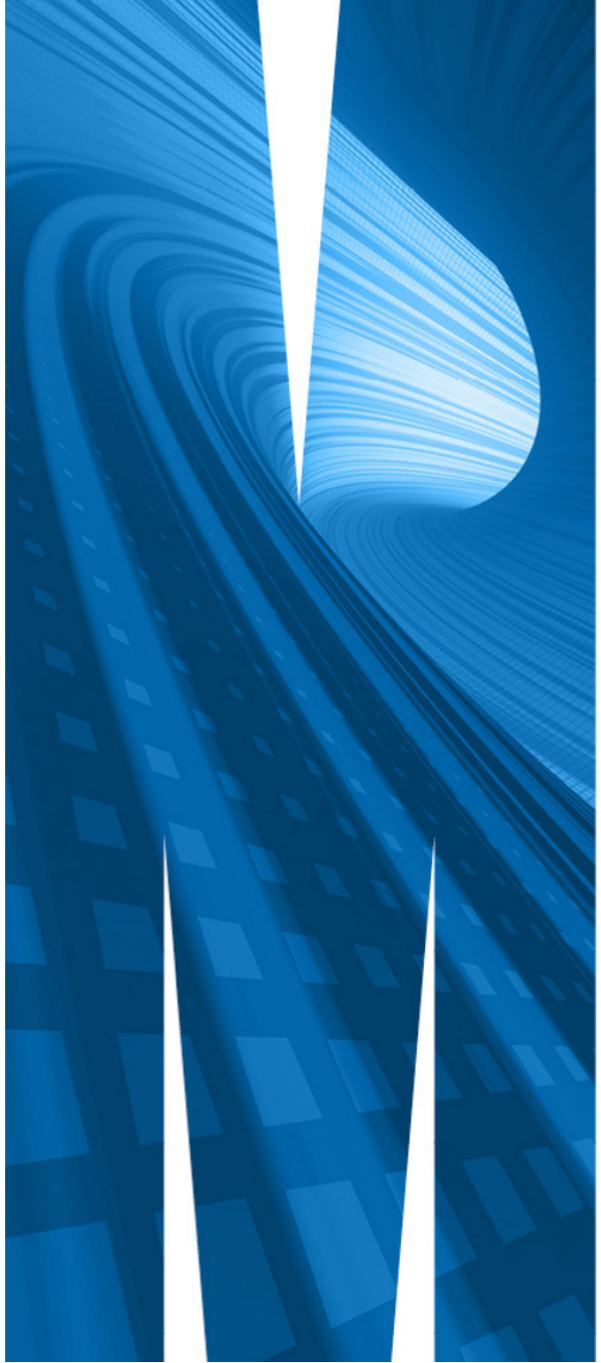
MONASH University

ETC3250/5250: Introduction to Machine Learning

Flexible regression



Lecturer: *Professor Di Cook*

Department of Econometrics and Business Statistics

ETC3250.Clayton-x@monash.edu

Week 2b

Moving beyond linearity

Sometimes the relationships we discover are not linear...

