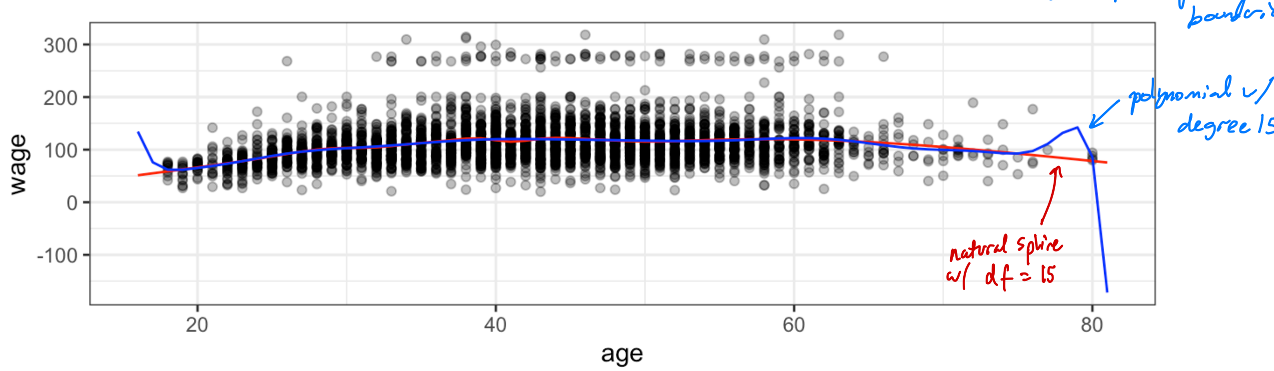
Use CV! Use L giving smallest CV MSE!

3.5 Comparison to Polynomial Regression

Regression splines often give superior results to polynomial regression.

Polynomial regression must use high degree to achieve same level of flexibility (i.e., x^{15})

but regression splines introduce flexibility through knots (degree fixed) $\Rightarrow$ more stability. (especially at boundaries).



extra flexibility of polynomial at boundary produces undesirable results, but the spline w/ same flexibility (df) 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

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

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 on the basis of 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: $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} + \varepsilon_i$

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

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

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

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

possibilities for f_j:

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

not covered, 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

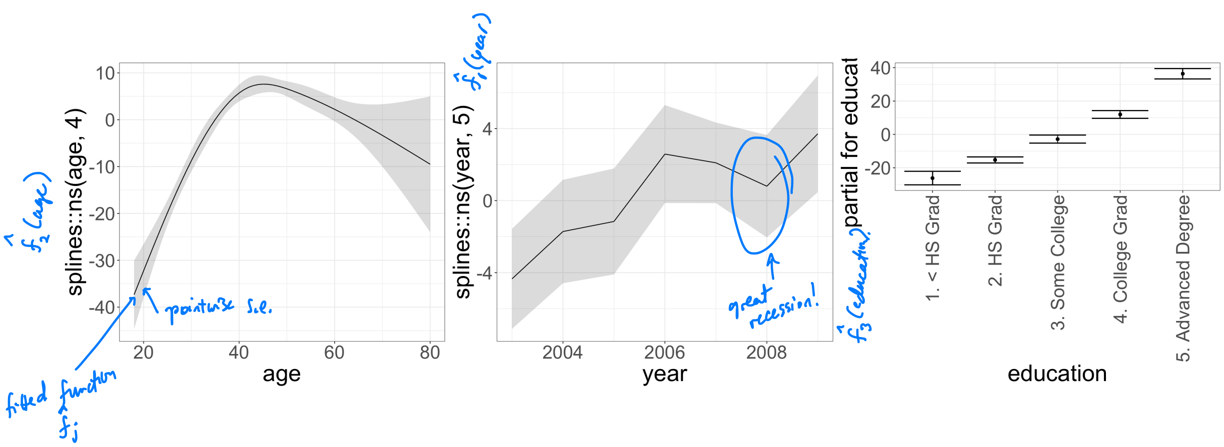
$$\text{Wage} = \beta_0 + f_1(\text{year}) + f_2(\text{age}) + f_3(\text{education}) + \varepsilon$$

where f_1 is natural spline w/ 4 df

f_2 is natural spline w/ 5 df

f_3 is identity of dummy variables created from education.

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



relationship between each variable and the response:

- age: holding year and education fixed, wage is low for young people and old people, highest for intermediate ages.

- year: holding age and education, wage tends to increase w/ year (inflation?)

- education: holding age and year fixed, ↑ education is associated w/ ↑ wage.

We could easily replace f_j w/ different smooth functions to get different fits.

just need to change the basis and use least squares.

Pros and Cons of GAMs

Advantages:

- GAMs allow nonlinear fits f_j to each predictor X_j model non-linear relationships that linear regression will miss.
- If there is truly a nonlinear relationship, can allow for more accurate prediction.
- 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 summarized by df.

Limitations:

- model is restricted to be additive

  i.e. with many variables, important interactions will be missed.

  solution: as with linear regression, we can manually add interaction terms by including additional predictors of the form $X_j \times X_k$

  or add low dimension interaction functions of form $f_{jk}(X_j, X_k)$

  two-dimensional splines (not covered).

For fully general models, we have to look for even more flexible approaches like random forests or boosted trees (next).

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

4.2 GAMs for Classification

assume y takes value 0 or 1 (generalizations exist for more categories).

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

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

log-odds are linear in predictors

$p(X) = P(Y=1 \mid X)$

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)

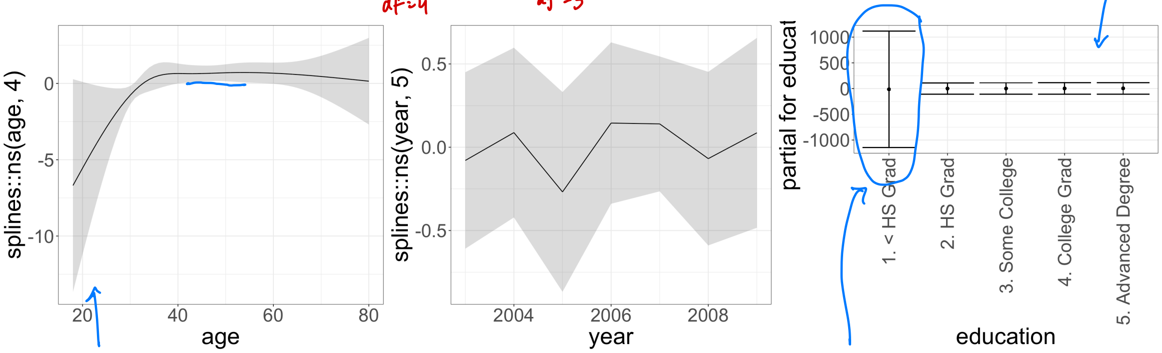
we could fit a gam:

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + f_1(\text{year}) + f_2(\text{age}) + f_3(\text{education})$$

natural splines df=4 df=5

dummy encoding.

increasing w/ education.



If you're under 32, you have less likely of being a high earner (year and education held fixed).

not much relationship

nobody in data set w/ < HS education and wage > 250k