MONASH University

ETC3250

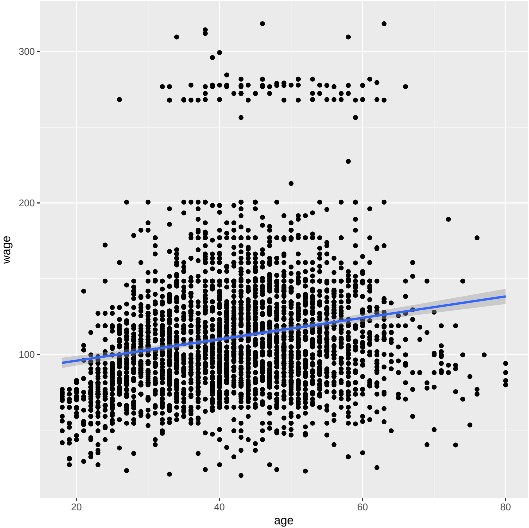
Business Analytics

Week 3
Flexible regression

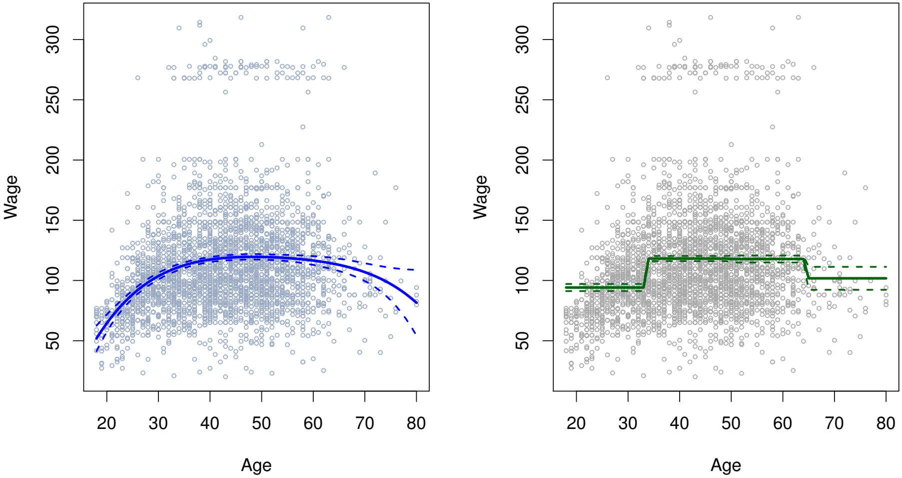
14 March 2018

Outline

Moving beyond linearity



Moving beyond linearity



Moving beyond linearity

The truth is never linear! Or almost never!

But often the linearity assumption is good enough.

When it's not . . .

- polynomials,
- step functions,
- **splines**,
- local regression, and
- **generalized additive models**

offer a lot of flexibility, without losing the ease and interpretability of linear models.

Basis functions

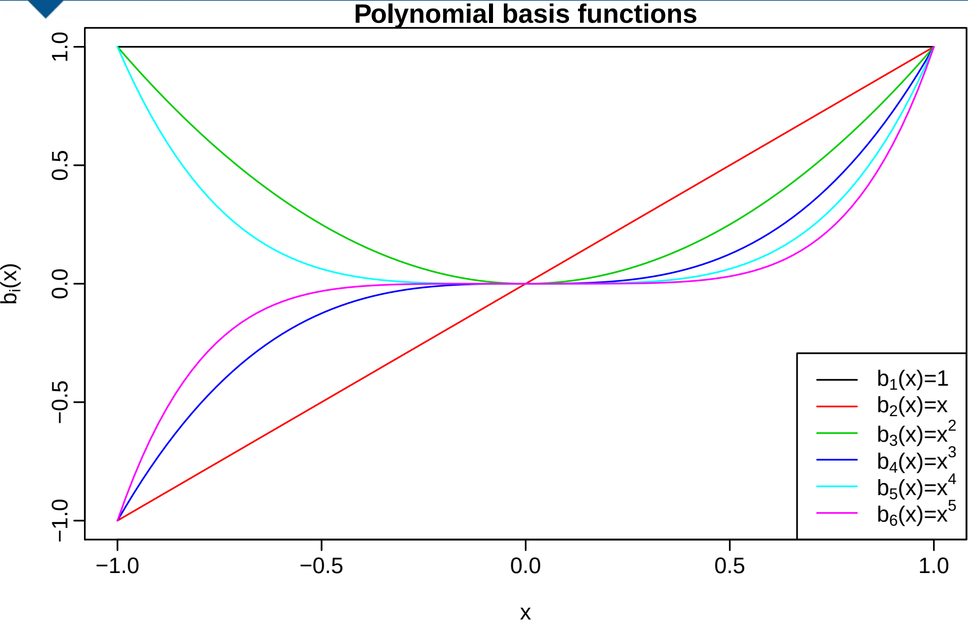
Instead of fitting a linear model (in X), we fit the model