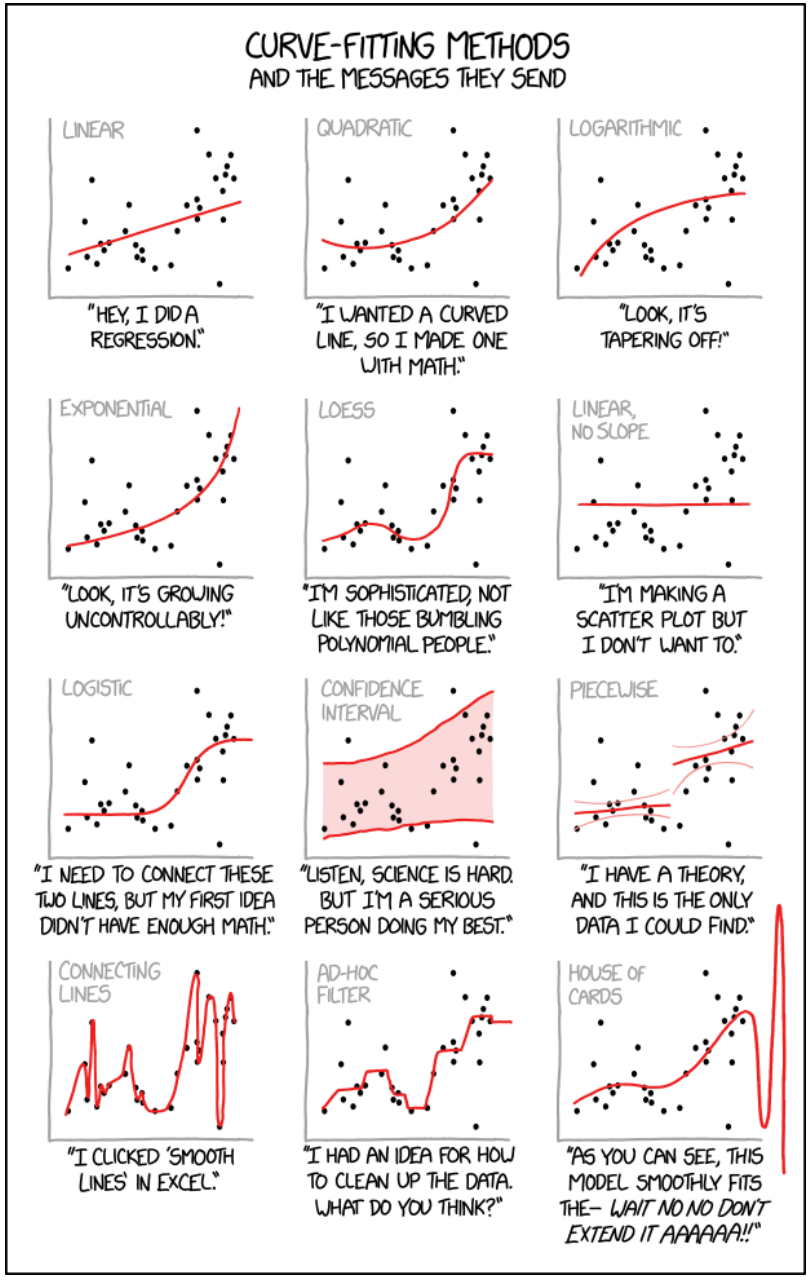
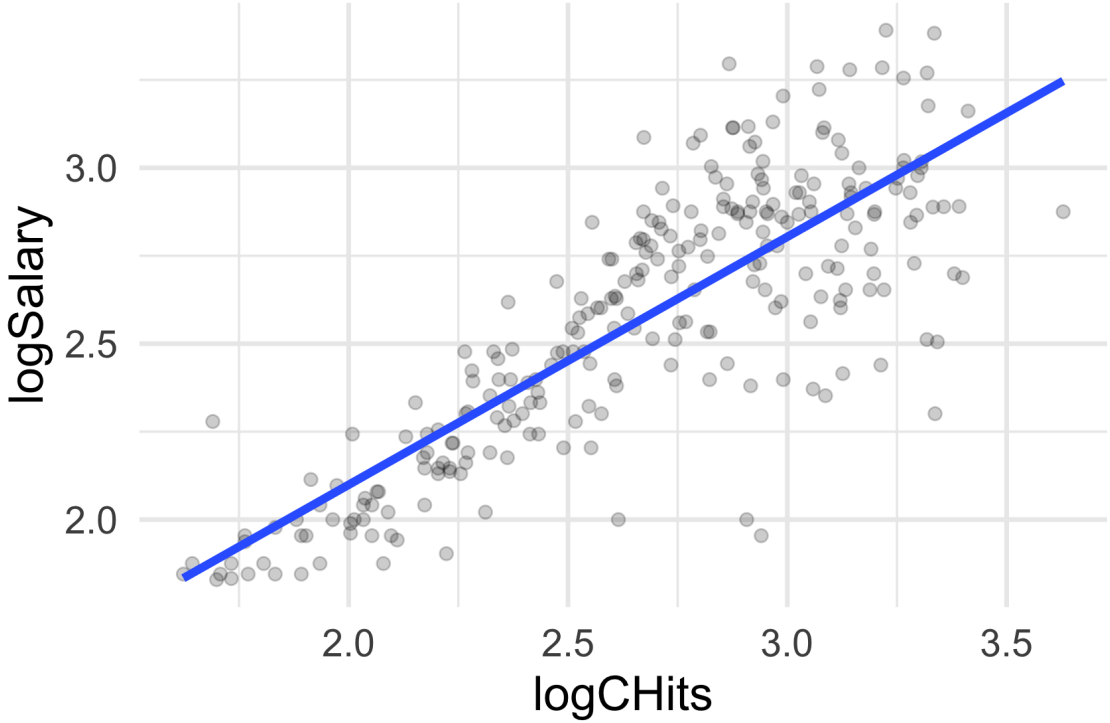


Image source: XKCD

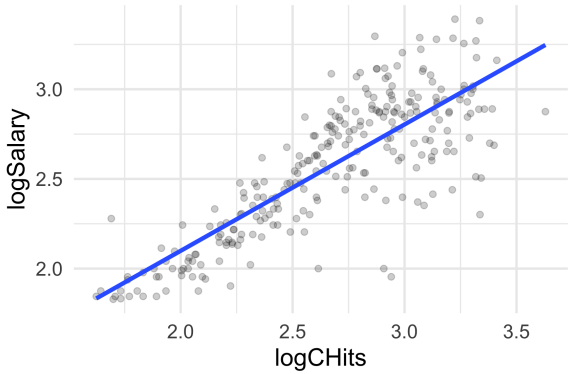
Moving beyond linearity

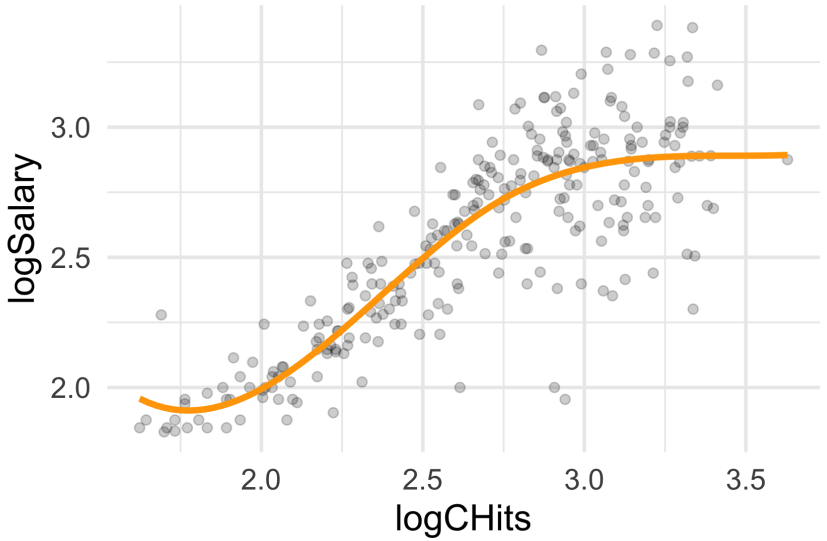
- Consider the following Major League Baseball data from the 1986 and 1987 seasons.
- Would a linear model be appropriate for modelling the relationship between Salary and Career hits, captured in the variables `logSalary` and `logCHits`?

