ETC3250: Flexible Regression

Semester 1, 2020

Professor Di Cook

Econometrics and Business Statistics
Monash University

Week 2 (b)

Moving beyond linearity

Sometimes the relationships we discover are not linear...

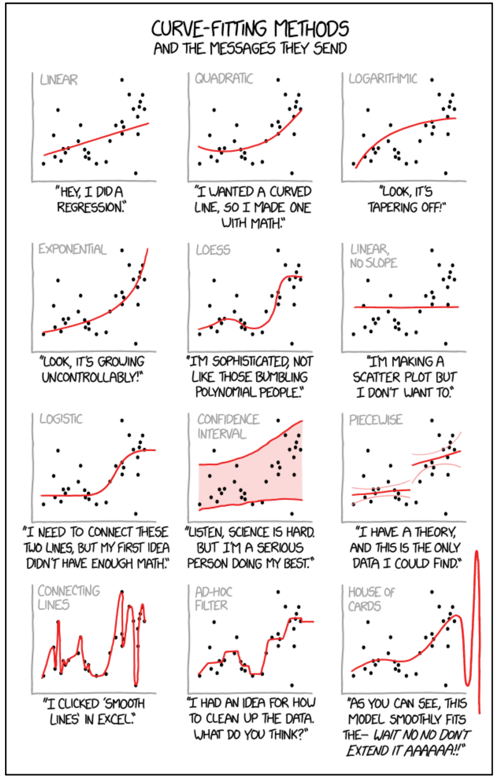
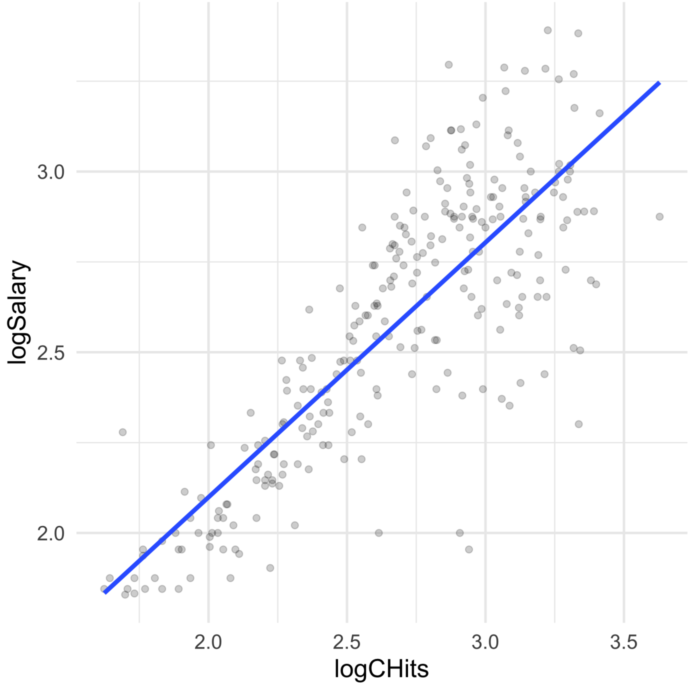


