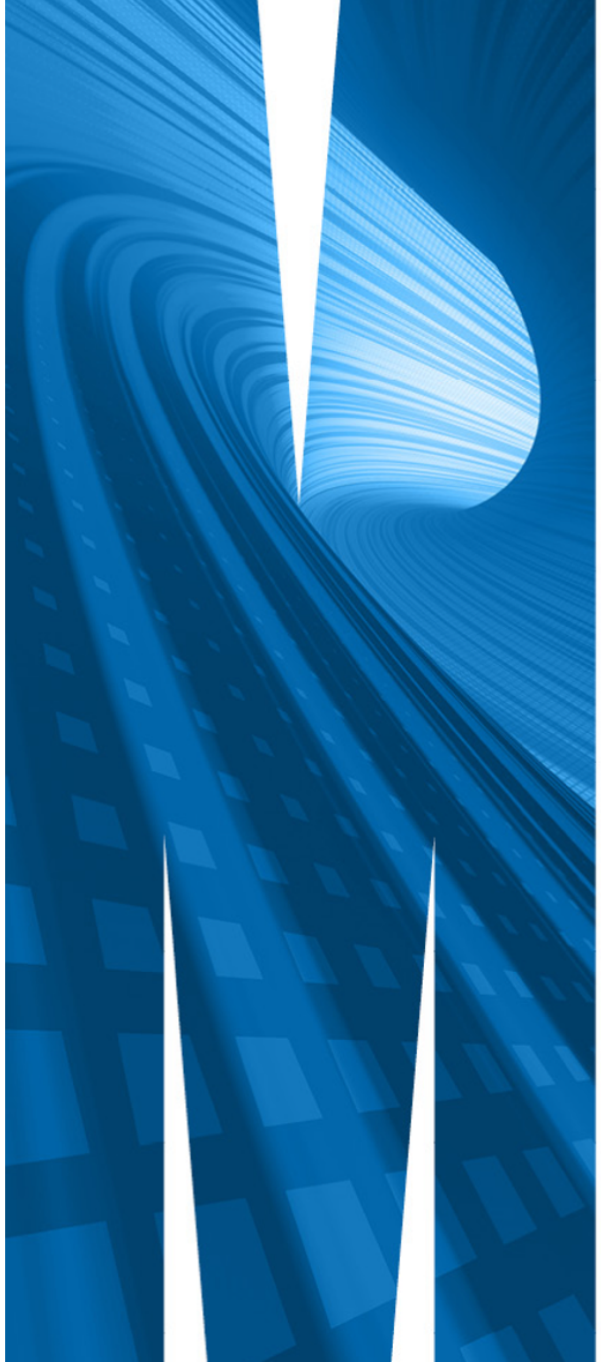
MONASH University

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...