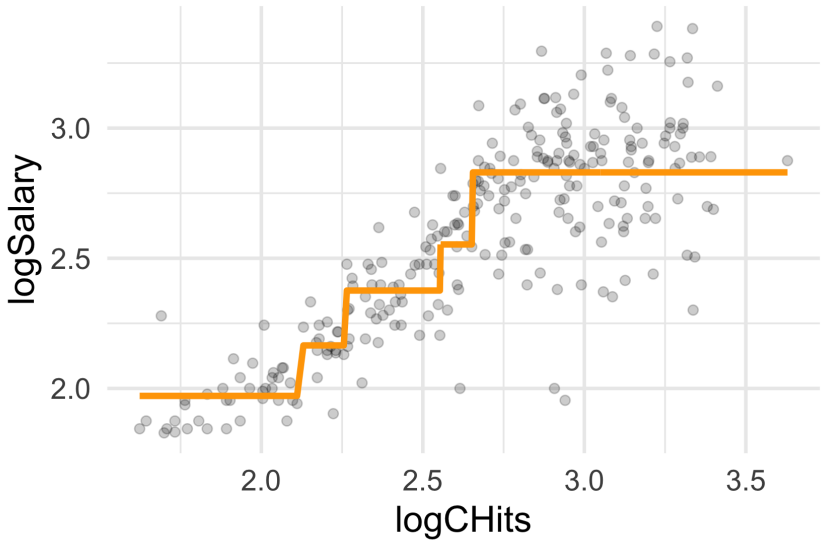


Moving beyond linearity

- Perhaps a more flexible regression model is needed!
- Which of these is a better fit for this data, do you think?







Flexible regression fits

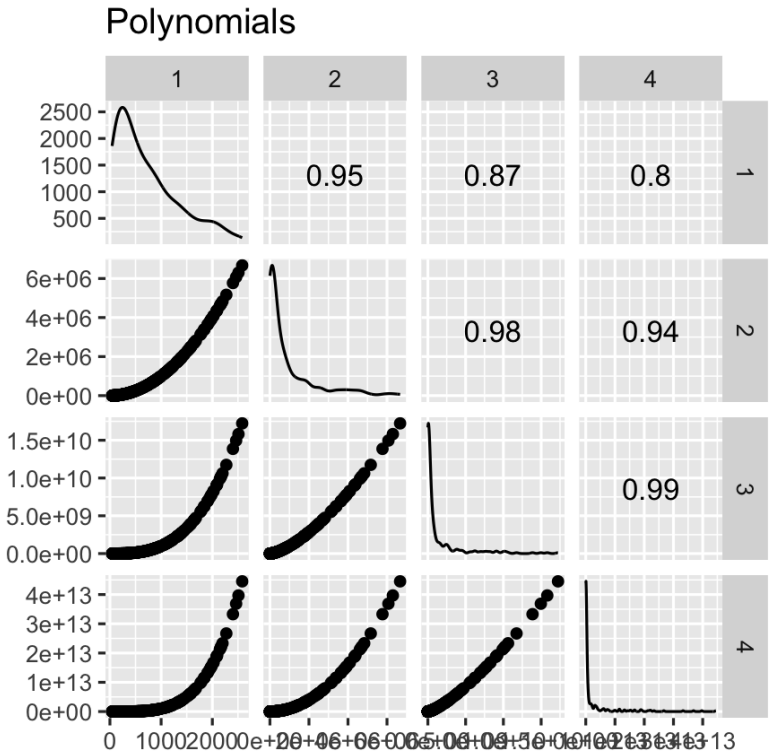
The truth is rarely linear, but often the linearity assumption is sufficient and simple. When it's not ...

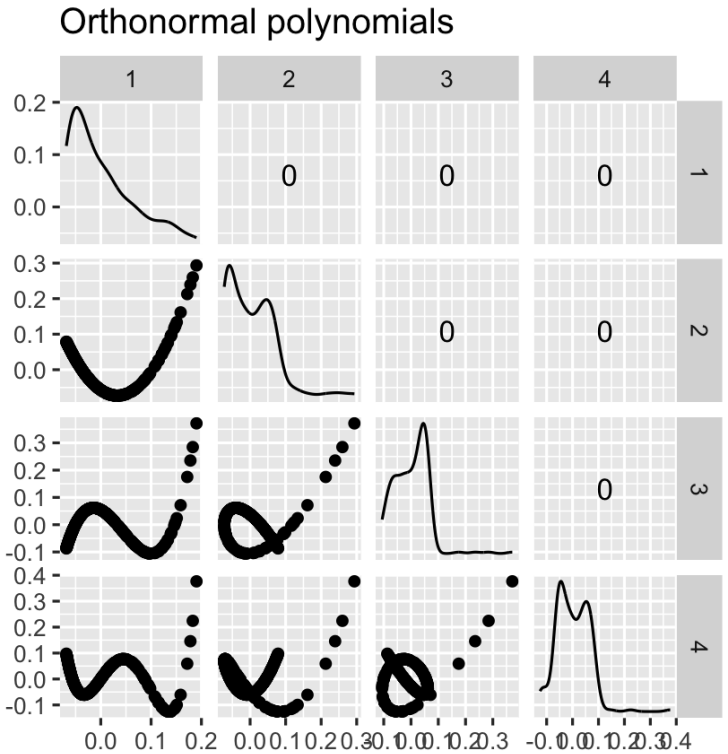
- polynomial regression, obtained by raising each of the original predictors to a power;
- step functions, cut the range of a predictor into distinct regions;
- regression splines, combine polynomials and step functions fit different functions to different subsets of a predictor;
- smoothing splines, regression splines plus a smoothness penalty;
- local regression, splines where the regions are allowed to overlap;
- generalized additive models, extend these approaches to multiple predictors.

offer a lot of flexibility, while maintaining the ease and interpretability of linear models.

