Chapter 7: Moving Beyond Linearity

So far we have mainly focused on linear models.

linear models are relatively easy to describe and implement.

Advantages: interpretation & inference

Disadvantages: can have limited predictive power because linearity is almost 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.

on relationship btw Y and X

We have already talked about this one.

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

e.g. cubic regression uses 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
+ easy to interpret
− 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 polynomials & step functions (extends both)

idea: cut the 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: extend above to deal w/ multiple predictors.

We will start w/ predicting Y on X (one predictor) and extend to multiple (4).

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

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 $c_1, \ldots, c_K$ in the range of X.

(2) Construct $K+1$ new variables

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

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

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

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

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

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

When $X < c_1$, all predictors $C_1, \ldots, C_K = 0$.

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

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

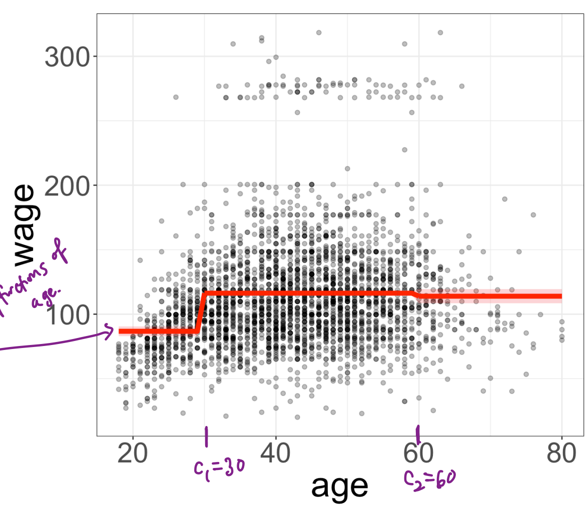
We can also fit the logistic regression model for classification

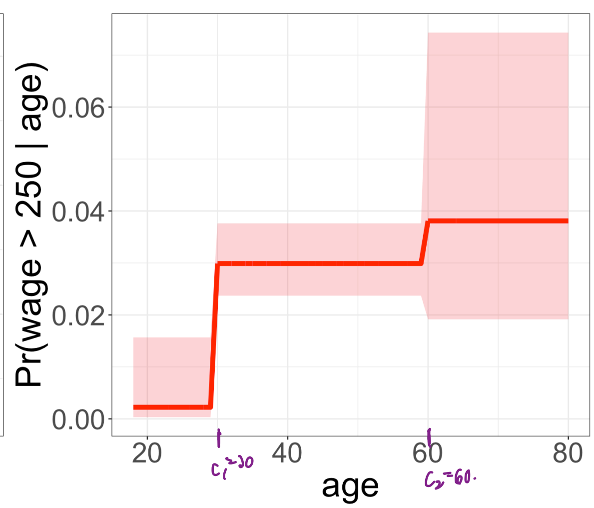
$$P(Y=1 \mid X) = \frac{\exp(\beta_0 + \beta_1 C_1(X) + \ldots + \beta_K C_K(X))}{1 + \exp(\beta_0 + \beta_1 C_1(X) + \ldots + \beta_K C_K(X))}.$$

Example: Wage data. 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.





logistic regression modeling prob of being a high earner given age.

Unless there are natural cutpoints in 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 function or transformations that can be applied to a predictor X

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

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

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

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.

$\Rightarrow$ 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 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.

e.g. fitting cubic polynomial over intervals break up global range.

↑ knots.

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

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

ie. fit two different polynomials to the data, one on subset for $x \le c$ one on subset for $x > c$.

each polynomial can be fit using least squares.

Using more knots leads to a more flexible piecewise polynomial.

If we place K knots $\Rightarrow$ fit $K+1$ polynomials (doesn't have to be cubic.

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

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

(# parameters we have 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 the knots.

To go further, we could add two more constraints

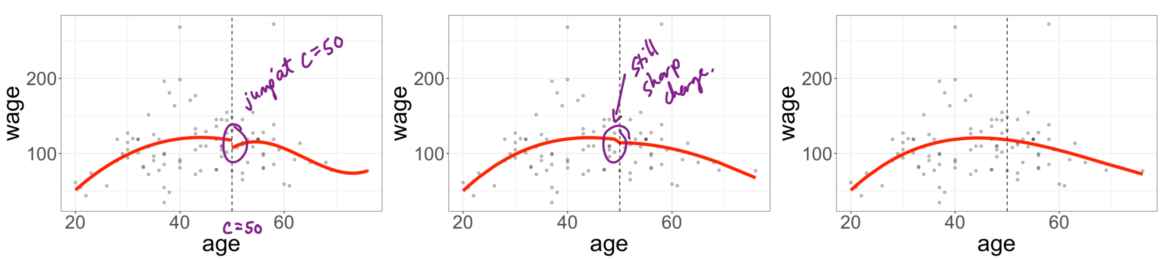
1. first derivatives of the piecewise polynomials are continuous at the knots
2. 2^{nd} derivatives of the piecewise polynomials are continuous at the knots

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, 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} derivatives

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

We can use the basis function idea 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 $b_1, \dots, b_{L+3}$

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.

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

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

→ This will lead to discontinuity in only the 3rd derivative at each ξ_j w/ continuous function and 1st and 2nd derivatives at each knot ξ_j.

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.*

$\Rightarrow$ "natural spline"

requires function to be linear at the boundary (where X is smaller than the smallest knot and bigger than biggest 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 function will vary rapidly and less where more stable.

more common in practice: place them uniformly.

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

How many knots should we use?

⇔ how flexible do we want our function?

