MONASH University

ETC3250

Business Analytics

3. Nonparametric regression

13 August 2015

Outline

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.

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.

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.

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

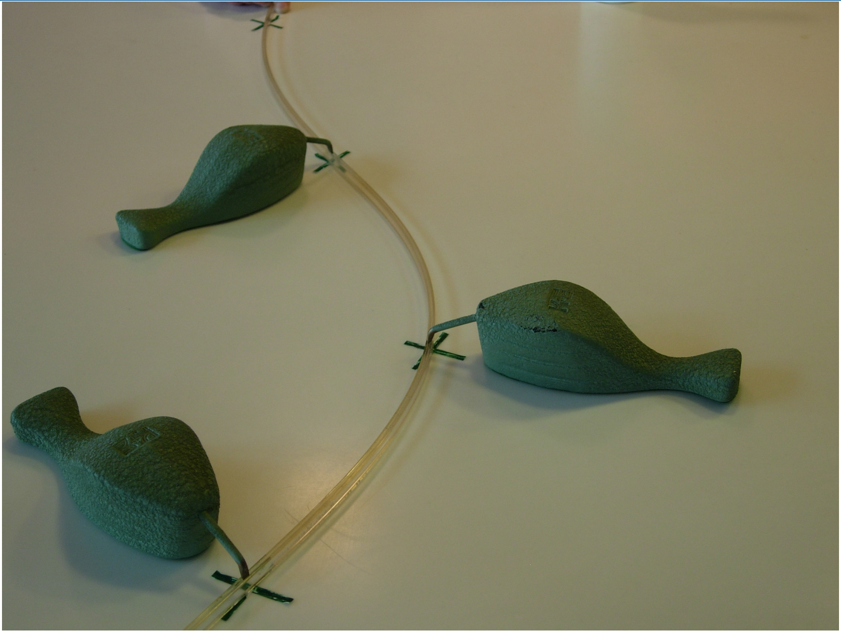
offer a lot of flexibility, without losing the ease and interpretability of linear models.

Nonlinear choices

1. Polynomials (beware)
2. Truncated power basis splines
3. Natural splines
4. B-splines
5. Smoothing splines
6. Radial basis functions
7. Kernel regression
8. Local regression
9. kNN

Outline

Splines



Splines

Splines

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

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.

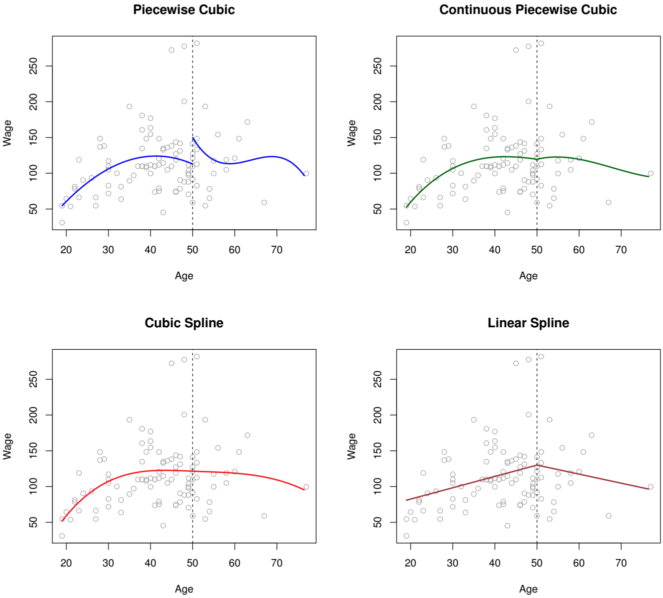
Splines

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

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

- Predictors: $x, \dots, x^p, (x - \kappa_1)_+^p, \dots, (x - \kappa_K)_+^p$
- 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