Polynomial regression

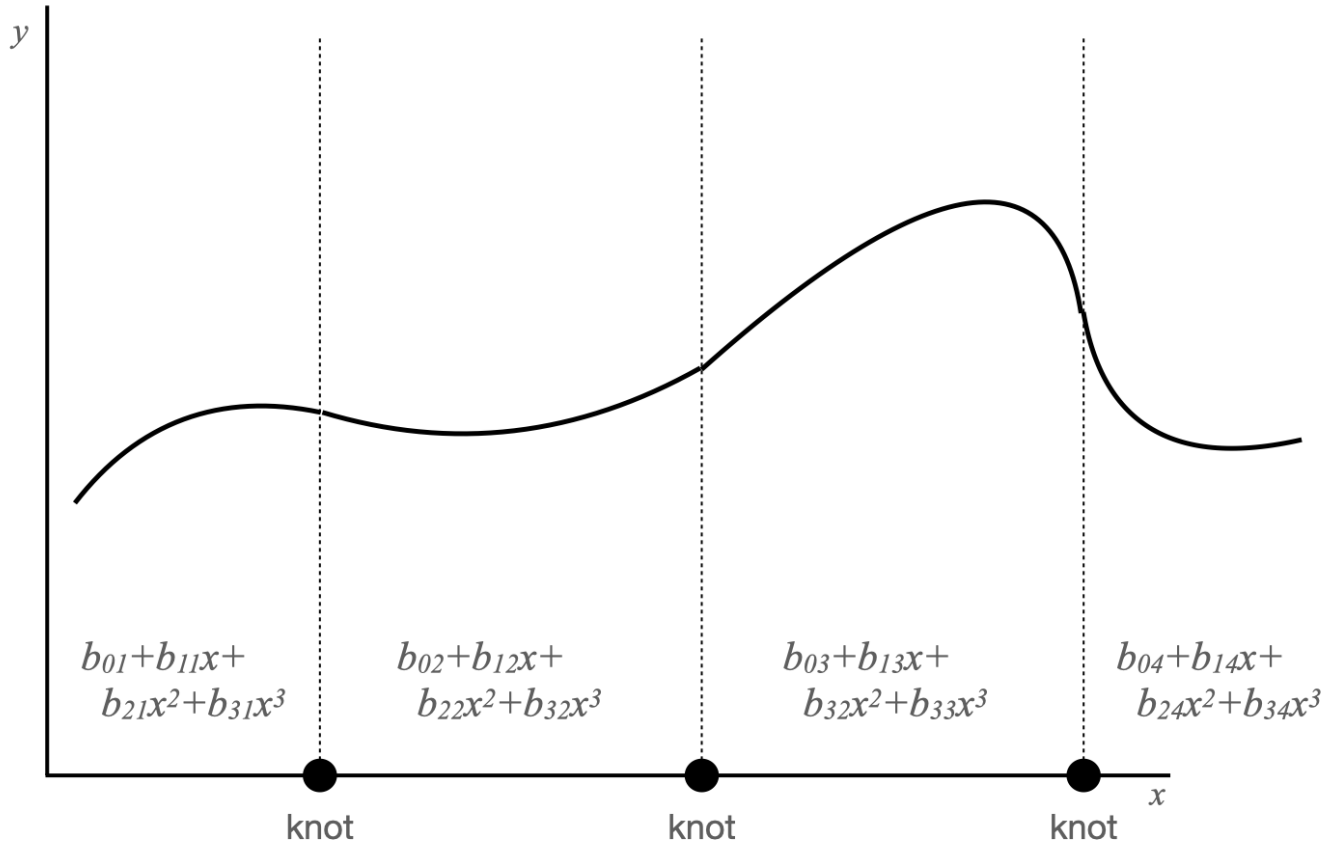
Although it is simple to add an extra x^2 or x^3 to the model, it induces a problem of collinearity among predictors. The solution is to use orthonormal polynomials.





Spline regression

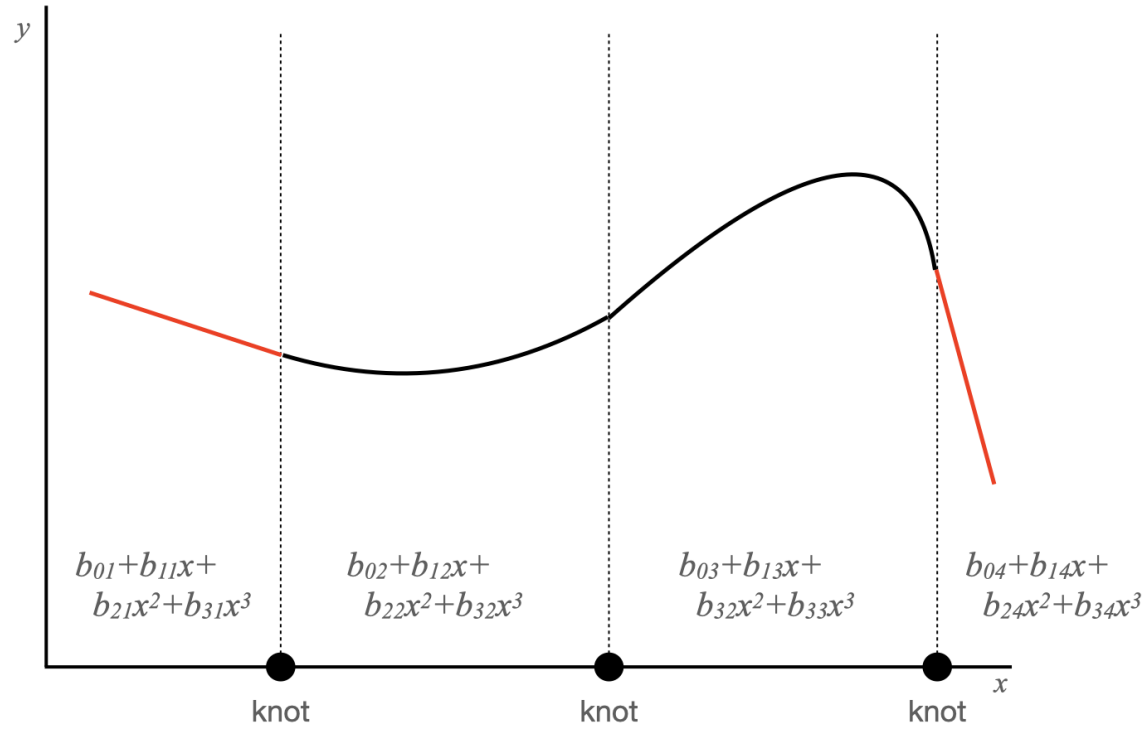
Fit a separate polynomial to different subsets.



