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

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

$$C_0(X) = \mathbb{I}(X < c_1)$$
$$C_1(X) = \mathbb{I}(c_1 \le X < c_2)$$
$$\vdots$$
$$C_K(X) = \mathbb{I}(c_K \le X)$$

"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 $X < C_1$.

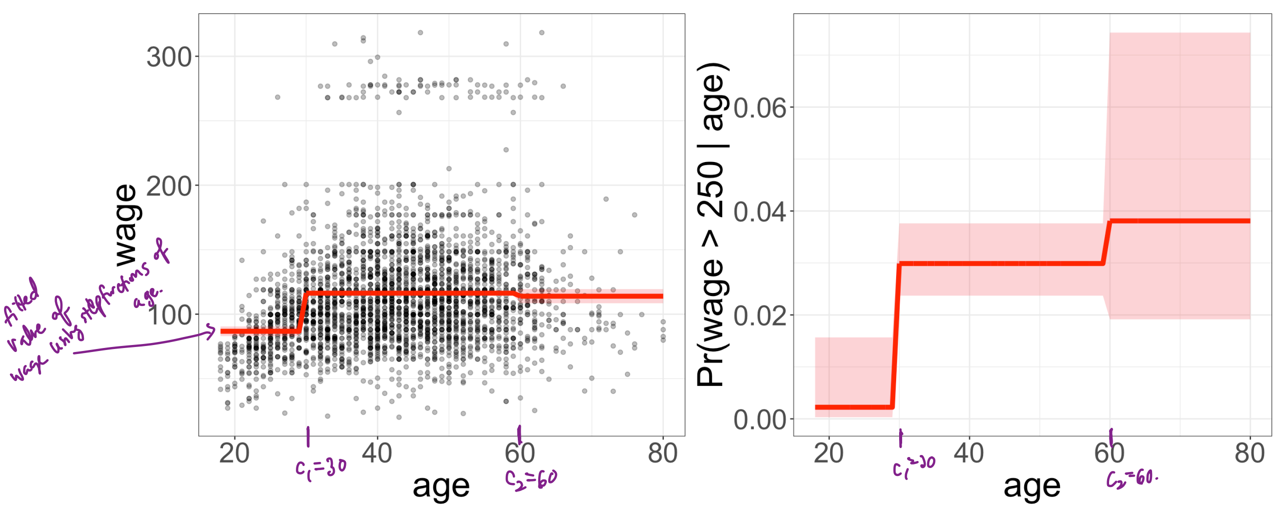
β_j represent the average increase in response for $X \in [C_j, C_{j+1})$ relative to $X < C_1$.

We can also fit the logistic regression model for classification

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

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



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:

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

Note that the basis functions are fixed and known.

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.

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.

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

Using more knots leads to a more flexible piecewise polynomial.

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

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.

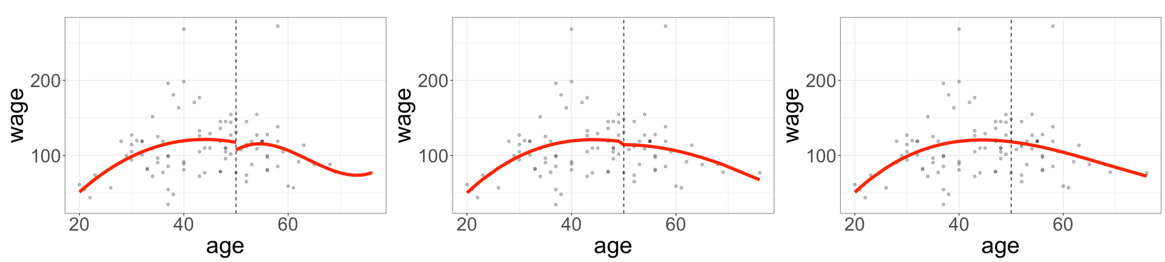
To go further, we could add two more constraints

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



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.

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.

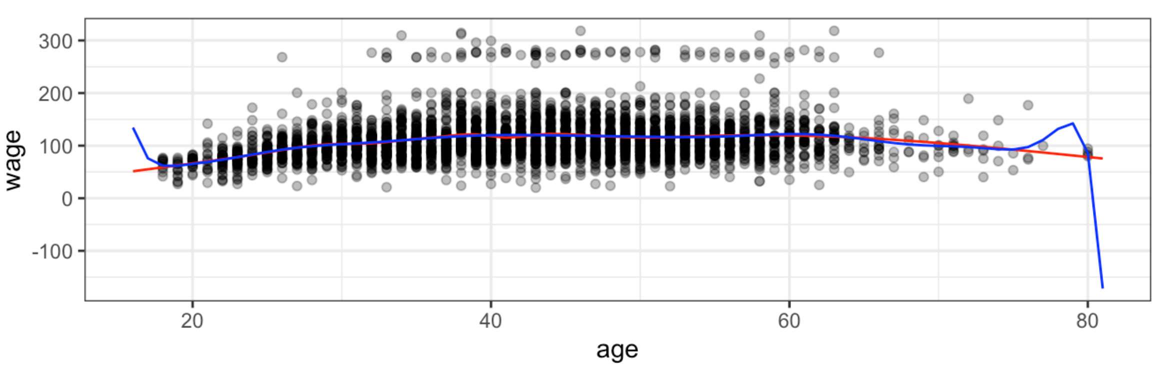
Unfortunately, splines can have high variance at the outer range of the predictors. One solution is to add *boundary constraints*.

3.4 Choosing the Knots

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

How many knots should we use?

3.5 Comparison to Polynomial Regression



4 Generalized Additive Models

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

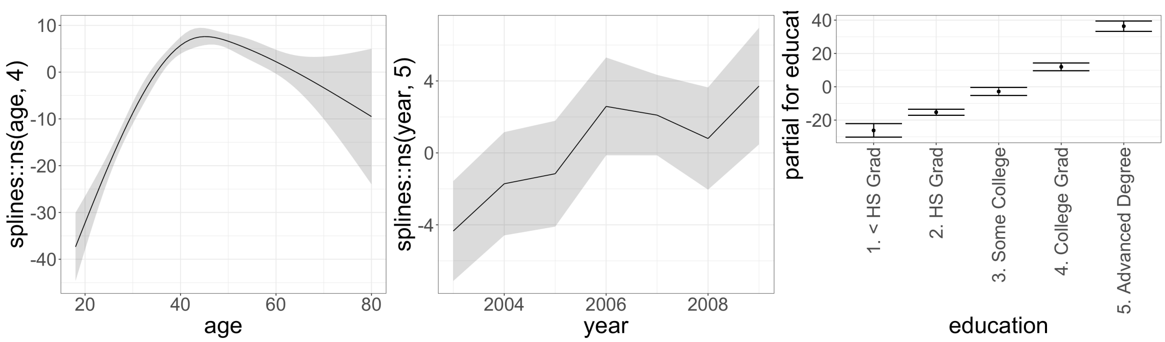
Generalized Additive Models (GAMs) provide a general framework for extending a standard linear regression model by allowing non-linear functions of each of the variables while maintaining *additivity*.

4.1 GAMs for Regression

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

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.



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.