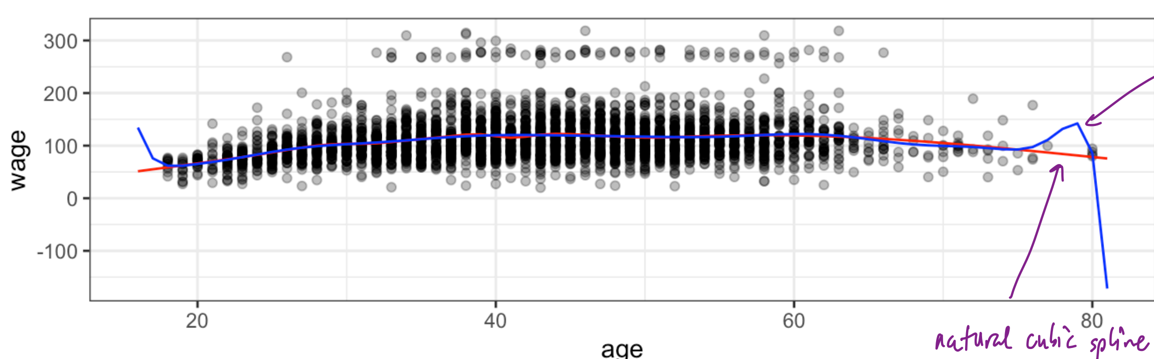
Use CV! Use L gives smallest CV error!

3.5 Comparison to Polynomial Regression

Regression splines (and natural splines) often give superior results to polynomial regression.

Polynomial regression must use high degree to achieve flexible fit (e.g. X^{15}), but

Regression splines introduce flexibility through knots (but degree fixed) ⇒ more stability (esp. at boundaries).



polynomial w/ degree 15

natural cubic spline w/ df = 15.

high degree of polynomial (to achieve flexibility) at the borders produces undesirable results.

The natural 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 $X_1, \dots, X_p$.

still additive models.

4.1 GAMs for Regression

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

idea: replace each linear component $\beta_j x_{ji}$ with a smooth nonlinear function of x_{ji}

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

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

"additive" because we calculate a separate f_j for each x_j and add them together!

possibilities for f_j:

- identity (leads to linear regression)
- polynomial function
- regression spline (natural spline)
- 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

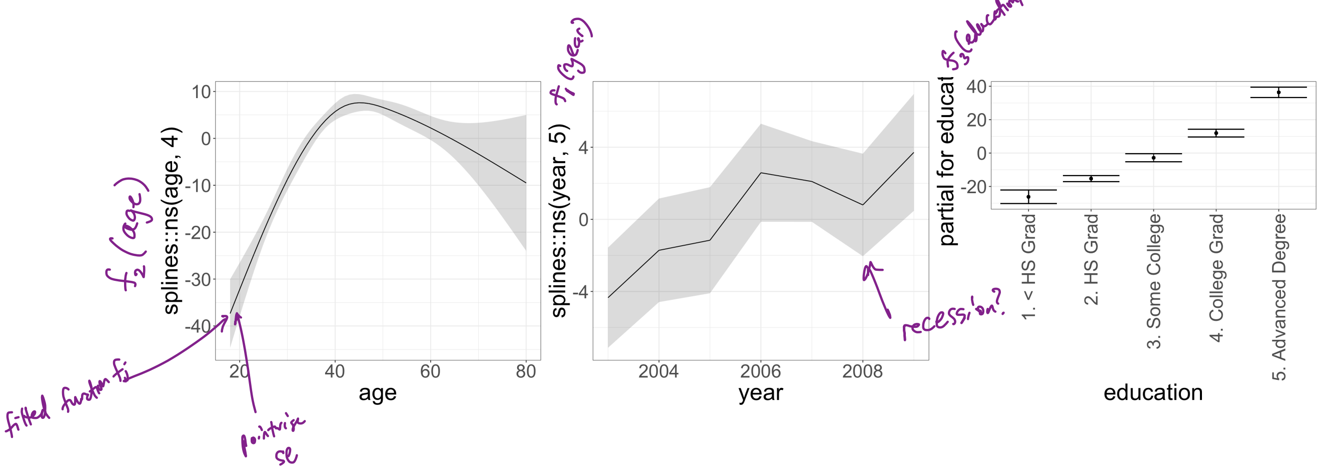
$$\text{Wage} = \beta_0 + f_1(\text{year}) + f_2(\text{age}) + f_3(\text{education}) + \varepsilon$$

where f_1 is natural spline w/ 4 df

f_2 is natural spline w/ 5 df

f_3 is identity of dummy variables created from education (piecewise constant).

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



fitted relationship btw each variable and the response.

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

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

- education: holding year and age fixed, wage tends to increase with education

We could have easily replaced f_j with different functions and gotten a different fit. Just need to change basis and use least squares. (choose best fit using CV - lowest error).

Pros and Cons of GAMs

Advantages:

- GAMs allow nonlinear fit f_j to each X_j (model non-linear relationships that linear regression will miss).
- allow for more accurate predictions if there truly is a non-linear relationship.
- additive model $\Rightarrow$ can still examine the effect of each X_j on Y individually while holding all others fixed.
  $\Rightarrow$ GAMs provide a useful respresentation for inference / interpretation.
- summarize flexibility of model by df.

Limitations:

- model is restricted to be additive
  i.e important interactions can be missed

  Solution: as with linear regression, we can manually add iteraction terms by including additional predictors of the form $X_j \times X_k$ or add interaction functions of the form $f_{jk}(x_j, x_k)$.

  two dimensional splines (not covered).

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

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

4.2 GAMs for Classification

assume Y takes 0 or 1 (generalizations exist to more categories).

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

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

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.