Image source: XKCD

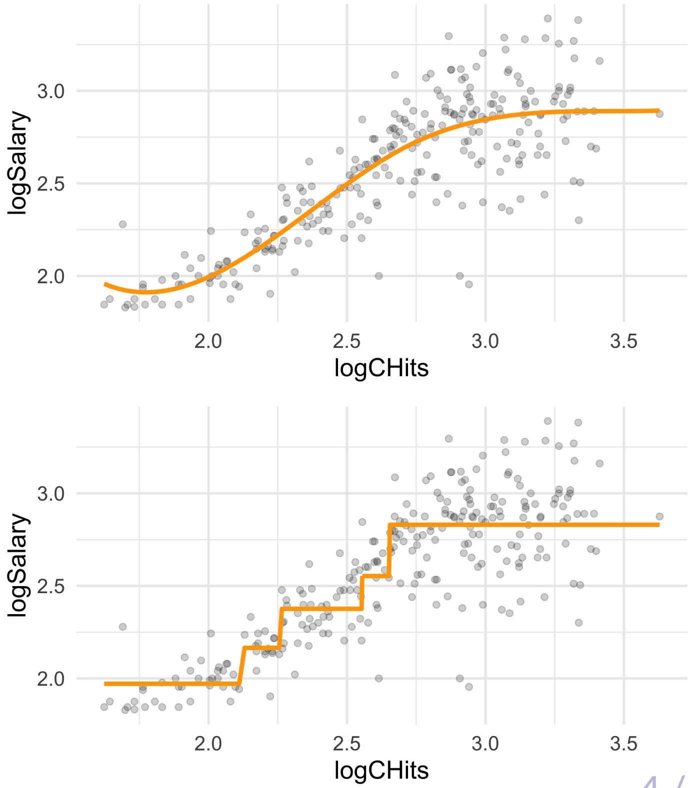
Moving beyond linearity

- Consider the following Major League Baseball data from the 1986 and 1987 seasons.
- Would a linear model be appropriate for modelling the relationship between Salary and Career hits, captured in the variables `logSalary` and `logCHits`?



Moving beyond linearity

📊 Perhaps a more flexible regression model is needed!



Flexible regression fits

The truth is rarely linear, 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.

Polynomial basis functions

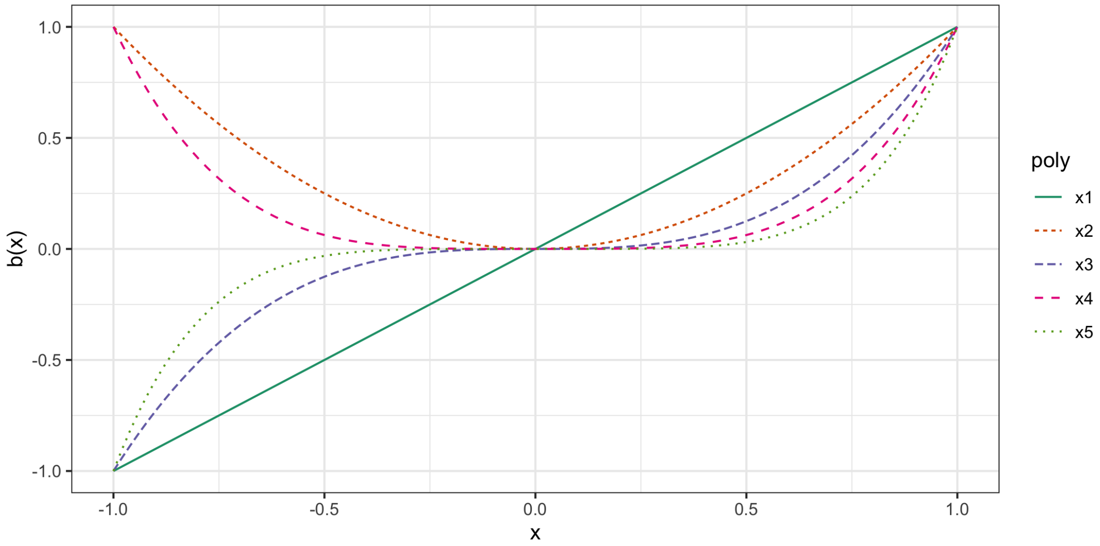
Instead of fitting a linear model (in X), we fit the model

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

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: $b_k(x_i) = I(c_k \leq x_i \leq c_{k+1})$

Polynomial basis functions



x1 $= x$, x2= x^2, x3= x^3, x4= x^4, x5= x^5

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 $x = \kappa_{j+1}$.