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

where $b_1(X), b_2(X), \ldots, b_K(X)$ are a family of functions or transformations that can be applied to a variable X, and $i = 1, \ldots, n$.

- Polynomial regression: $b_k(x_i) = x_i^k$
- Piecewise constant functions: $b_k(x_i) = I(c_k \leq x_i \leq c_{k+1})$
- ...

Basis functions - polynomial



Outline

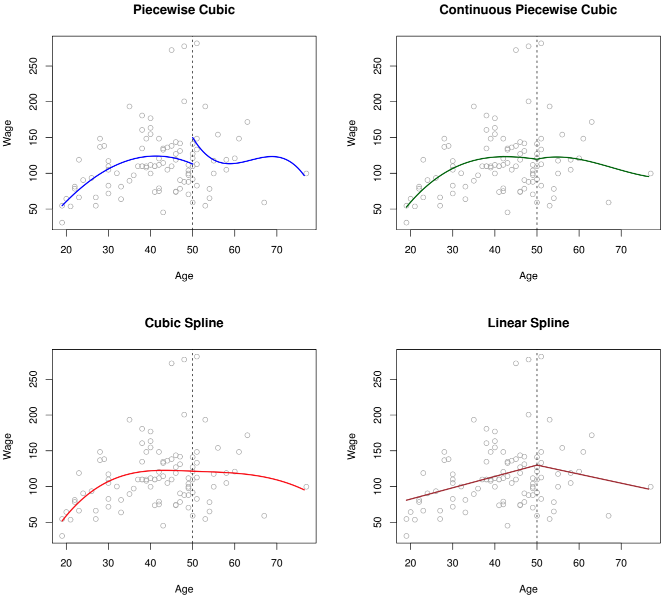
Splines

Knots: $\kappa_1, \ldots, \kappa_K$.

A spline is a continuous function $f(x)$ consisting of polynomials between each consecutive pair of 'knots' $x = \kappa_j$ and $x = \kappa_{j+1}$.

- Parameters constrained so that $f(x)$ is continuous.
- Further constraints imposed to give continuous derivatives.

Splines

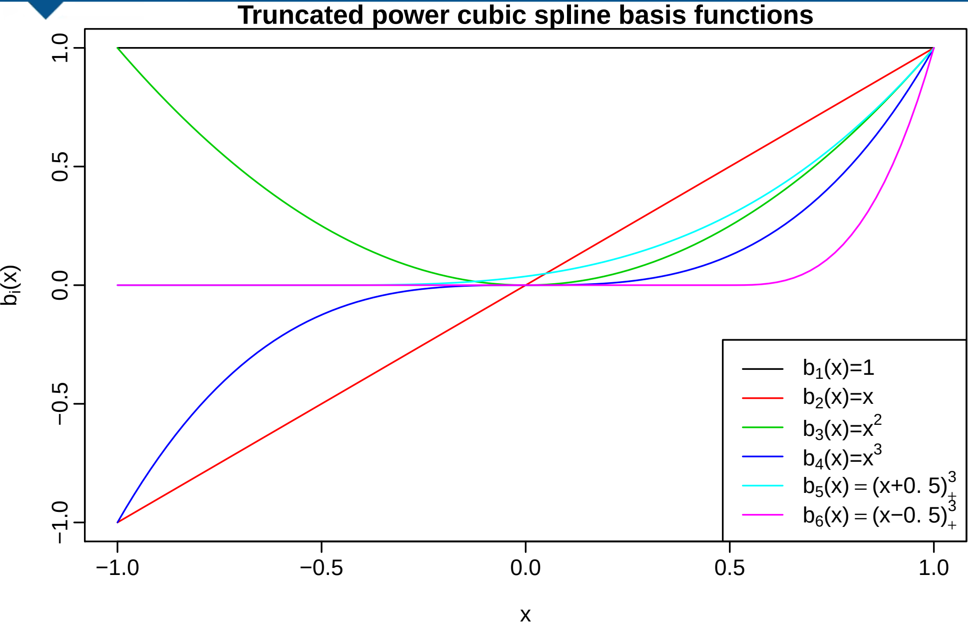


Piecewise Cubic
Continuous Piecewise Cubic
Cubic Spline
Linear Spline
Wage
Age
50
100
150
200
250
20
30
40
50
60
70