- 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

- 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

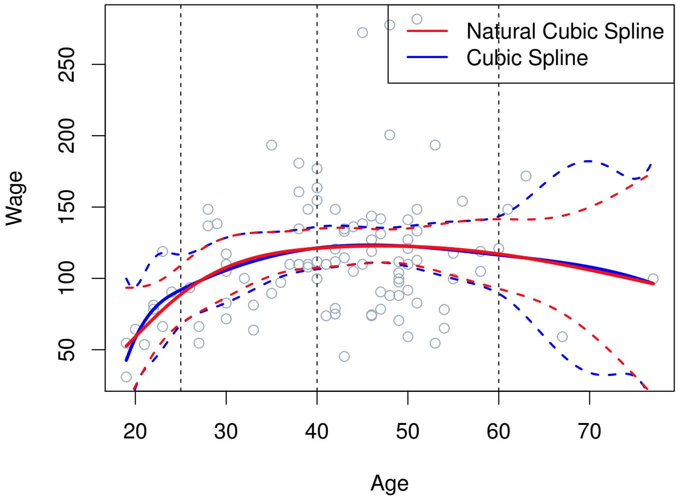
Truncated power basis

- 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

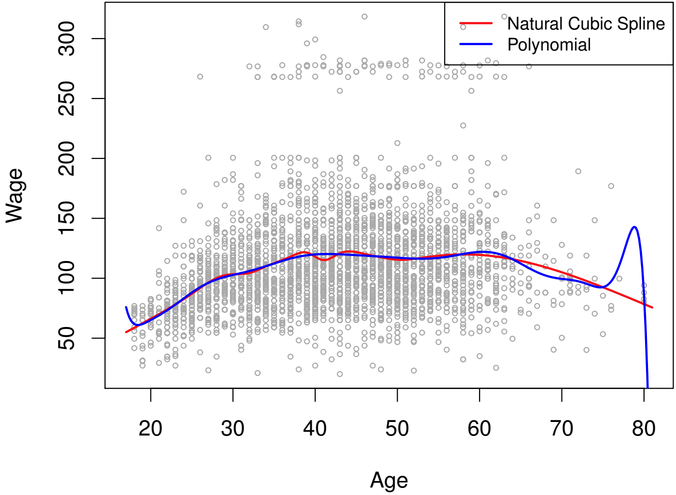
Natural splines

- 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



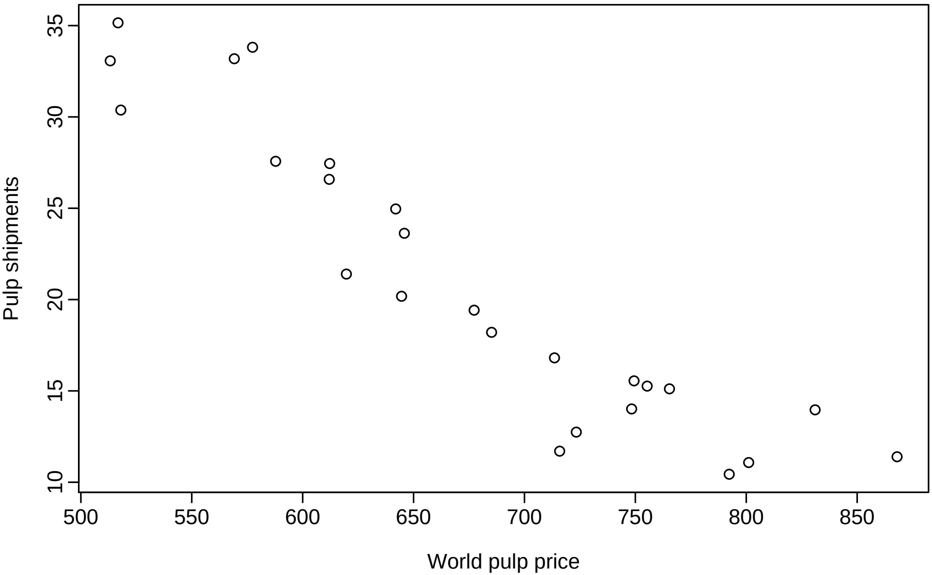
Natural splines



Knot placement

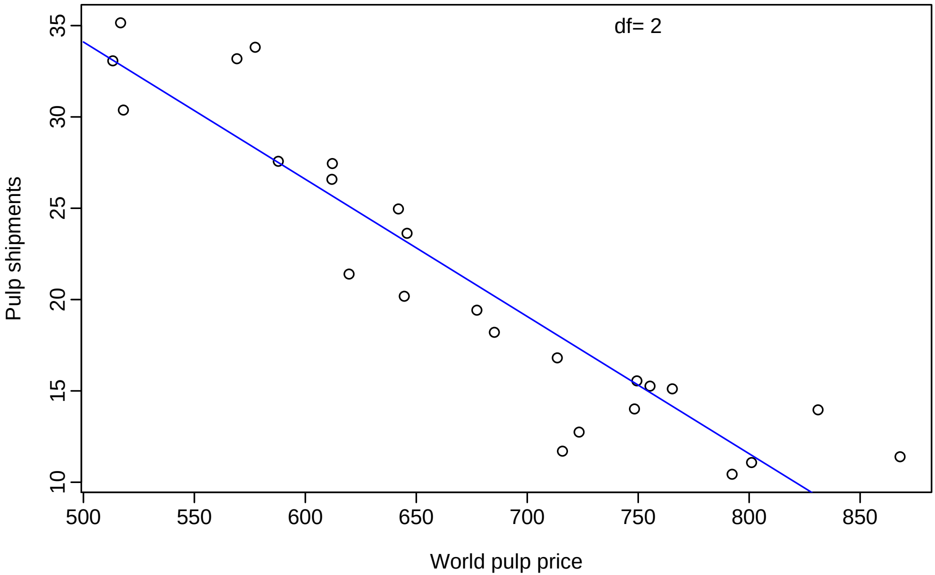
- Strategy 1: specify `df` (equivalently K) and let `ns` place them at appropriate quantiles of the observed X.
- Strategy 2: choose K and their locations.

Splines

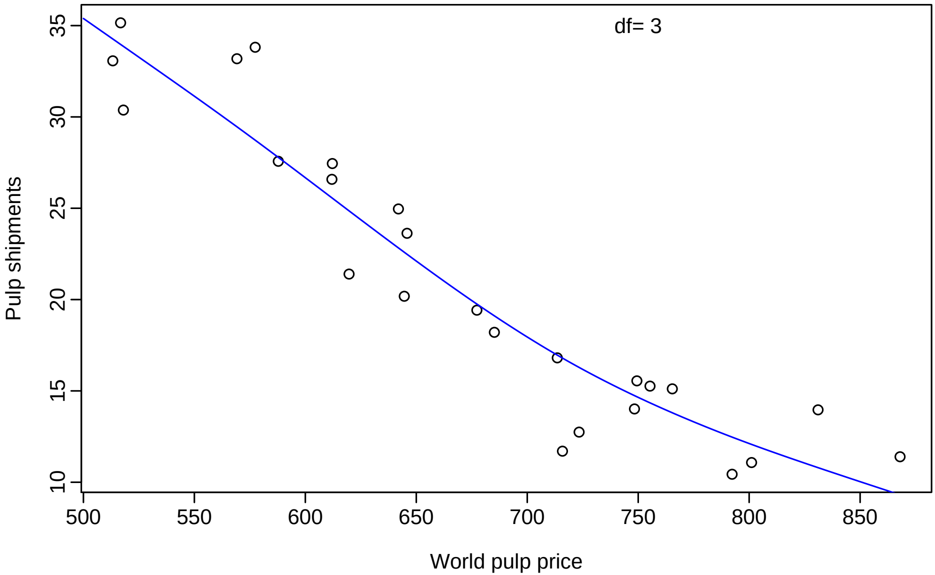


Pulp shipments
World pulp price
10
15
20
25
30
35
500
550
600
650
700
750
800
850

Splines

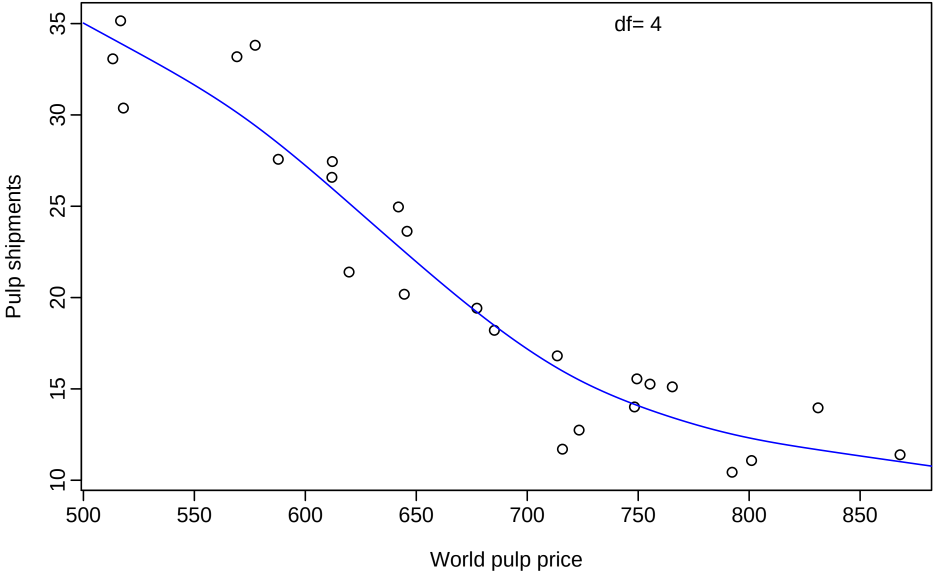

df= 2
Pulp shipments
World pulp price
10
15
20
25
30
35
500
550
600
650
700
750
800
850

Splines

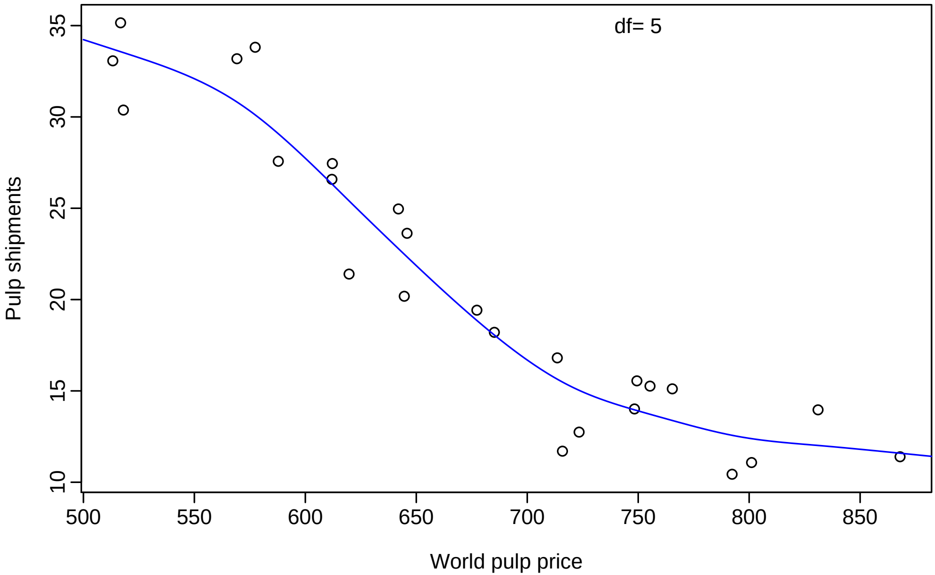


df= 3
Pulp shipments
World pulp price
10
15
20
25
30
35
500
550
600
650
700
750
800
850