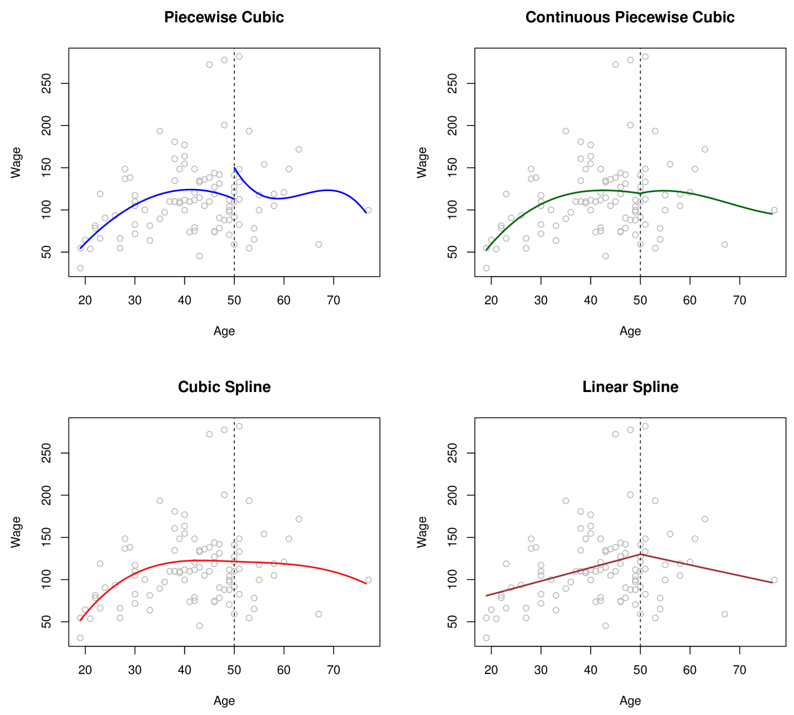
- Parameters constrained so that $f(x)$ is continuous.
- Further constraints imposed to give continuous derivatives.

Piecewise Cubic Poly Spline

Piecewise cubic polynomial with a single knot at a point c:

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

Piecewise Poly



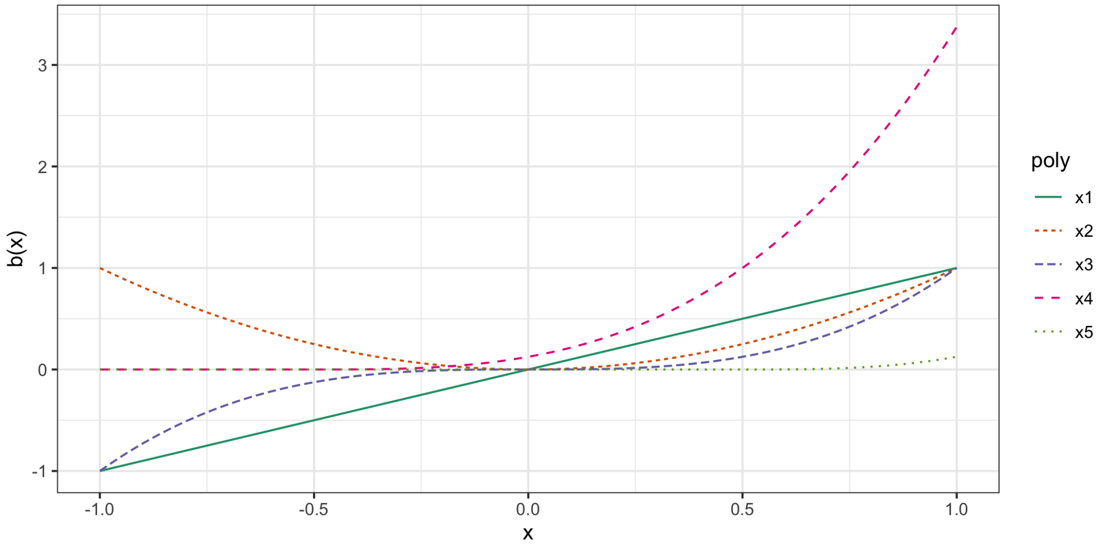
Basis Functions

- Truncated power basis
- Predictors: $x, ..., x^p, (x - \kappa_1)_+^p, ..., (x - \kappa_K)_+^p$

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

Basis functions



x1 $= x$, x2= x^2, x3= x^3, x4= $(x + 0.5)^3_+$, x5= $(x - 0.5)^3_+$

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.