Spline basis representation

- Truncated power basis
- Predictors: $x, \ldots, x^p, (x - \kappa_1)_+^p, \ldots, (x - \kappa_K)_+^p$
- Then the regression is piecewise order-p polynomials.
- $p - 1$ continuous derivatives.
- Usually choose $p = 1$ or $p = 3$.
- $p + K + 1$ degrees of freedom

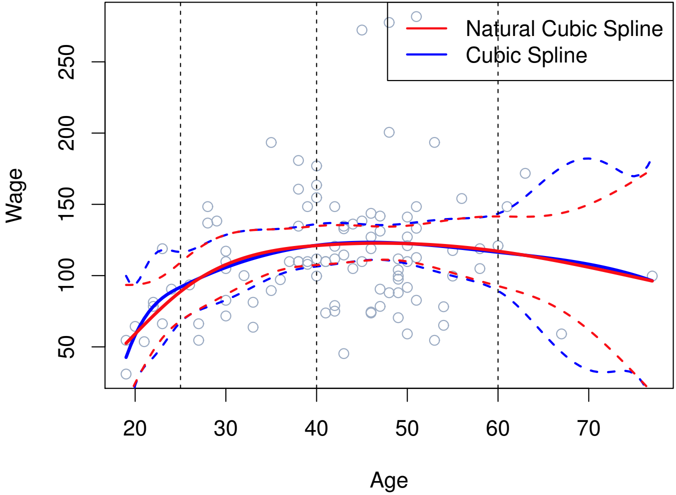
Truncated power basis



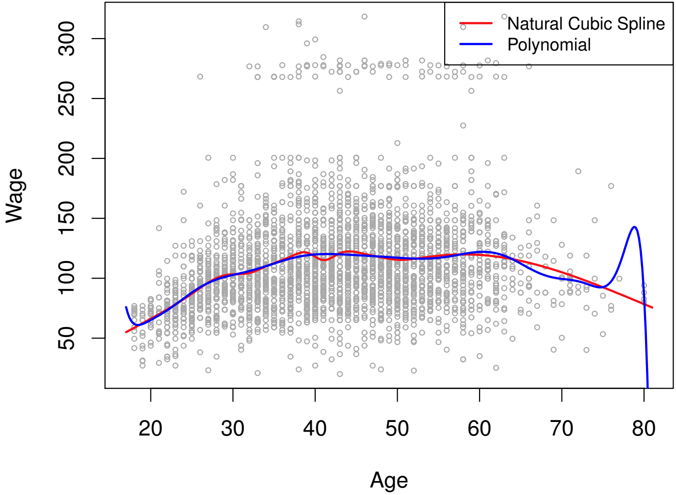
Natural splines

- Splines based on truncated power bases have high variance at the outer range of the predictors.
- Natural splines are similar, but have additional boundary constraints: the function is linear at the boundaries. This reduces the variance.
- Degrees of freedom $\mathrm{df} = K$.
- Create predictors using ns function in R (automatically chooses knots given df).