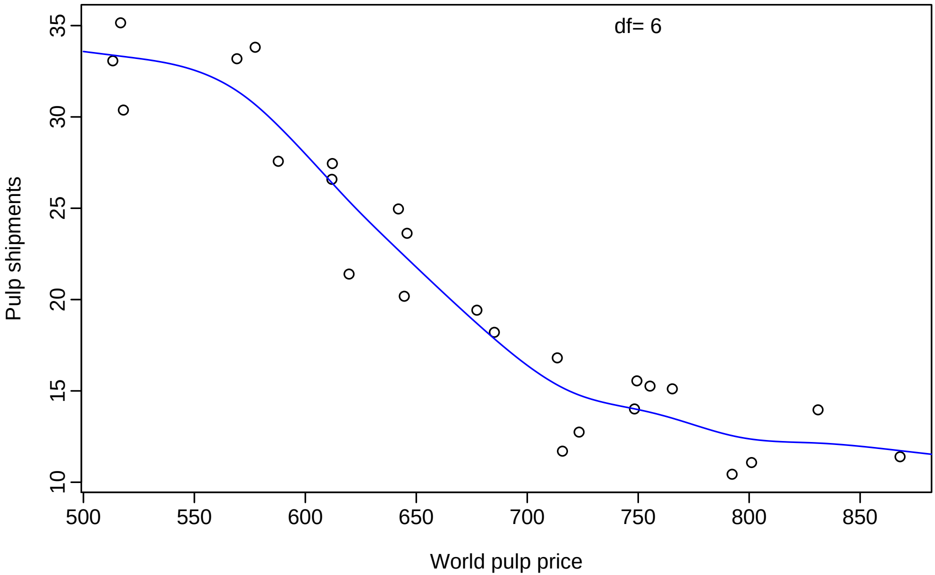
Splines


df= 4
Pulp shipments
World pulp price
10
15
20
25
30
35
500
550
600
650
700
750
800
850

Splines

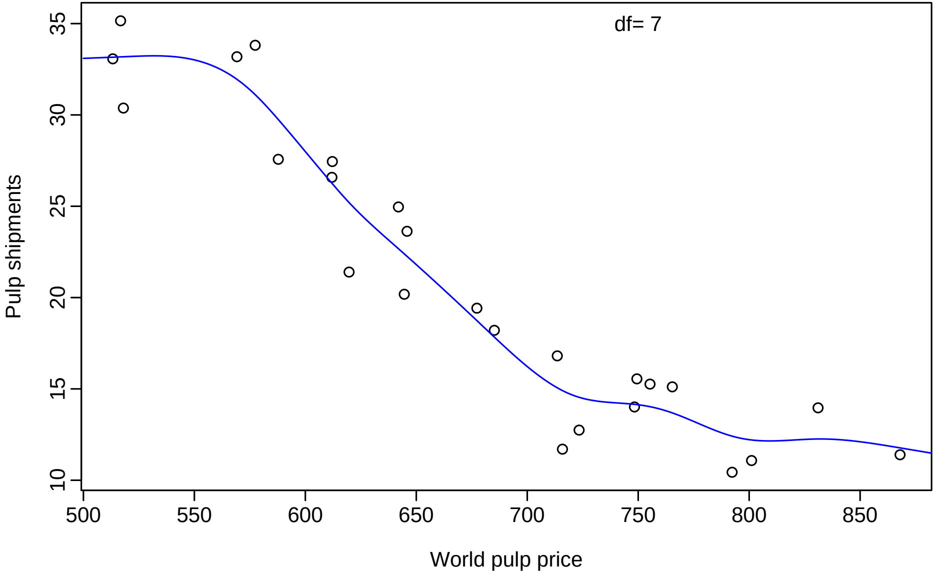

df= 5
Pulp shipments
World pulp price
10
15
20
25
30
35
500
550
600
650
700
750
800
850

Splines



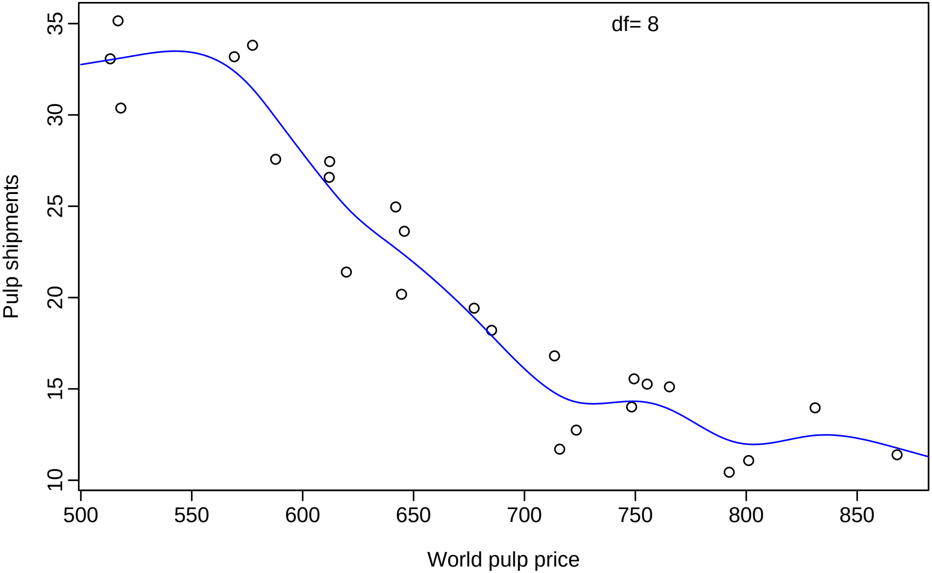
df= 6
Pulp shipments
World pulp price
10
15
20
25
30
35
500
550
600
650
700
750
800
850

Splines



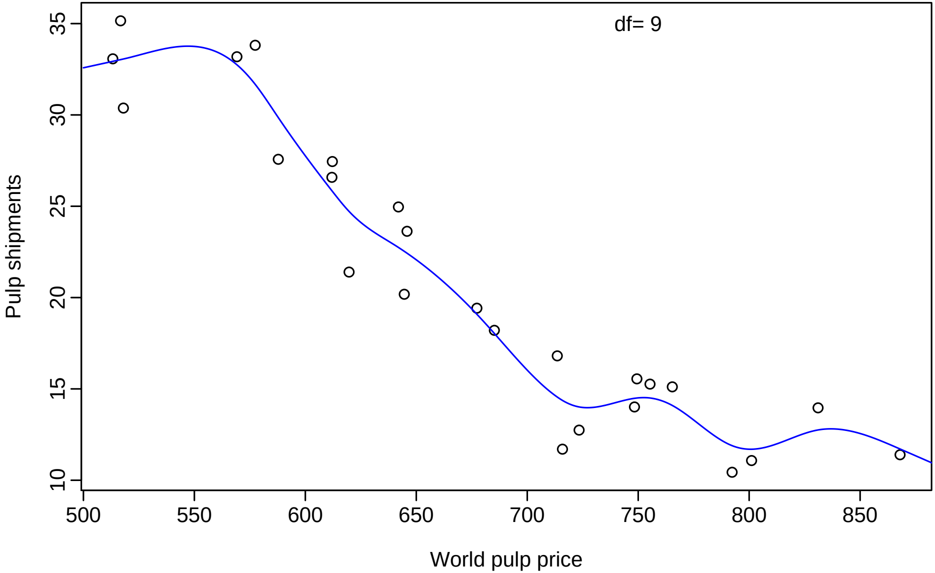
df= 7
Pulp shipments
World pulp price
10
15
20
25
30
35
500
550
600
650
700
750
800
850