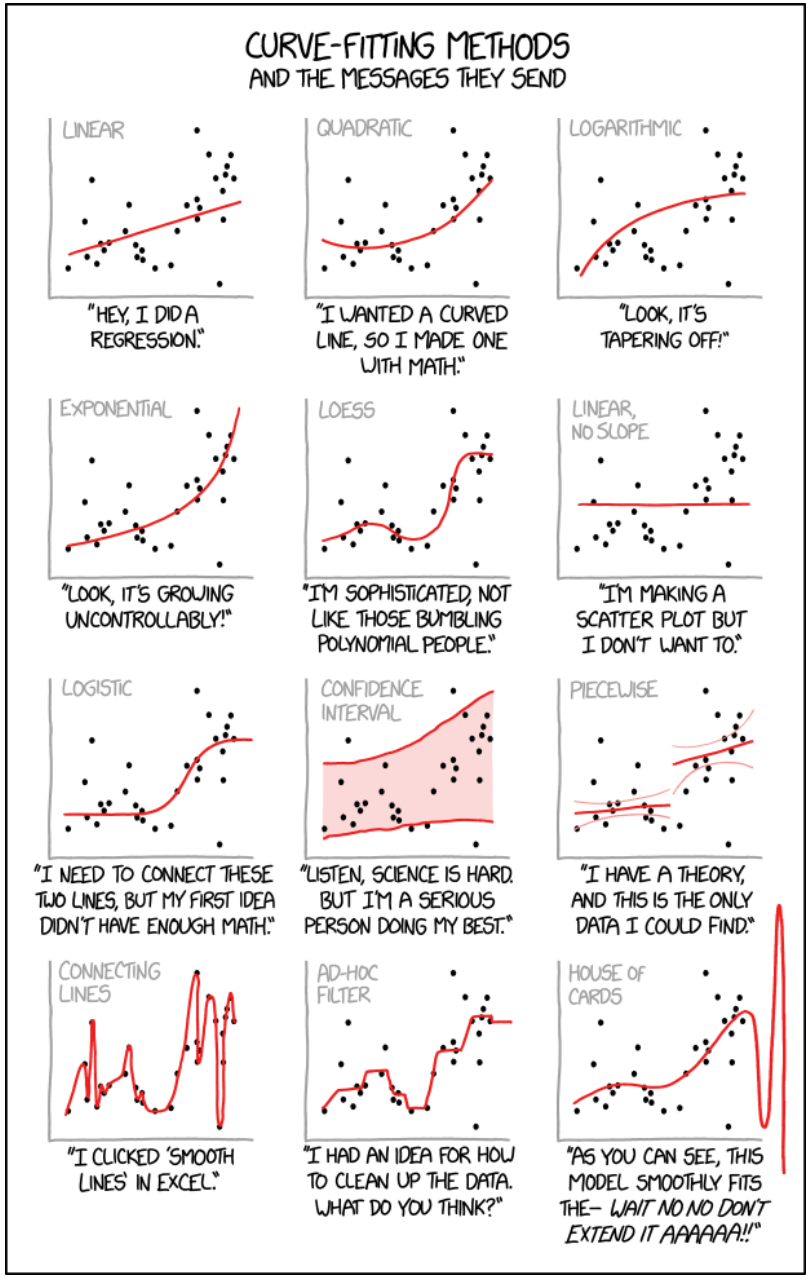
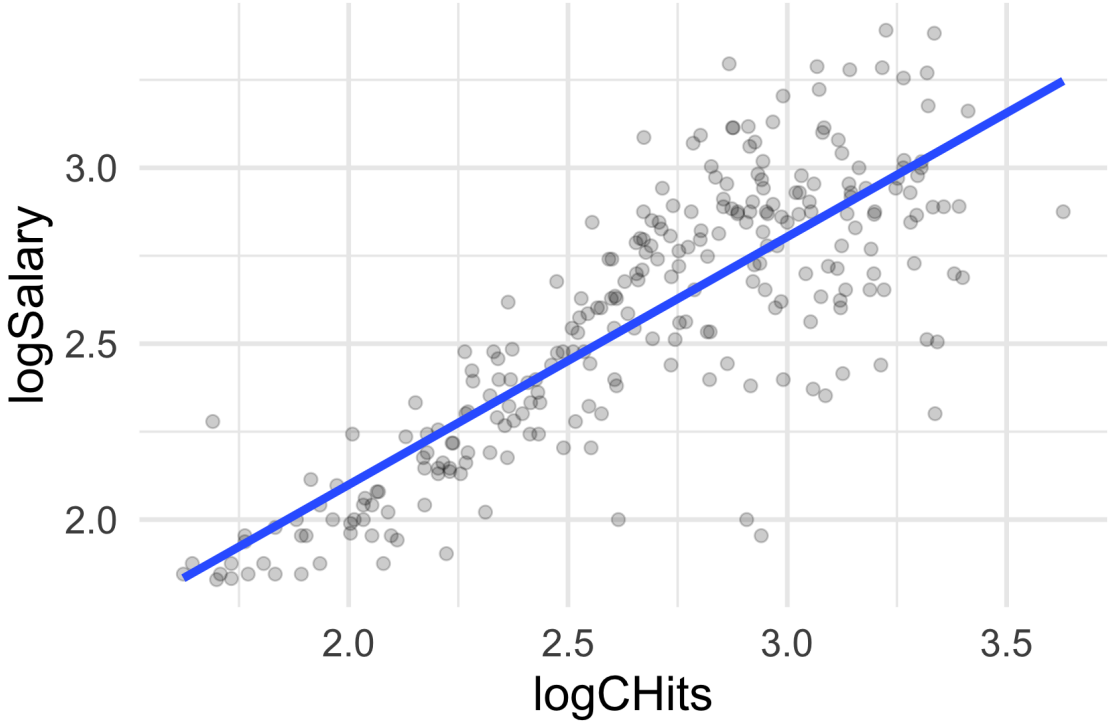