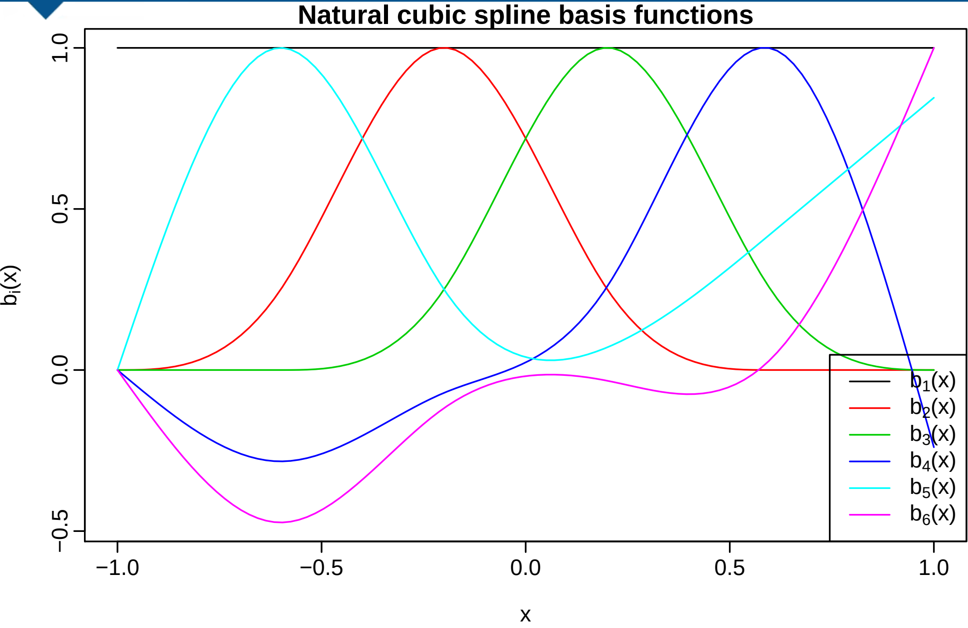
Natural splines



Natural splines



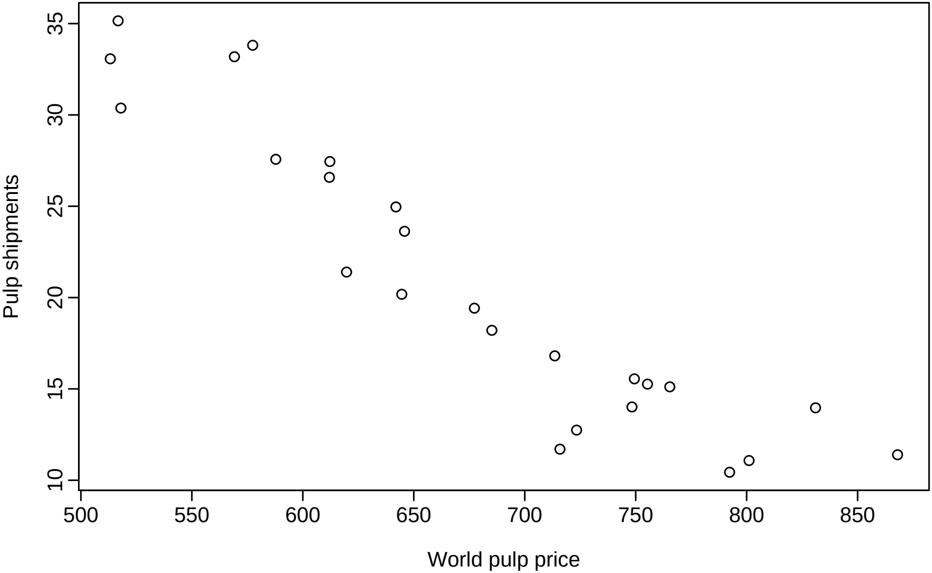
Natural splines



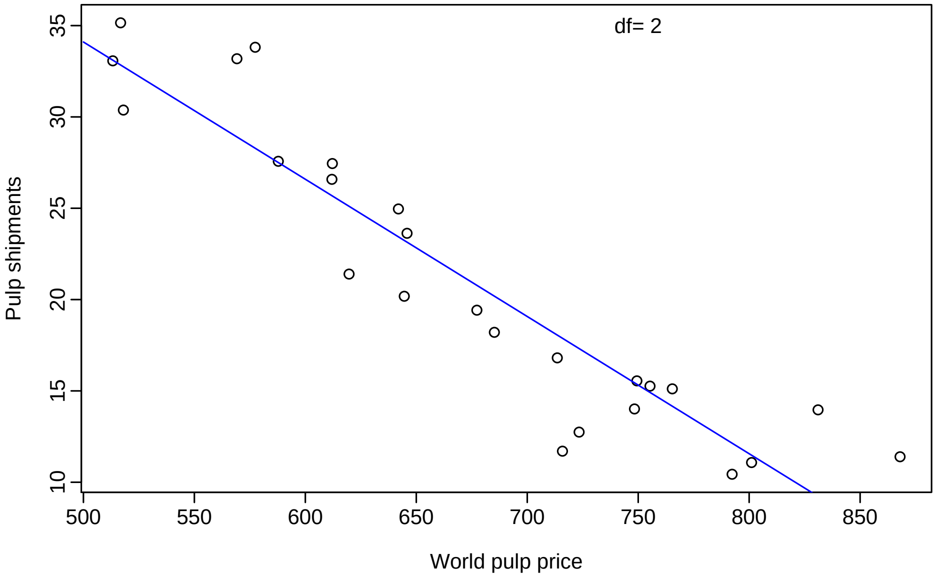
Knot placement

- Strategy 1: specify `df` (equivalently K) and let `ns` place them at appropriate quantiles of the observed X.
- Strategy 2: choose K and their locations.