Splines

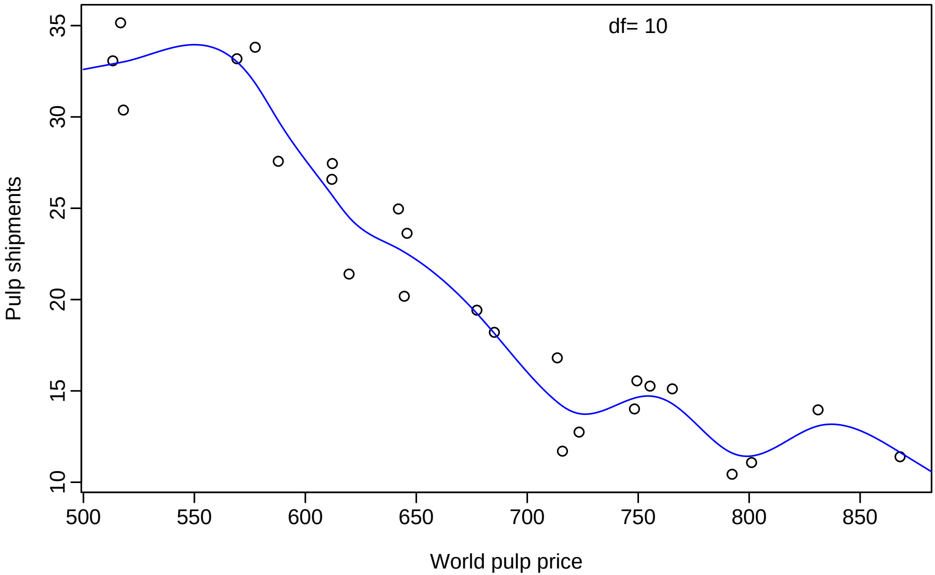


df= 8
Pulp shipments
World pulp price
10
15
20
25
30
35
500
550
600
650
700
750
800
850

Splines

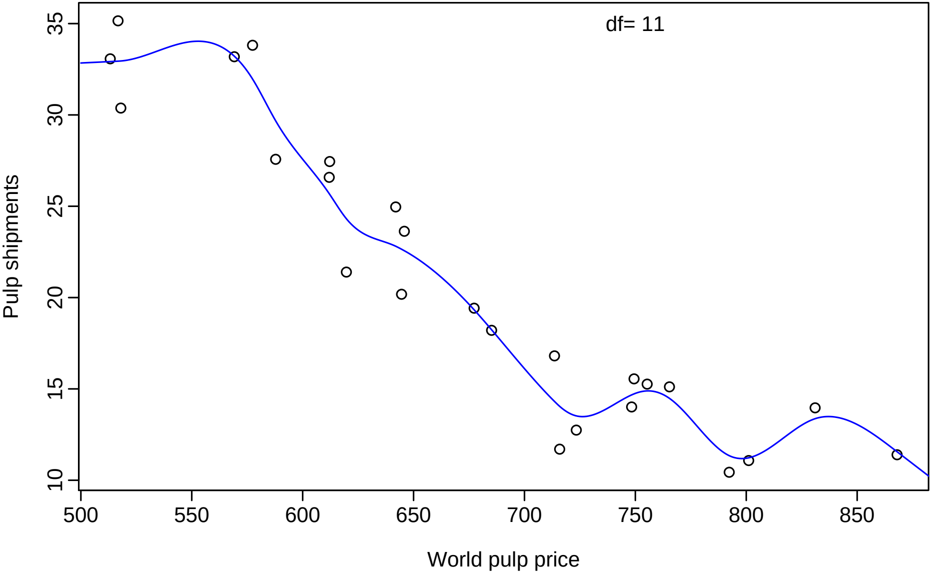

df= 9
Pulp shipments
World pulp price
35
30
25
20
15
10
500
550
600
650
700
750
800
850

Splines

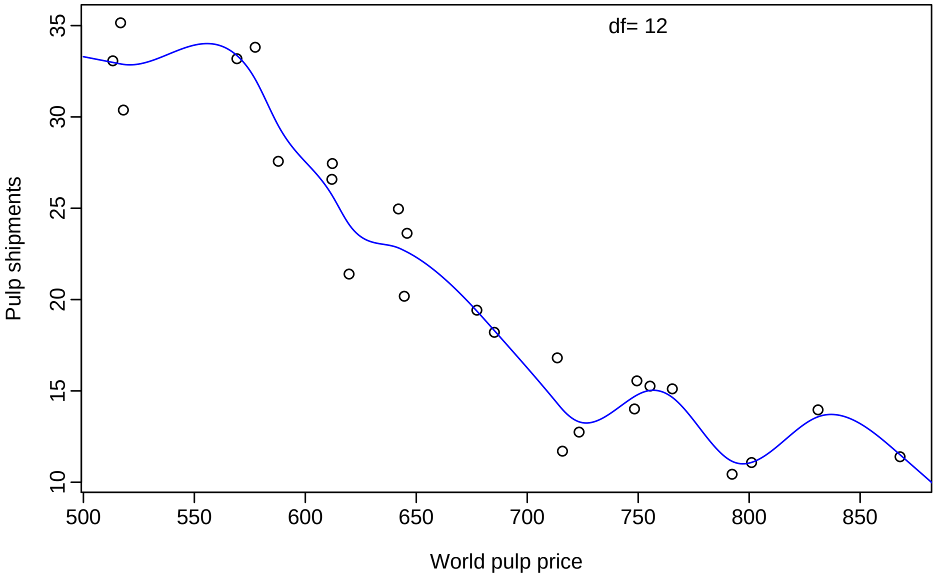


df= 10
Pulp shipments
World pulp price
10
15
20
25
30
35
500
550
600
650
700
750
800
850

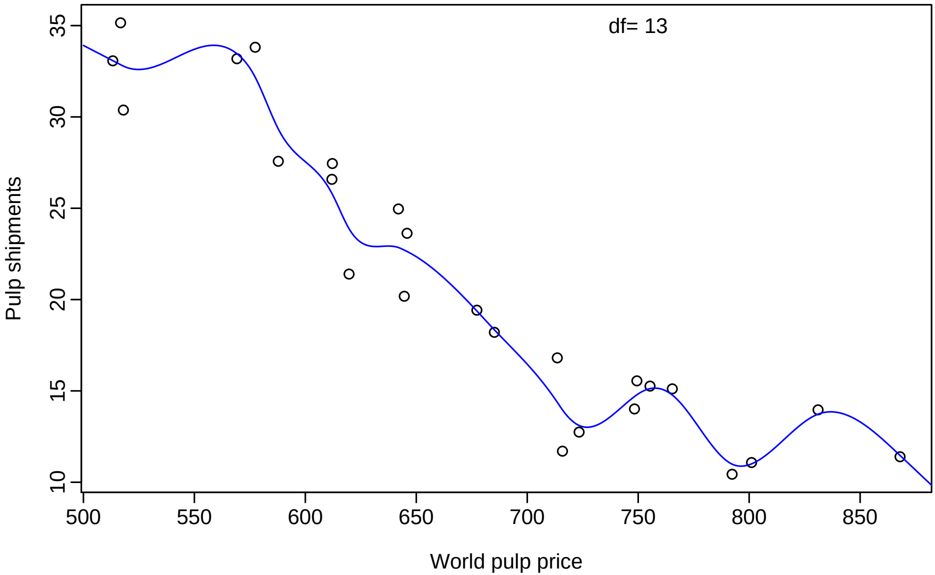
Splines


df= 11
Pulp shipments
World pulp price
10
15
20
25
30
35
500
550
600
650
700
750
800
850