Image source: XKCD

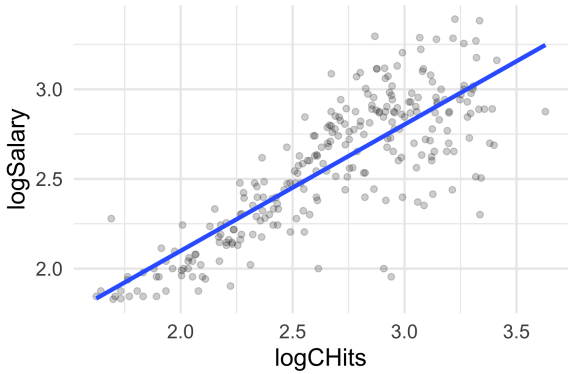
Moving beyond linearity

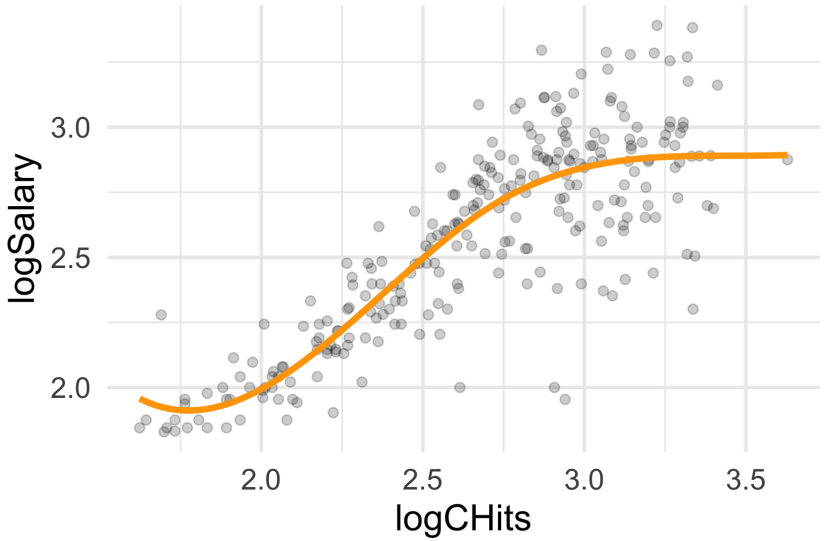
- Consider the following Major League Baseball data from the 1986 and 1987 seasons.
- Would a linear model be appropriate for modelling the relationship between Salary and Career hits, captured in the variables `logSalary` and `logCHits`?