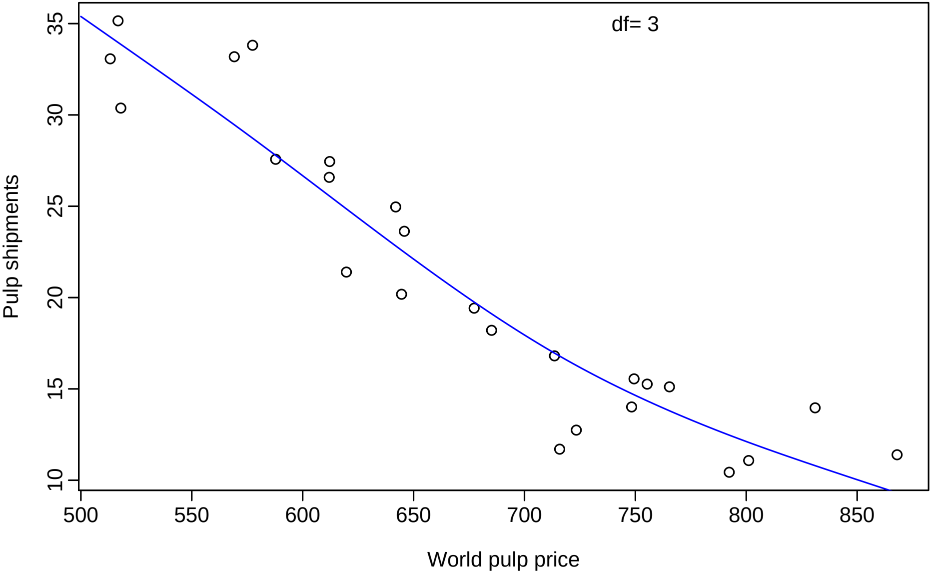
Splines - degree of freedom



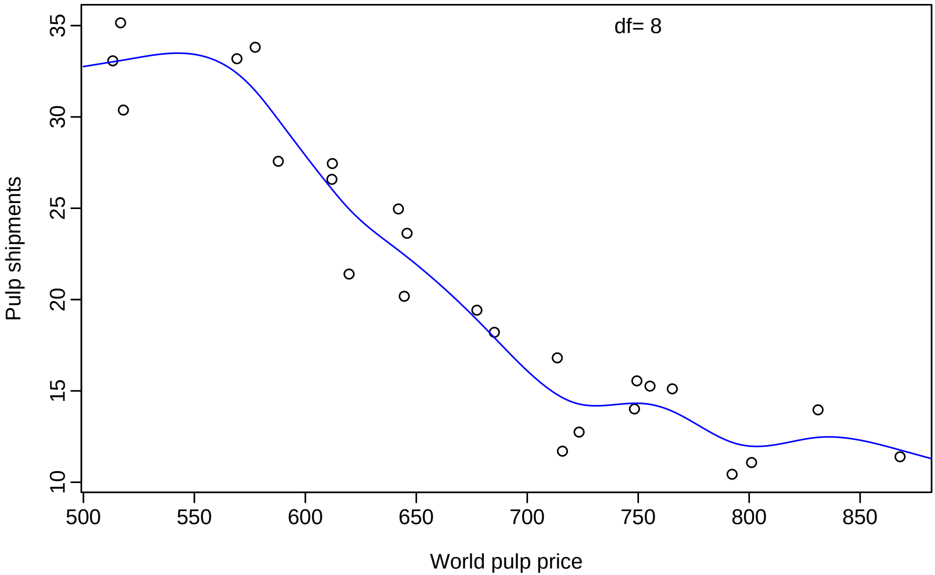
Splines - degree of freedom



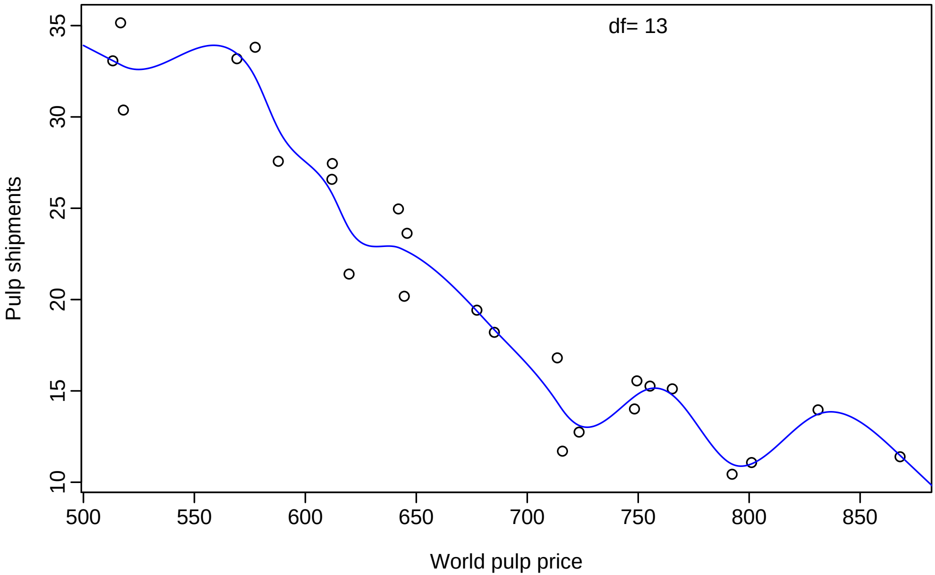
Splines - degree of freedom



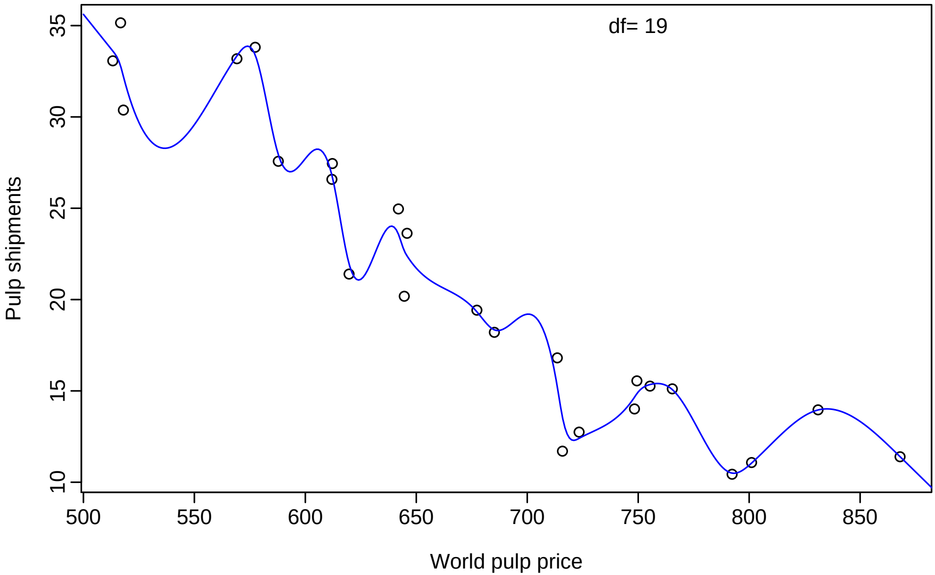
Splines - degree of freedom



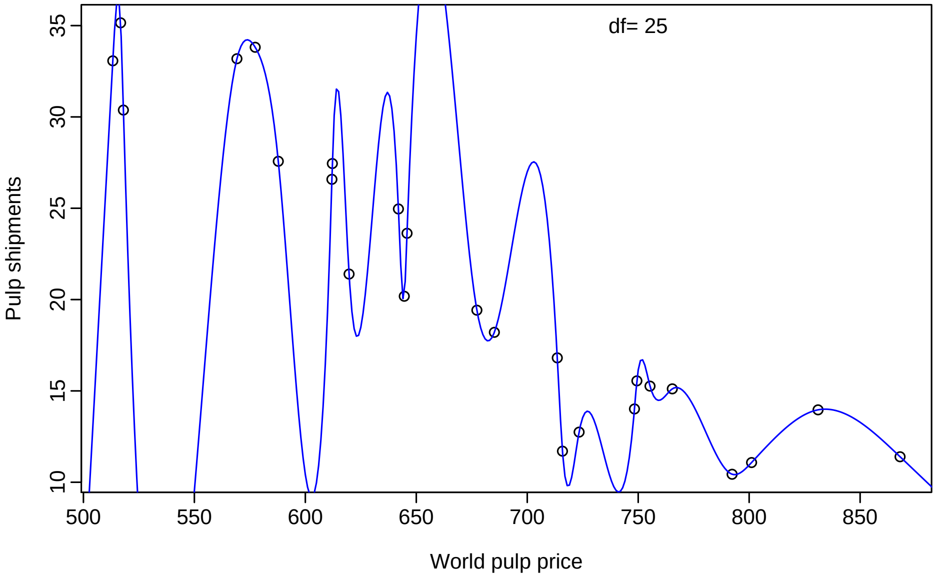
Splines - degree of freedom



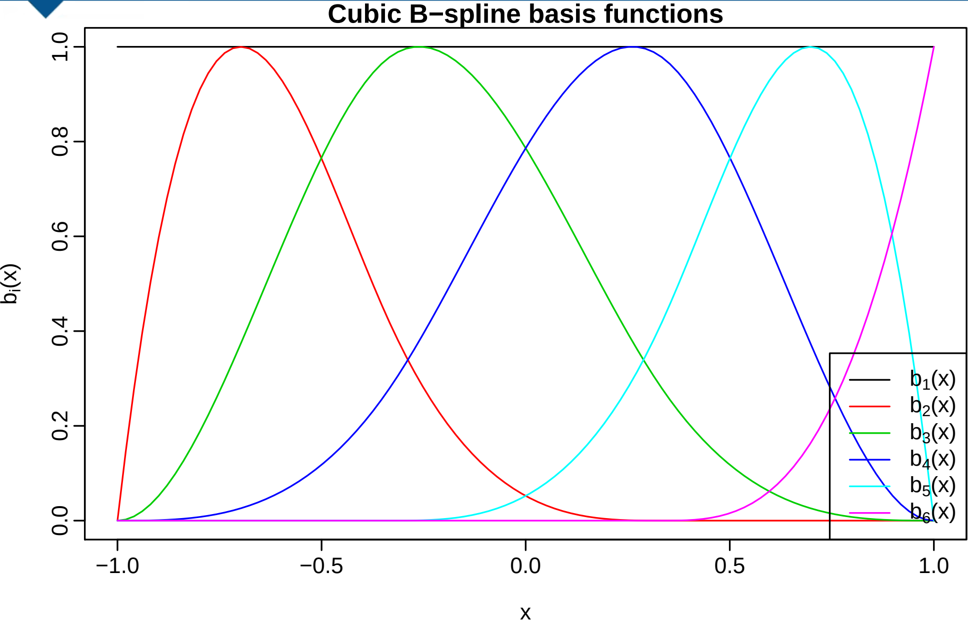
Splines - degree of freedom



Splines - degree of freedom



Other basis functions



Outline

The curse of dimensionality

Why is it hard to fit models of the form

$$y = f(x_1, x_2, \ldots, x_p) + e?$$

- Data is very sparse in high-dimensional space.