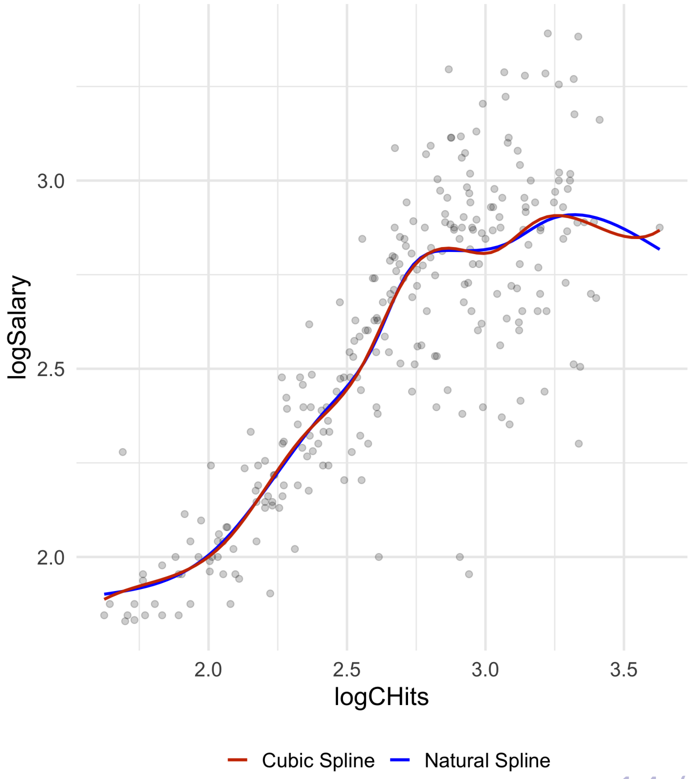
Degrees of freedom `df` $= K$.

Create predictors using `ns` function in R (automatically chooses knots given `df`).

Comparison with Cubic splines

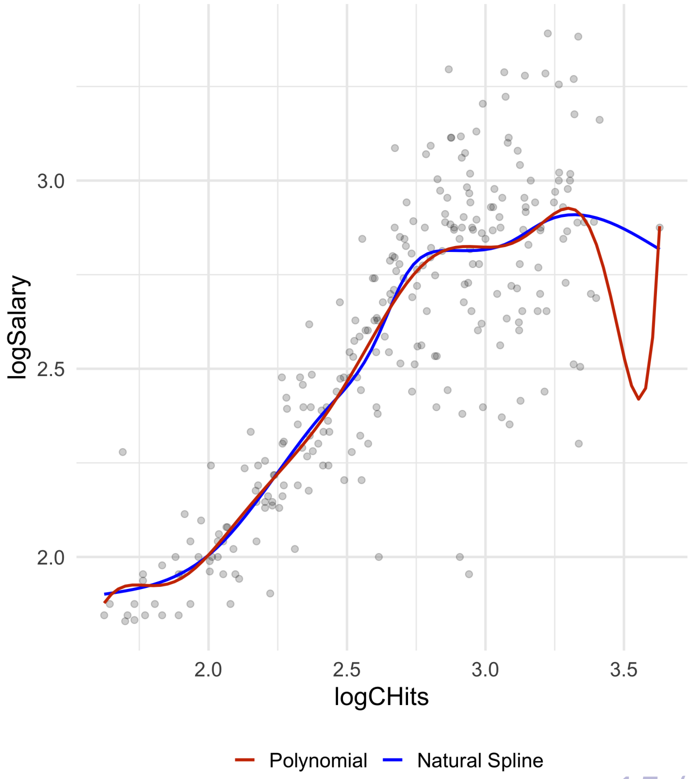
We can fit a cubic spline in `R` using `splines::bs()`, and fit a natural cubic spline using `splines::ns()`.

📊 Notice the difference between the fits towards the end of the curves.