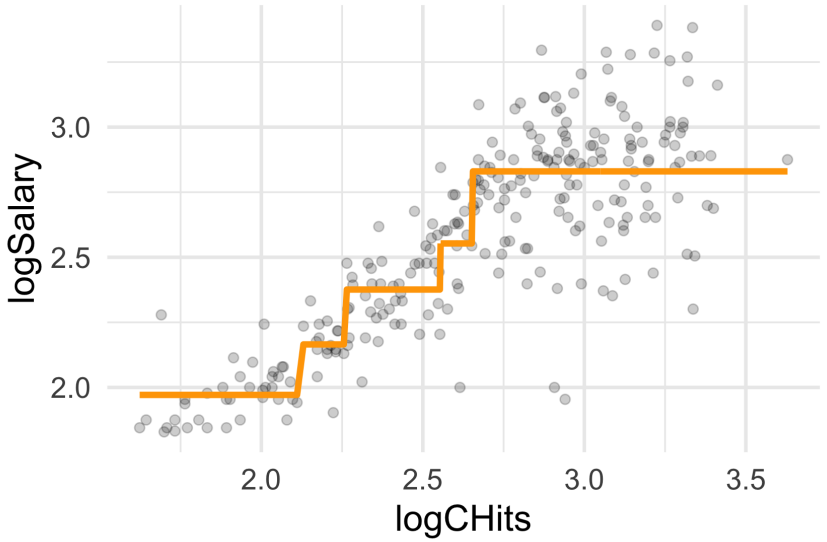


Moving beyond linearity

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





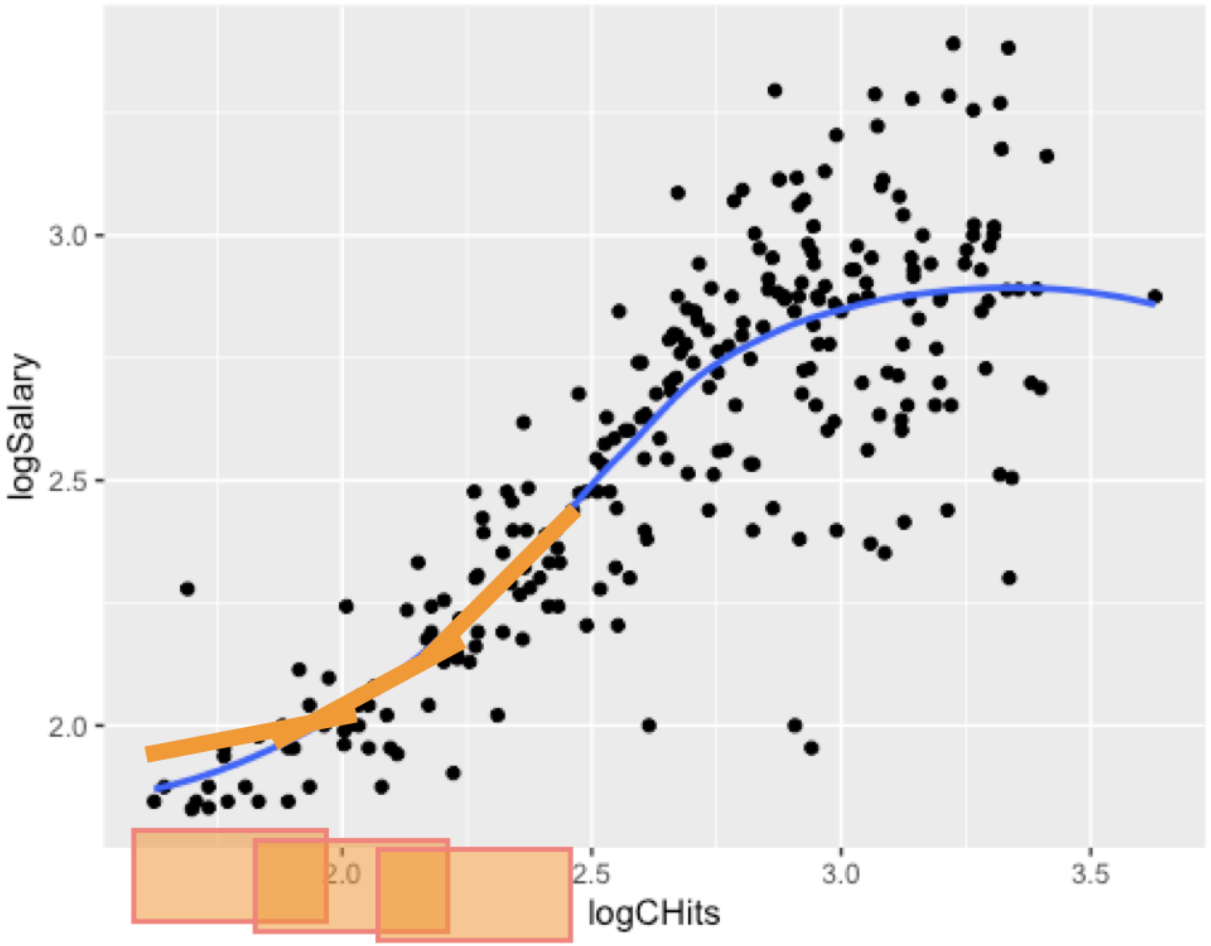


Flexible regression fits

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

- local regression, sliding window with regression fitted to subsets;