Splines

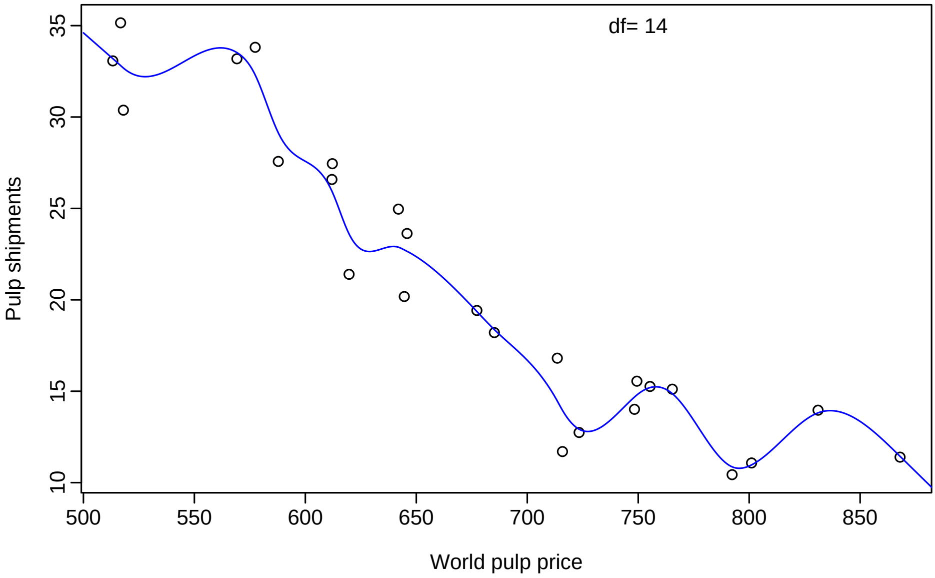

df= 12
Pulp shipments
10
15
20
25
30
35
500
550
600
650
700
750
800
850
World pulp price

Splines



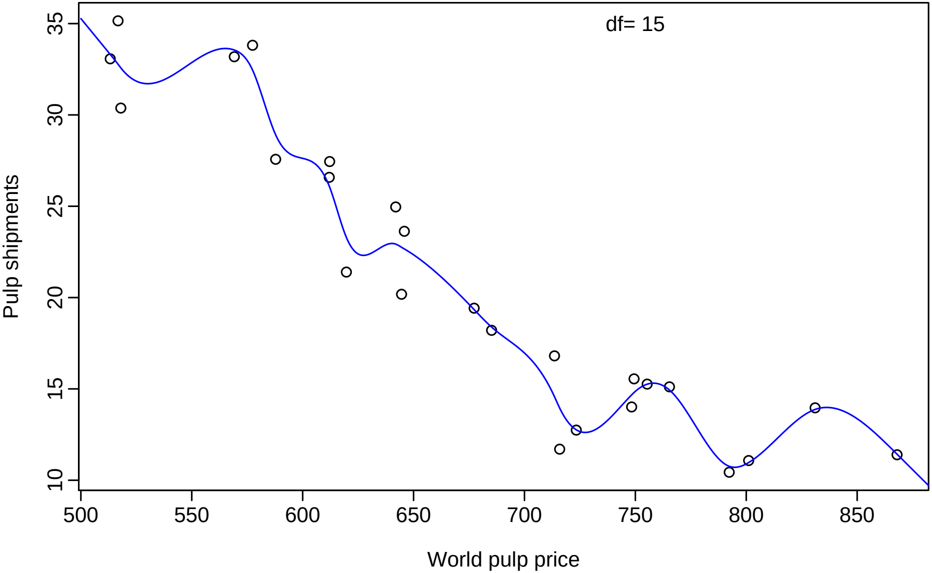
df= 13
Pulp shipments
World pulp price
10
15
20
25
30
35
500
550
600
650
700
750
800
850

Splines



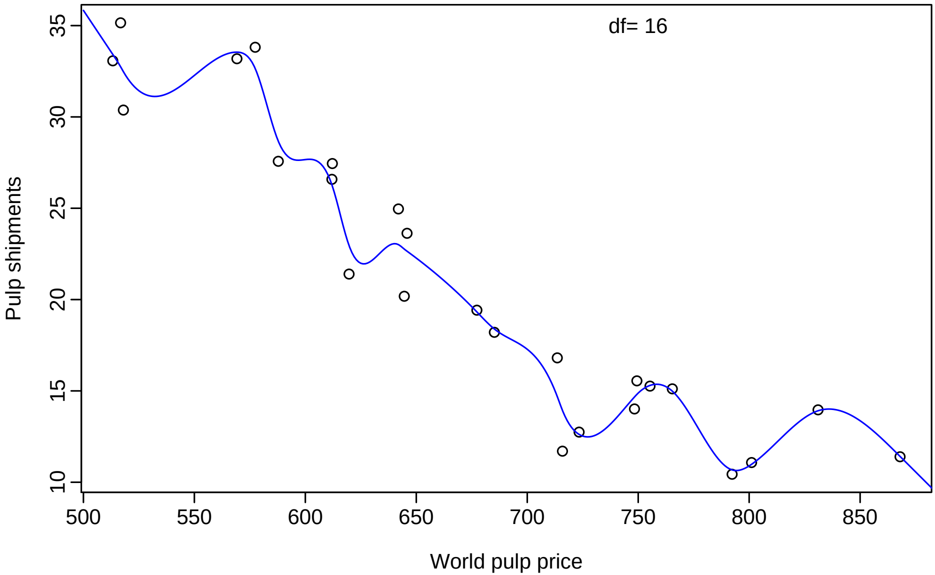
df= 14
Pulp shipments
World pulp price
10
15
20
25
30
35
500
550
600
650
700
750
800
850

Splines



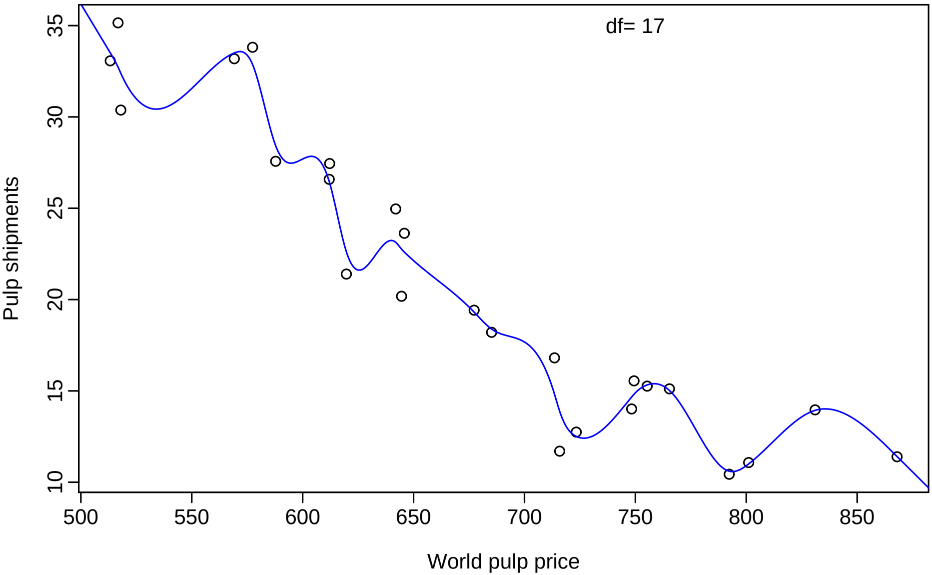
df= 15
Pulp shipments
World pulp price
10
15
20
25
30
35
500
550
600
650
700
750
800
850