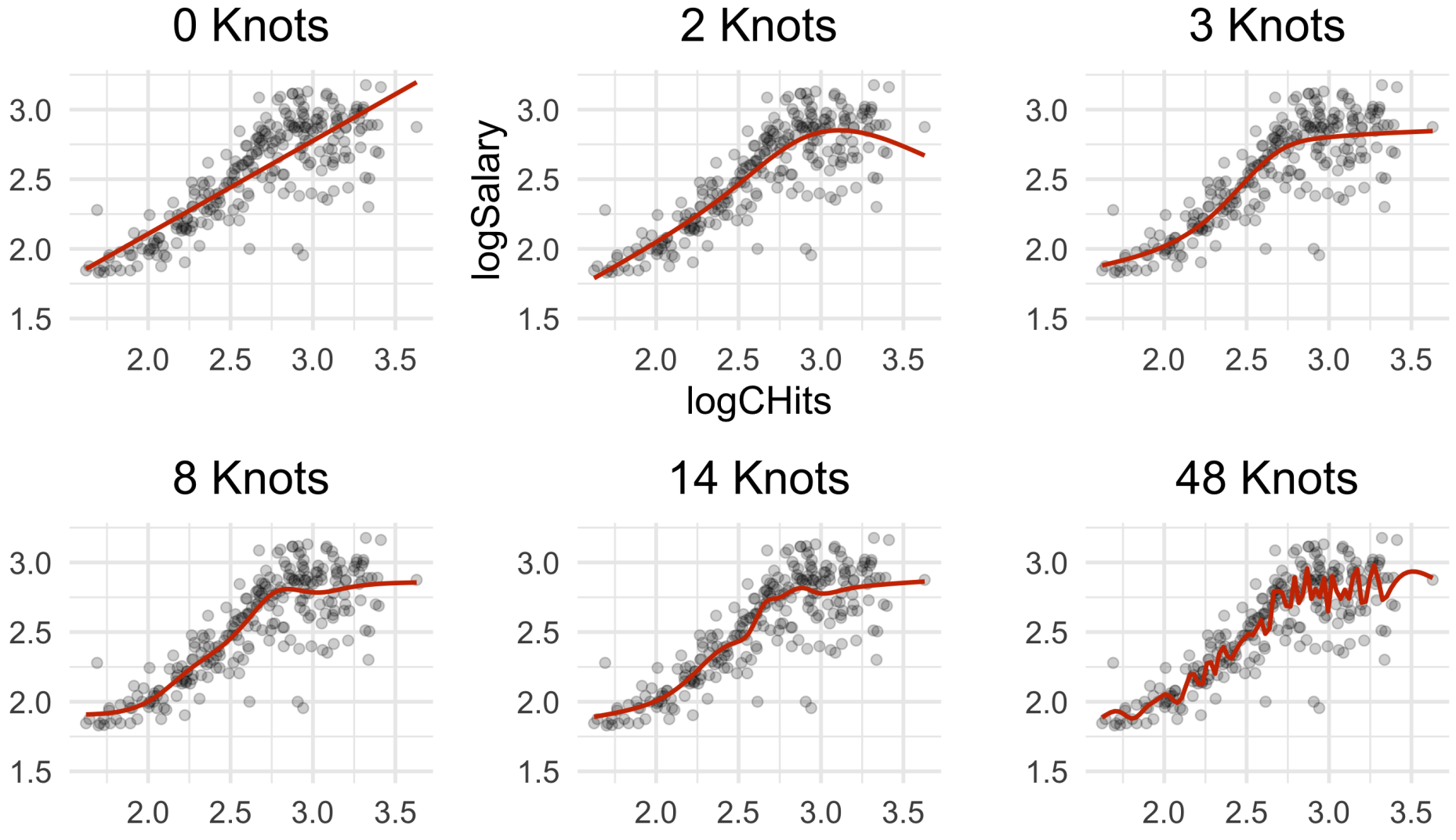
Data Science Deciphered: What is a Spline? has a lovely explanation.

Natural splines

Fit a separate polynomial to different subsets, and force a linear fit at the boundary.