- 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;
- generalized additive models, extend these approaches to multiple predictors.

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



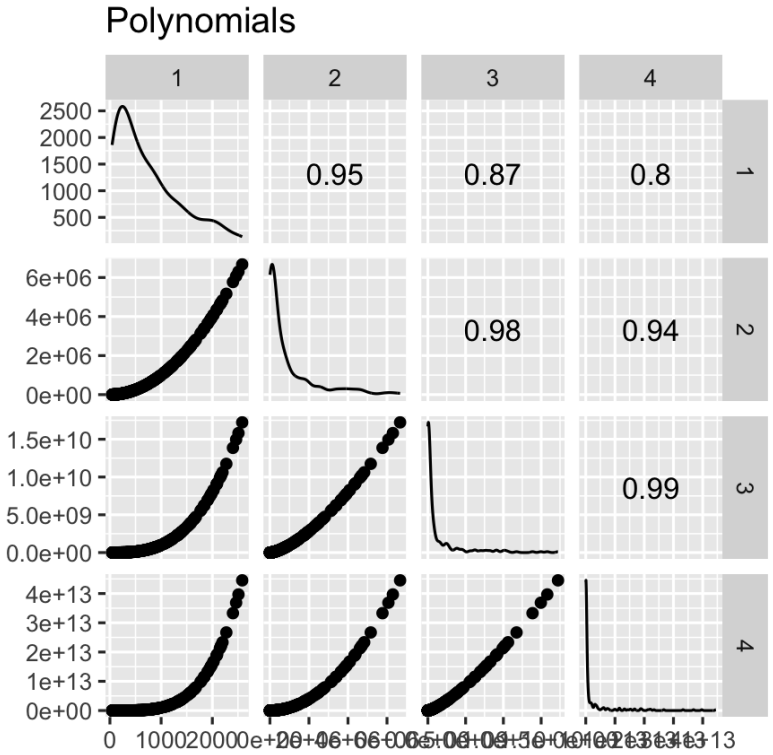
Local regression (smoothers)