- Model assumes p-way interactions which are hard to estimate.

The curse of dimensionality

Why is it hard to fit models of the form

$$y = f(x_1, x_2, \ldots, x_p) + e?$$

- Data is very sparse in high-dimensional space.
- Model assumes p-way interactions which are hard to estimate.

Generalized Additive Models

Allows for flexible nonlinearities in several variables, but retains the additive structure of linear models.

$$y_i = \beta_0 + f_1(x_{i,1}) + f_2(x_{i,2}) + \cdots + f_p(x_{p,1}) + e_i$$

- Each f_i is a smooth univariate function.

Generalized Additive Models

- Can fit a GAM simply using, e.g. natural splines:

  ```
  lm(wage ~ ns(year,df=5) + ns(age,df=5) + education)
  ```
- Coefficients not that interesting; fitted functions are.
- Use `plot.gam` from gam package.
- Can mix terms — some linear, some nonlinear — and use `anova()` to compare models.
- GAMs are additive, although low-order interactions can be included in a natural way using, e.g. bivariate smoothers or interactions of the form ns(age,df=5):ns(year,df=5).

Interactions and additivity

- Additive models assume no interactions.
- Add bivariate smooths for two-way interactions.
- Graphically check for interactions using faceting.

```
qplot(age, wage, data = Wage) + facet_wrap(~ year)
```