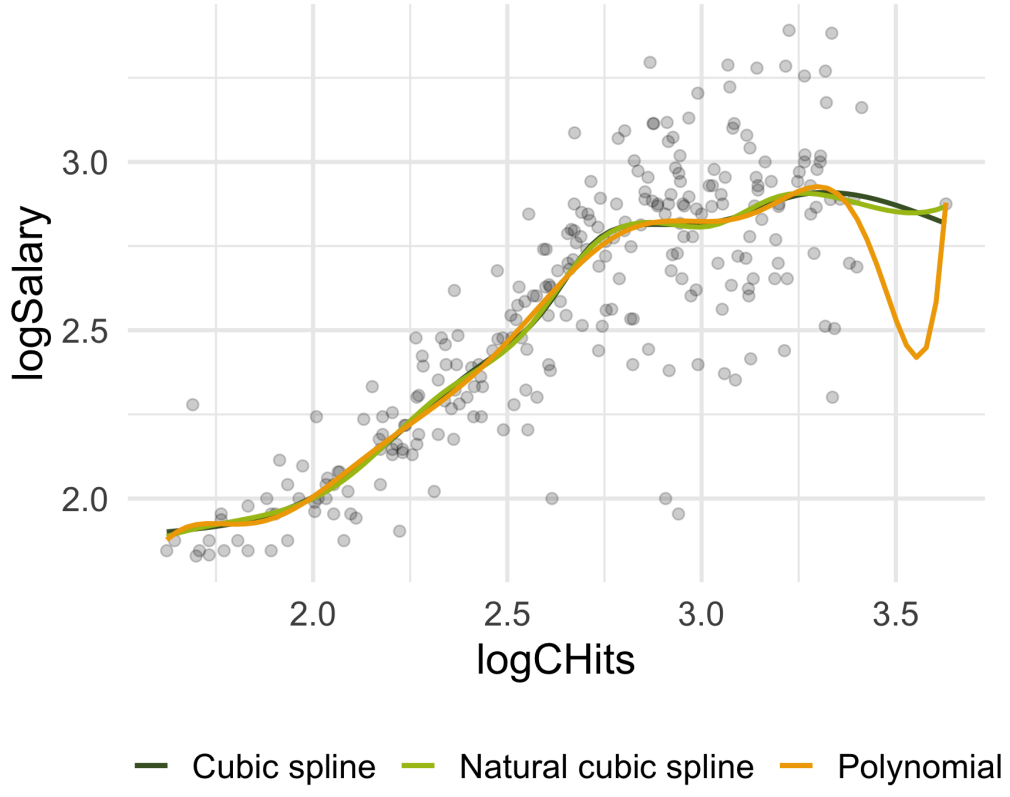
Natural cubic splines with differing number of knots



Comparison between splines and polynomials

We can fit a polynomial with `poly()`, cubic spline using `splines::bs()`, and fit a natural cubic spline using `splines::ns()`. Notice end of the curves, and the beginning.

- Polynomial is fitting $x, x^2, \ldots, x^{10}$.
- Spline is fitting degree 3 polynomial with added knots (breaks) for different functions in different subsets.
- Natural spline is fitting degree 3 polynomial, and knots with boundary forced to be linear.

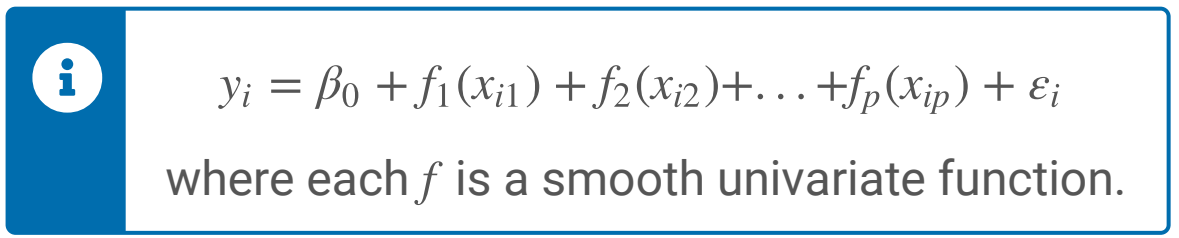


Generalised additive models (GAMs)

Its really hard to fit a model of the form

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

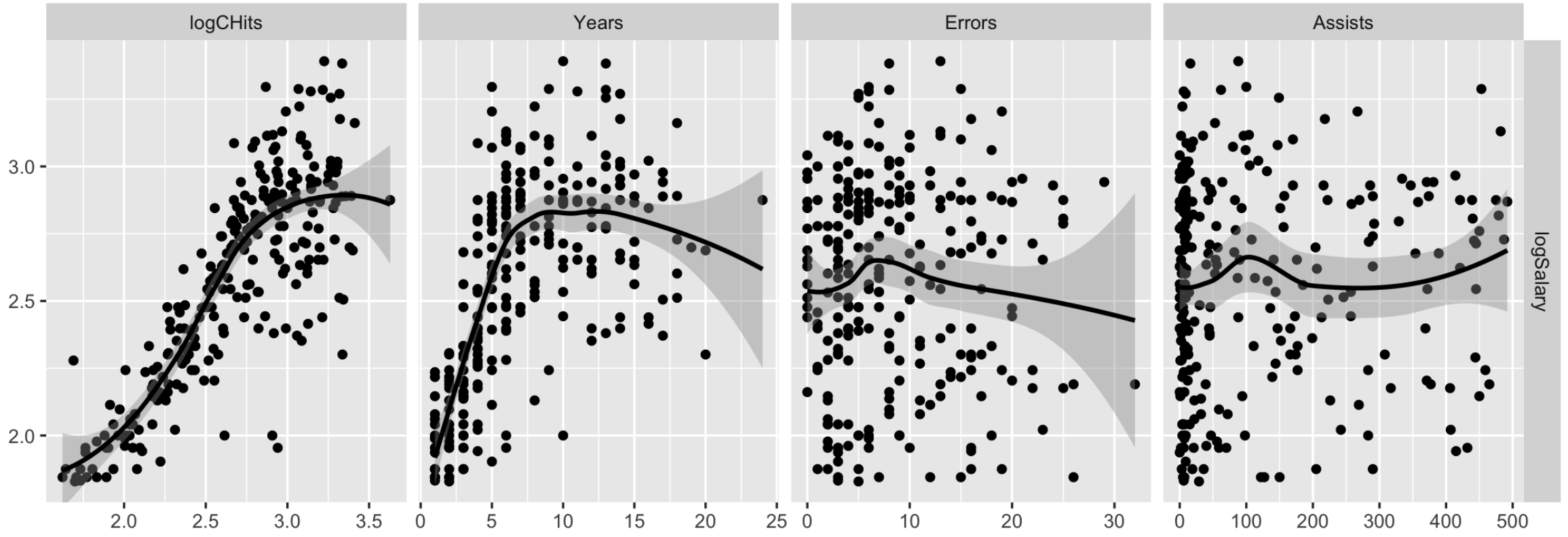
- Data is very sparse in high-dimensional space.
- Model assumes p-way interactions which are hard to estimate.
- Fit the model additively, is simpler, and still flexible, yet interpretable



$$y_i = \beta_0 + f_1(x_{i1}) + f_2(x_{i2}) + \ldots + f_p(x_{ip}) + \varepsilon_i$$

where each f is a smooth univariate function.