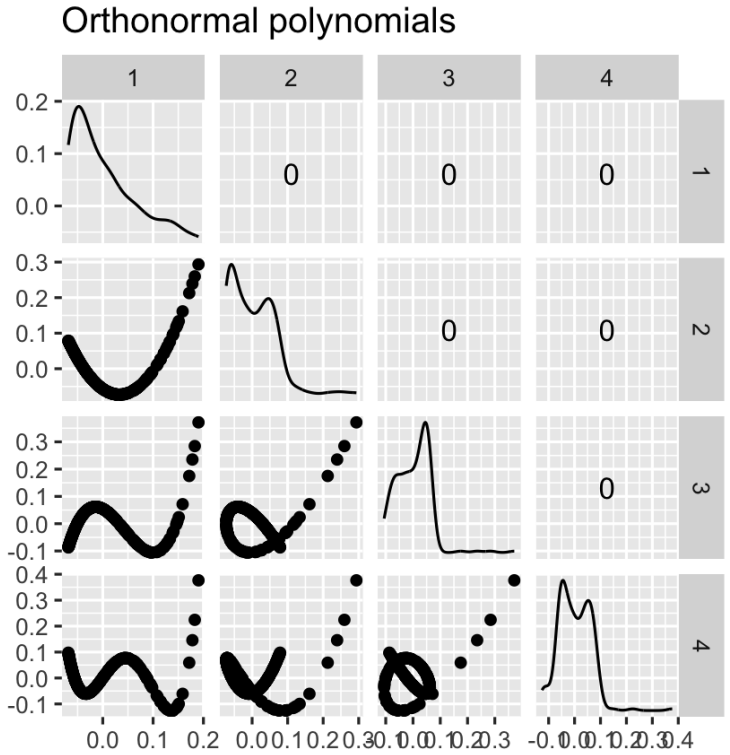
Overlapping subsets of data, (weighted) regression on each subset. Overlap helps to smooth the fitted model.

A drawback of this approach is that it does not produce a functional form of the fitted model.

Polynomial regression

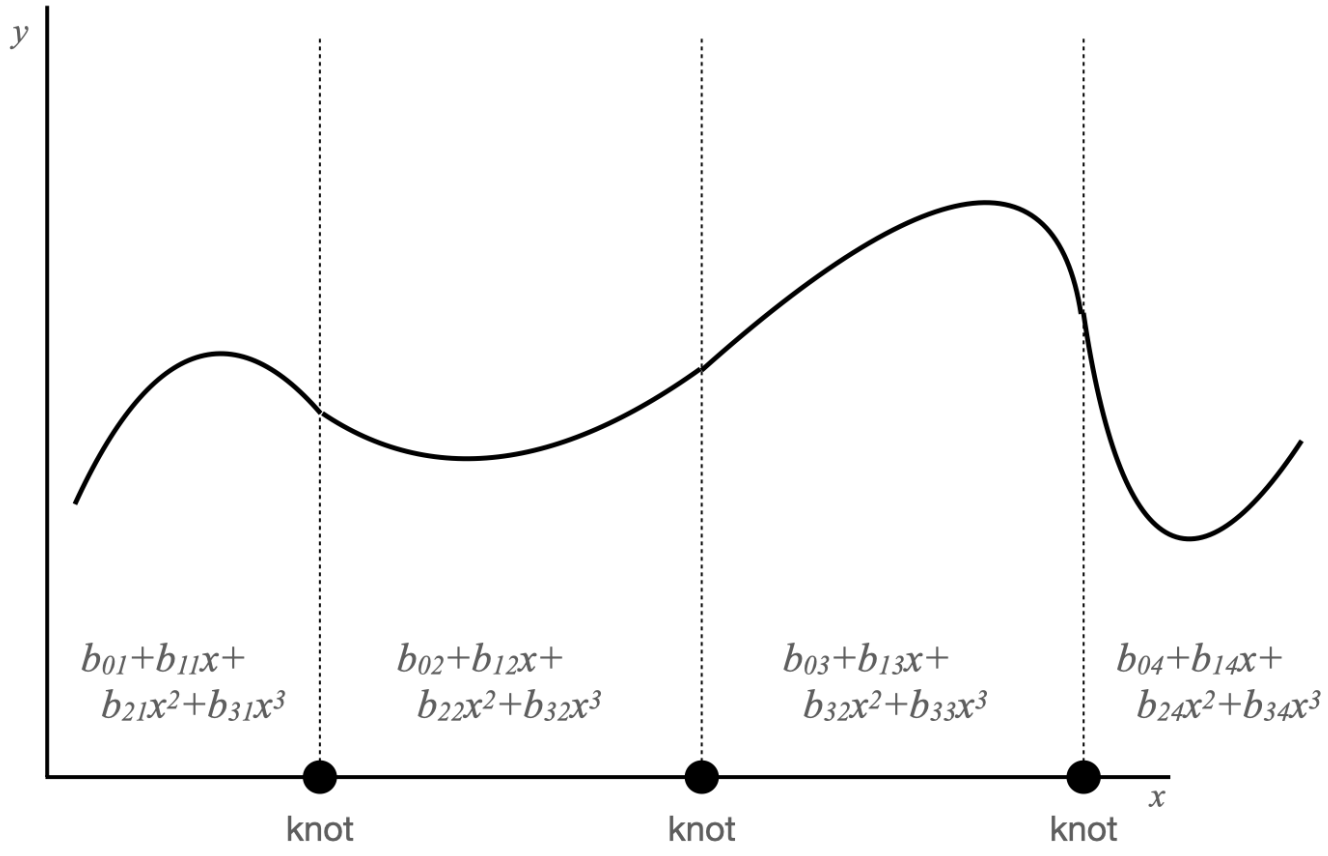
Although it is simple to add an extra x^2 or x^3 to the model, it induces a problem of collinearity among predictors. The solution is to use orthogonal polynomials.