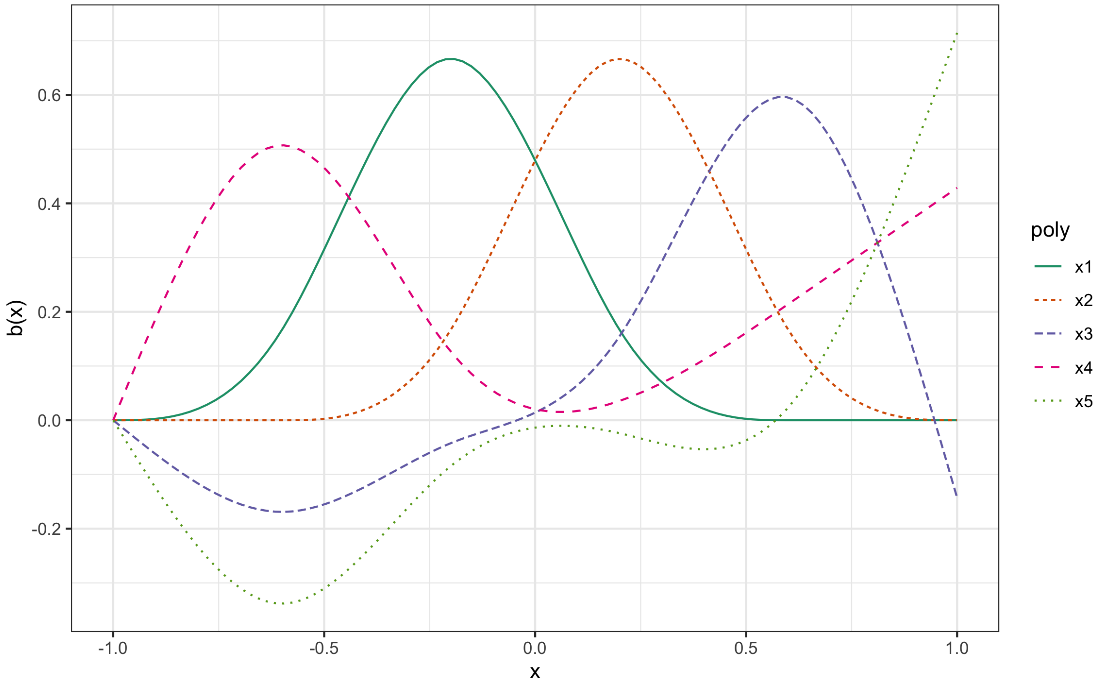


Comparison with Polynomial Regression

📊 Notice the difference between the fits towards the end of the curves.