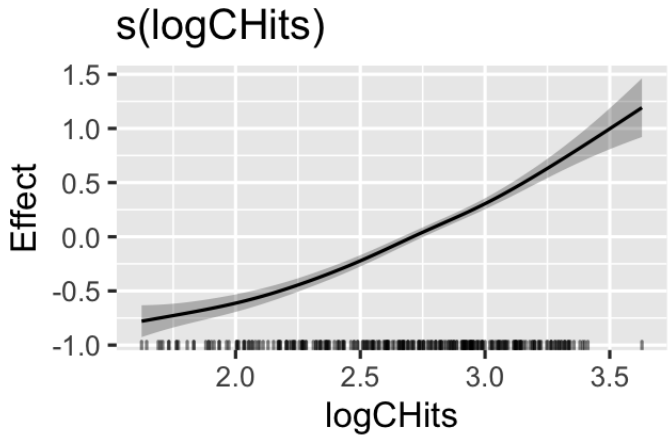
Example: here's the data

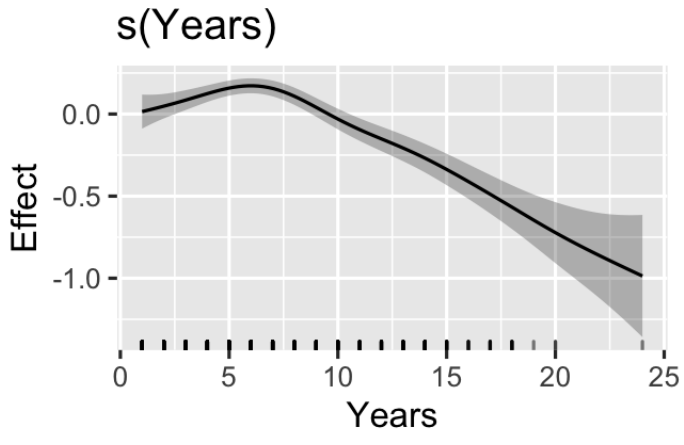


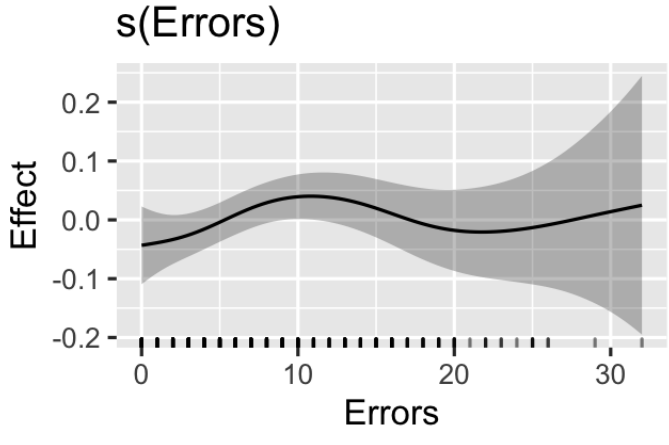
Example

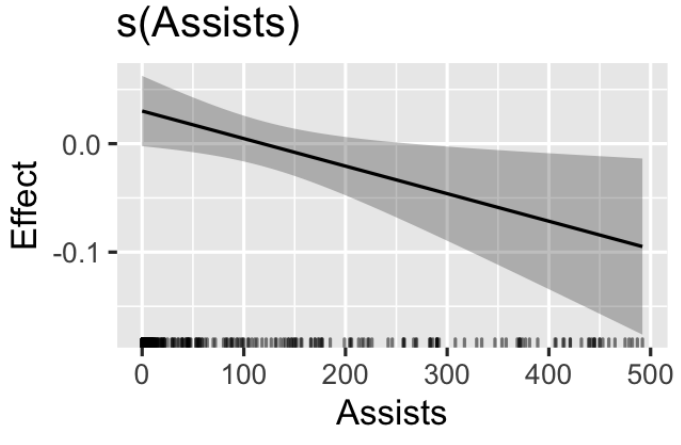
$$\log(\text{Salary}) = \beta_0 + f_1(\log(\text{CHits})) + f_2(\text{Years}) + f_3(\text{Errors}) + f_4(\text{Assists}) + \varepsilon$$