Splines



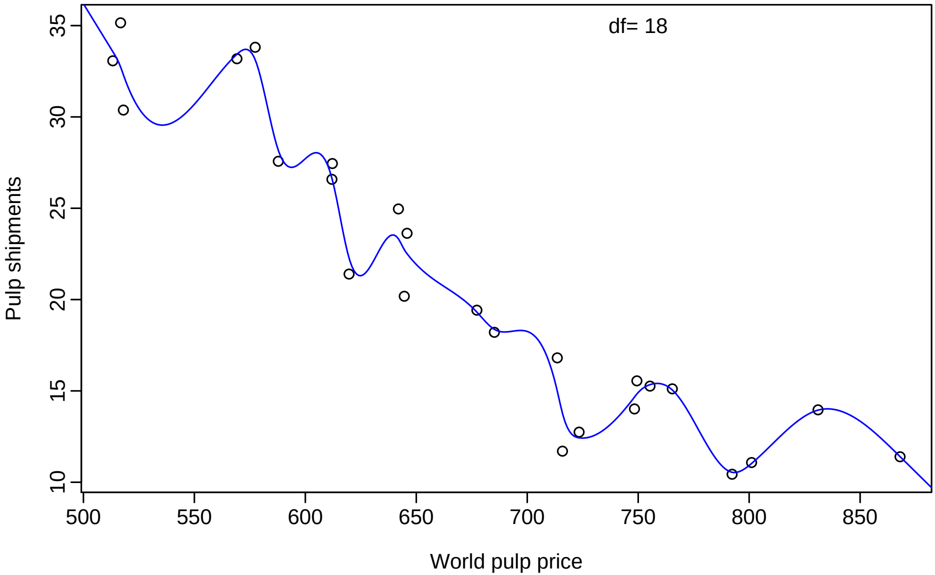
df= 16
Pulp shipments
World pulp price
10
15
20
25
30
35
500
550
600
650
700
750
800
850

Splines

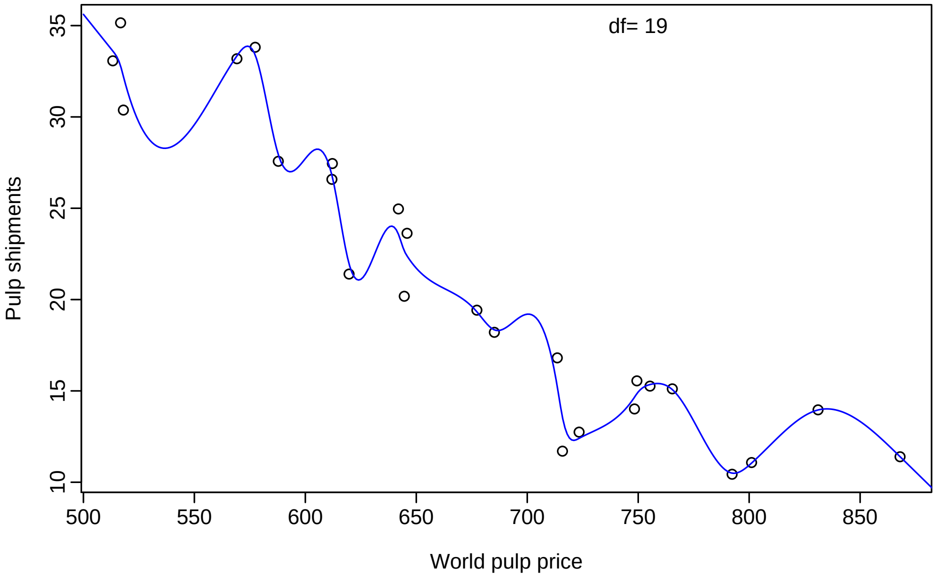


df= 17
Pulp shipments
World pulp price
10
15
20
25
30
35
500
550
600
650
700
750
800
850

Splines

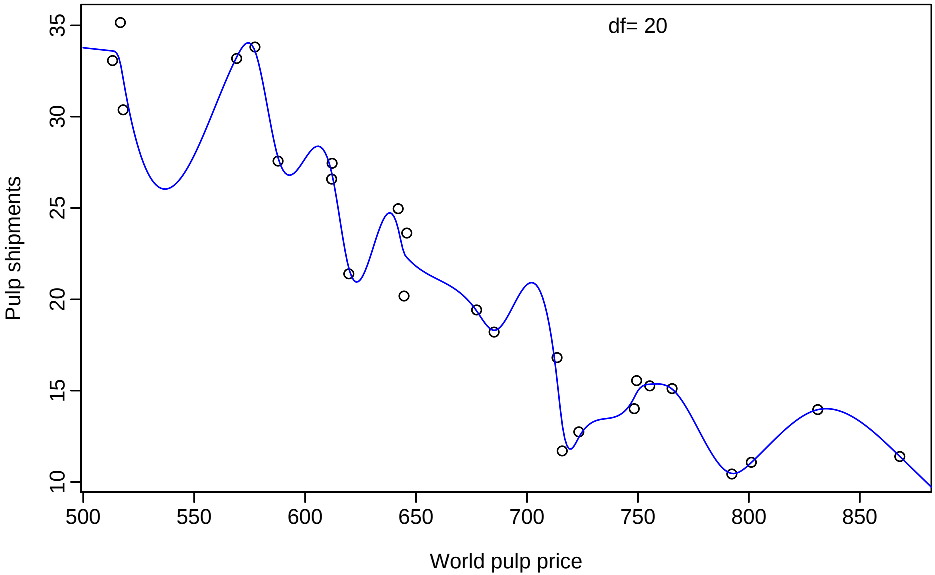

df= 18
Pulp shipments
10
15
20
25
30
35
500
550
600
650
700
750
800
850
World pulp price

Splines



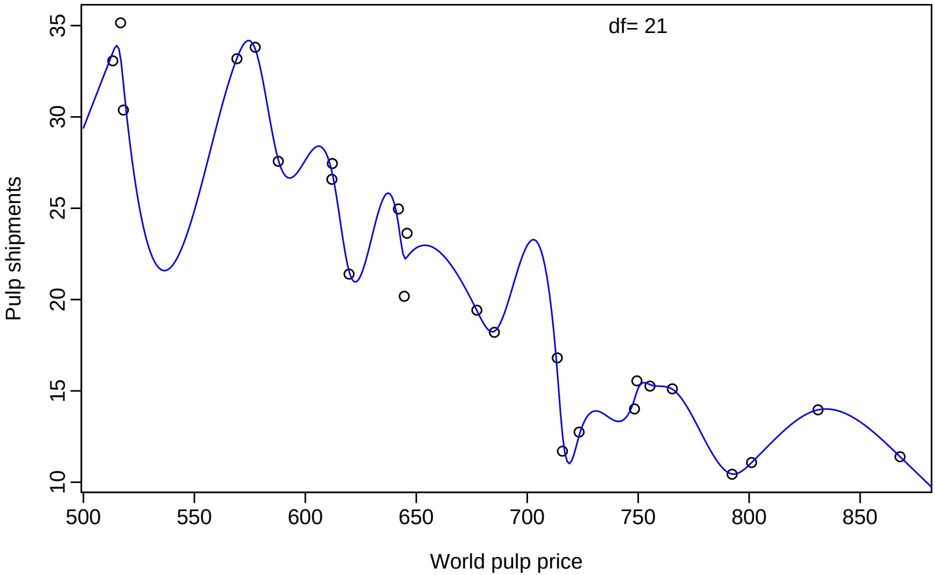
df= 19
Pulp shipments
World pulp price
10
15
20
25
30
35
500
550
600
650
700
750
800
850