Natural cubic splines

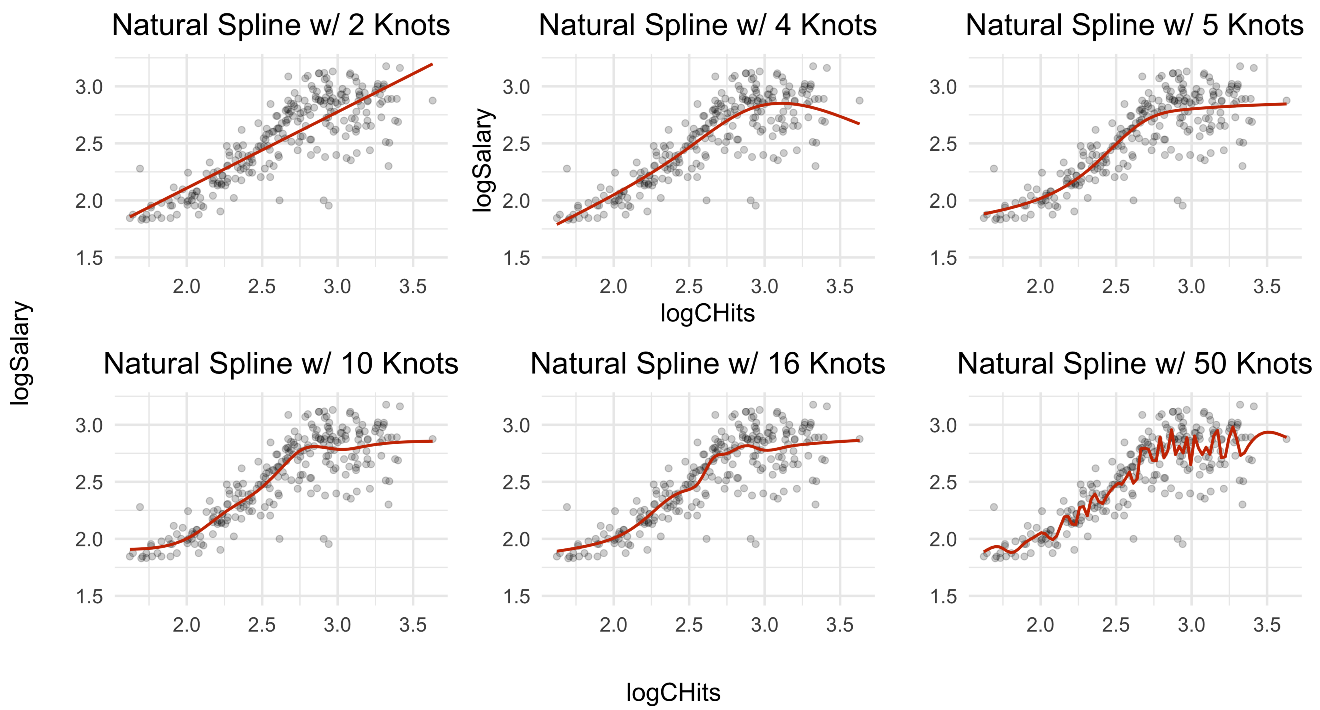


Knot placement

Strategy 1: specify `df` (which creates `df-1` internal knots and 2 boundary knots, so that `df` = $K + 1$) and let `ns()` place them at appropriate quantiles of the observed X.

Strategy 2: choose K and their locations.

Natural cubic splines with differing knots



Generalised additive models (GAMs)

Why is it hard to fit models of the form

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

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

Additive functions

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

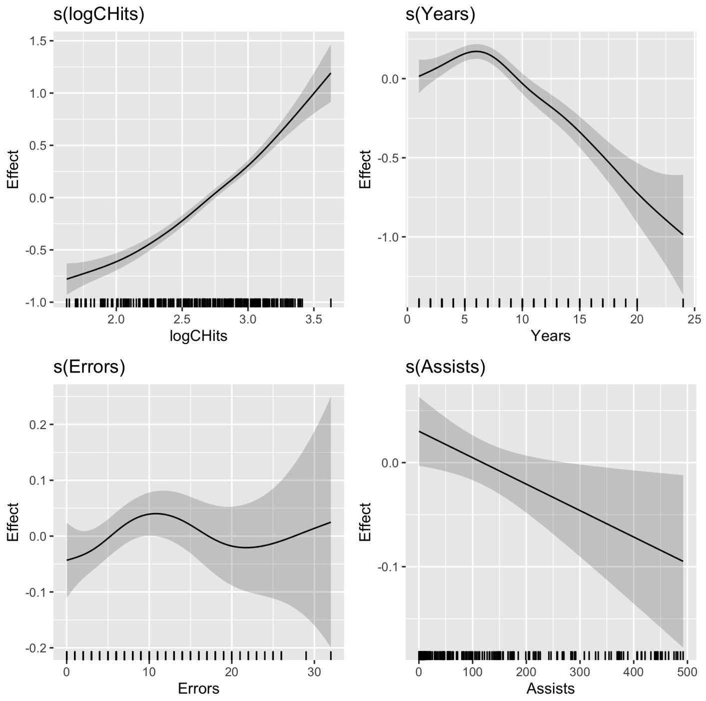
where each f is a smooth univariate function.

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

Additive functions

$$\log(\text{Salary}) = \beta_0 + f_1(\log(\text{CHits})) + f_2(\text{Years}) + f_3(\text{Errors}) + f_4(\text{Assists}) + \varepsilon$$

```
my_gam <- gam(logSalary~s(logCHits) +
              s(Years)+ s(Errors) +
              s(Assists),data = hits)
```



Generalisations

- Can fit a GAM simply using, e.g. natural splines:
- Coefficients not that interesting; fitted functions are.
- Use `draw` from `gratia` package to plot GAMs fitted in `mgcv` 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)`.

Can we include interaction effects?

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

🧑‍💻 Made by a human with a computer

Slides at https://iml.numbat.space.

Code and data at https://github.com/numbats/iml.

Created using R Markdown with flair by **xaringan**, and **kunoichi** (female ninja) style.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.