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



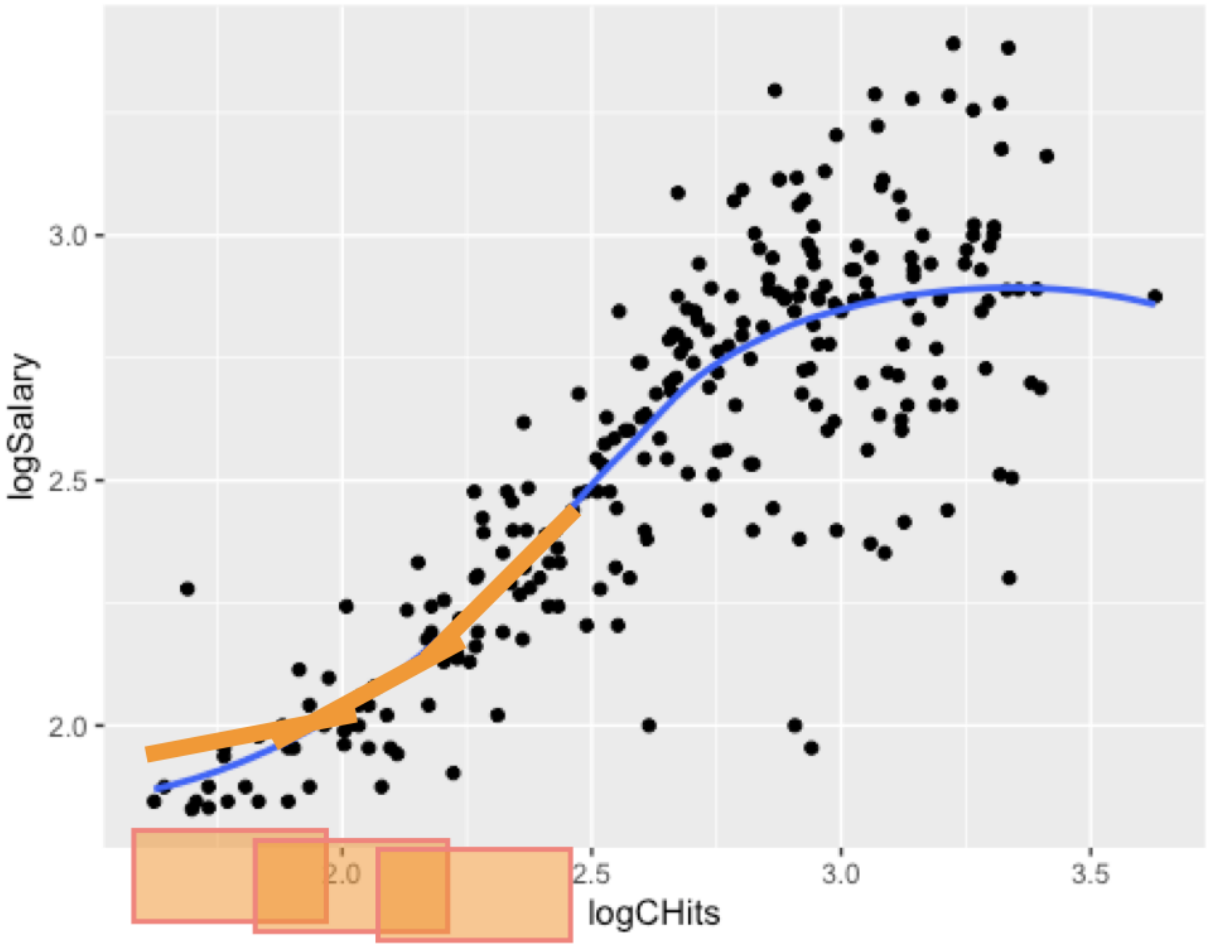






Generally

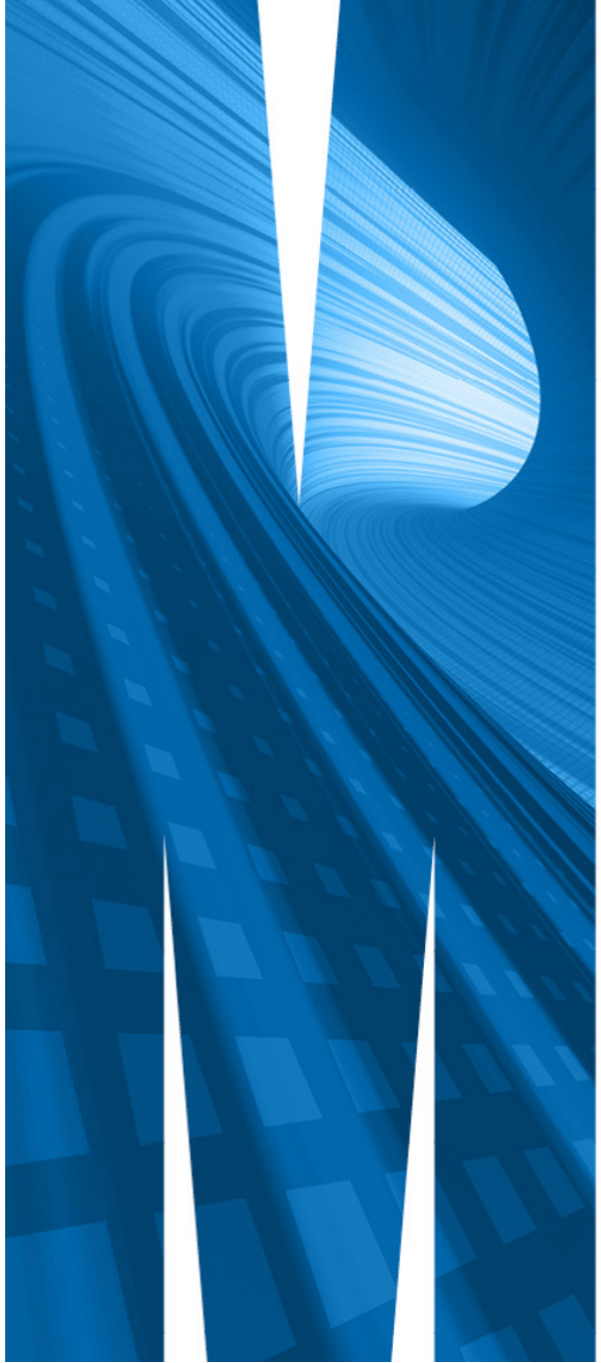
- You can fit a GAM manually using natural splines, for example.
- Coefficients are not interesting, the fitted functions are.
- Use `draw` from `gratia` package to plot the functions that are fitted as GAMs in `mgcv` package.
- The model can contain a mix of terms --- some linear, some nonlinear.
- 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)`.



Local regression (smoothers)

Overlapping subsets of data, (weighted) regression on each subset. Overlap helps to smooth the fitted model.

MONASH University



Lecturer: *Professor Di Cook*

Department of Econometrics and Business Statistics

ETC3250.Clayton-x@monash.edu

Week 2b