Splines



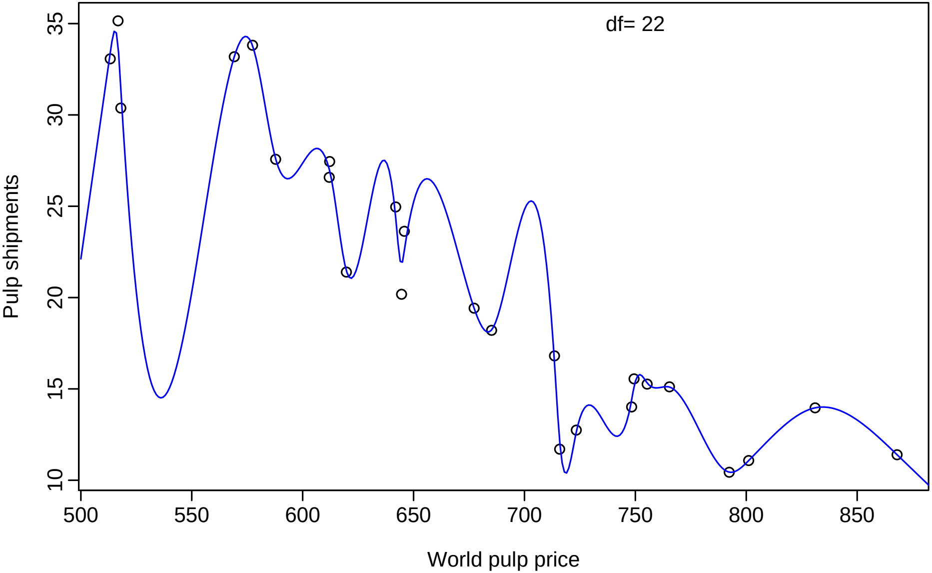
df= 20
Pulp shipments
World pulp price
10
15
20
25
30
35
500
550
600
650
700
750
800
850

Splines



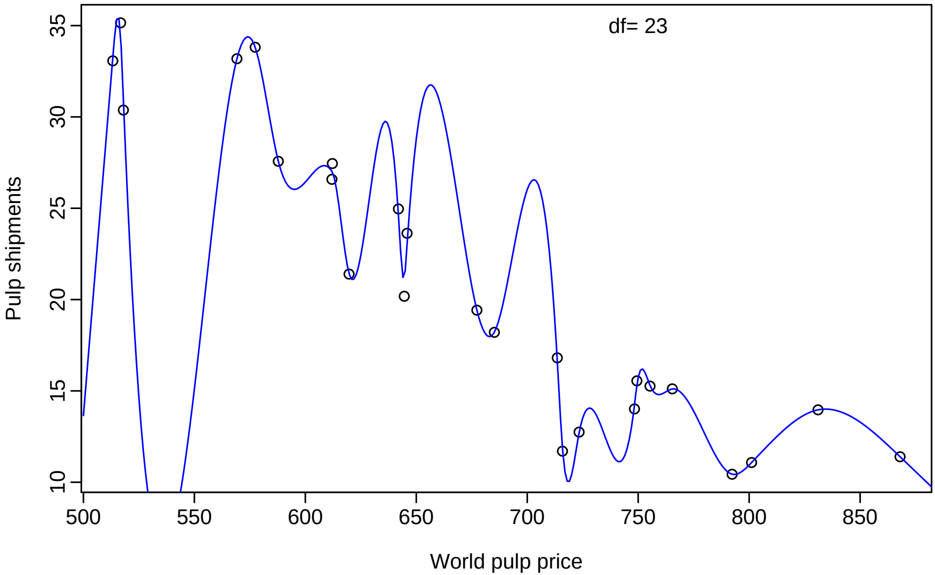
df= 21
Pulp shipments
World pulp price
10
15
20
25
30
35
500
550
600
650
700
750
800
850

Splines



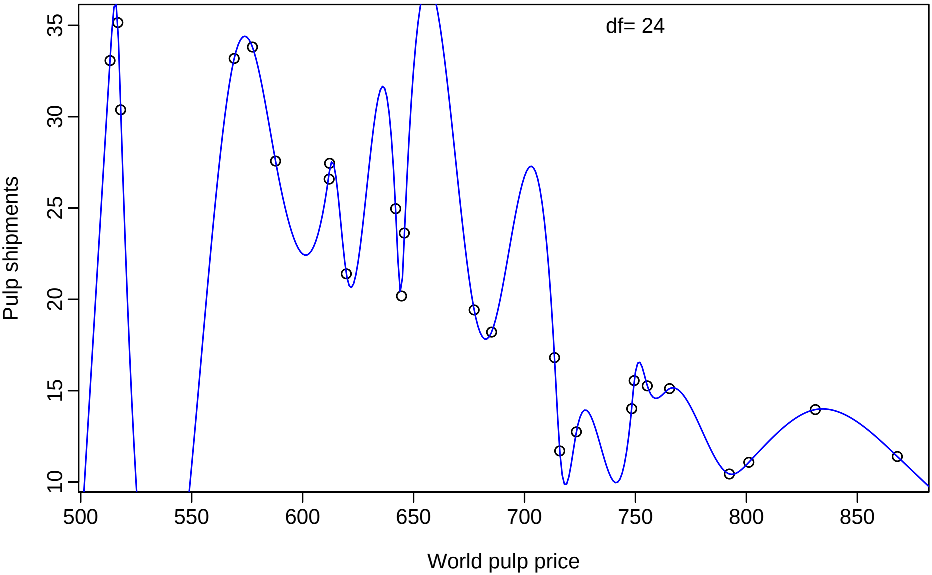
df= 22
Pulp shipments
World pulp price
10
15
20
25
30
35
500
550
600
650
700
750
800
850

Splines

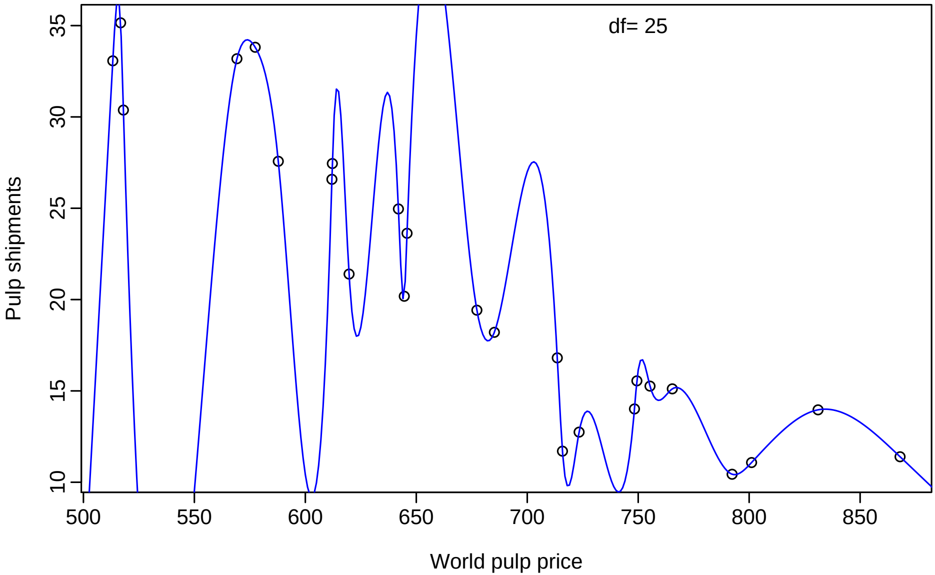


df= 23
Pulp shipments
World pulp price
10
15
20
25
30
35
500
550
600
650
700
750
800
850

Splines

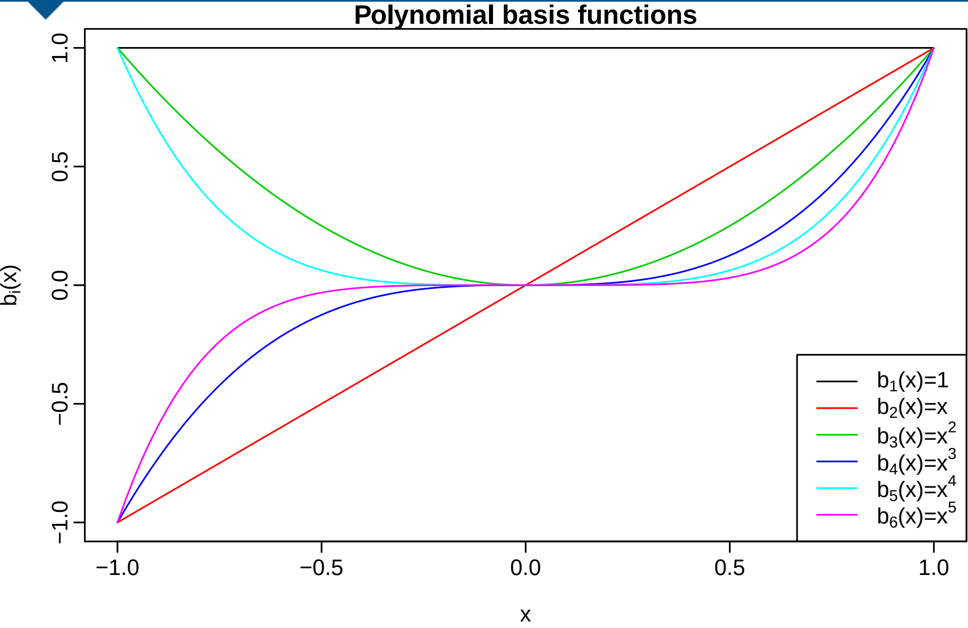

df= 24
Pulp shipments
10
15
20
25
30
35
500
550
600
650
700
750
800
850
World pulp price

Splines

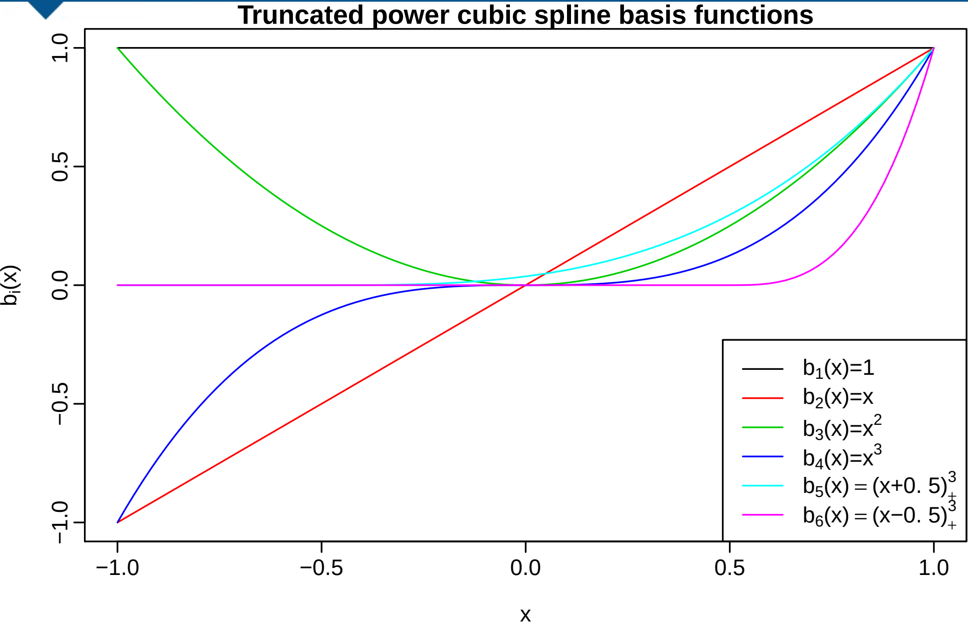


df= 25
Pulp shipments
World pulp price
10
15
20
25
30
35
500
550
600
650
700
750
800
850

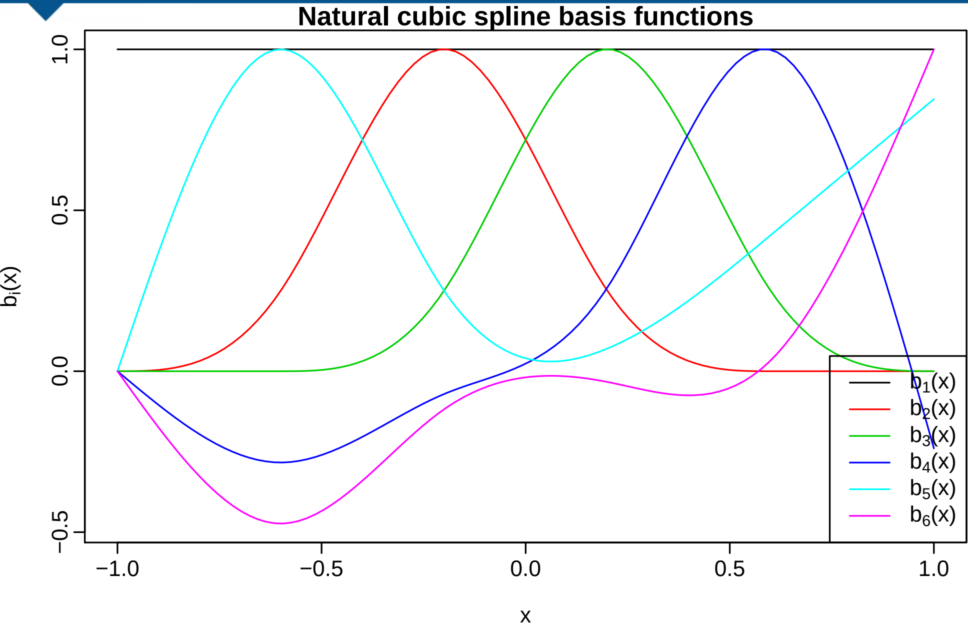
Basis functions



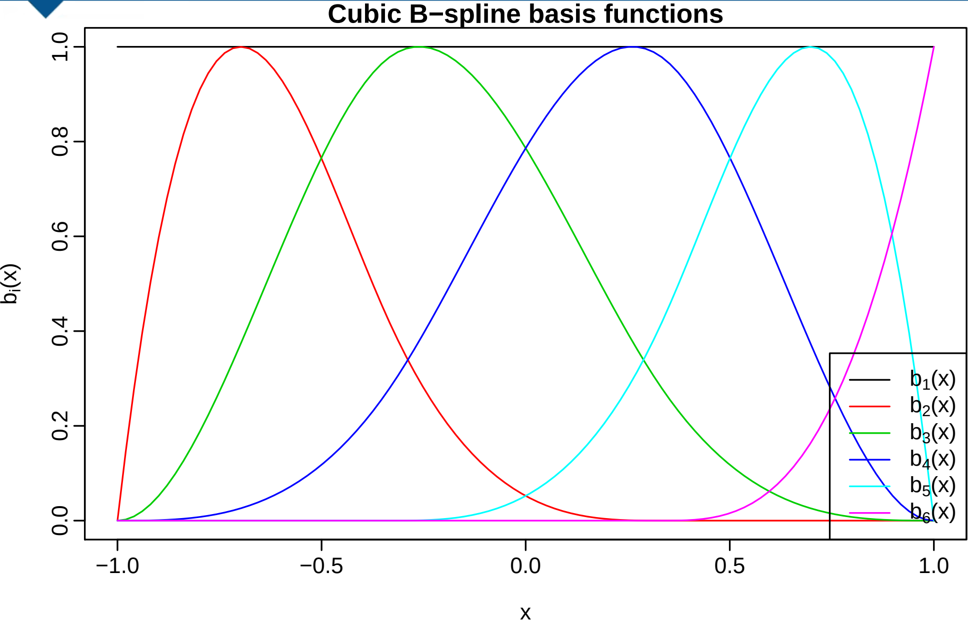
Basis functions



Basis functions



Basis functions



Outline

The curse of dimensionality

Why can't we 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 almost impossible to estimate.

The curse of dimensionality

Why can't we 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 almost impossible 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

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)
```

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)
```