Spline regression

Fit a separate polynomial to different subsets.

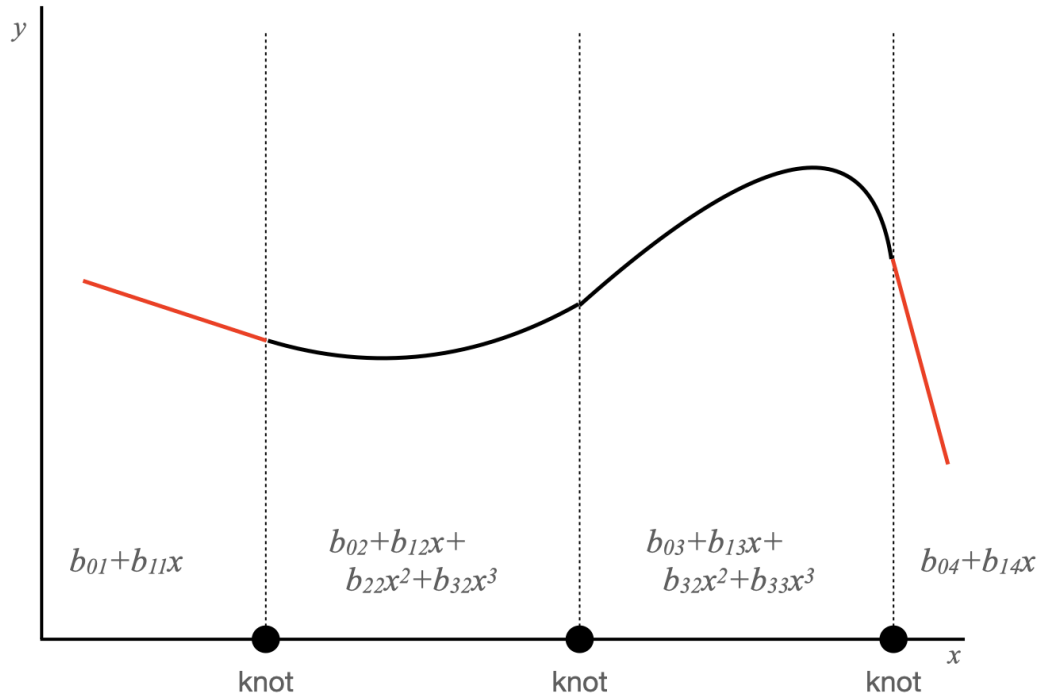


Data Science Deciphered: What is a Spline? has a lovely explanation.

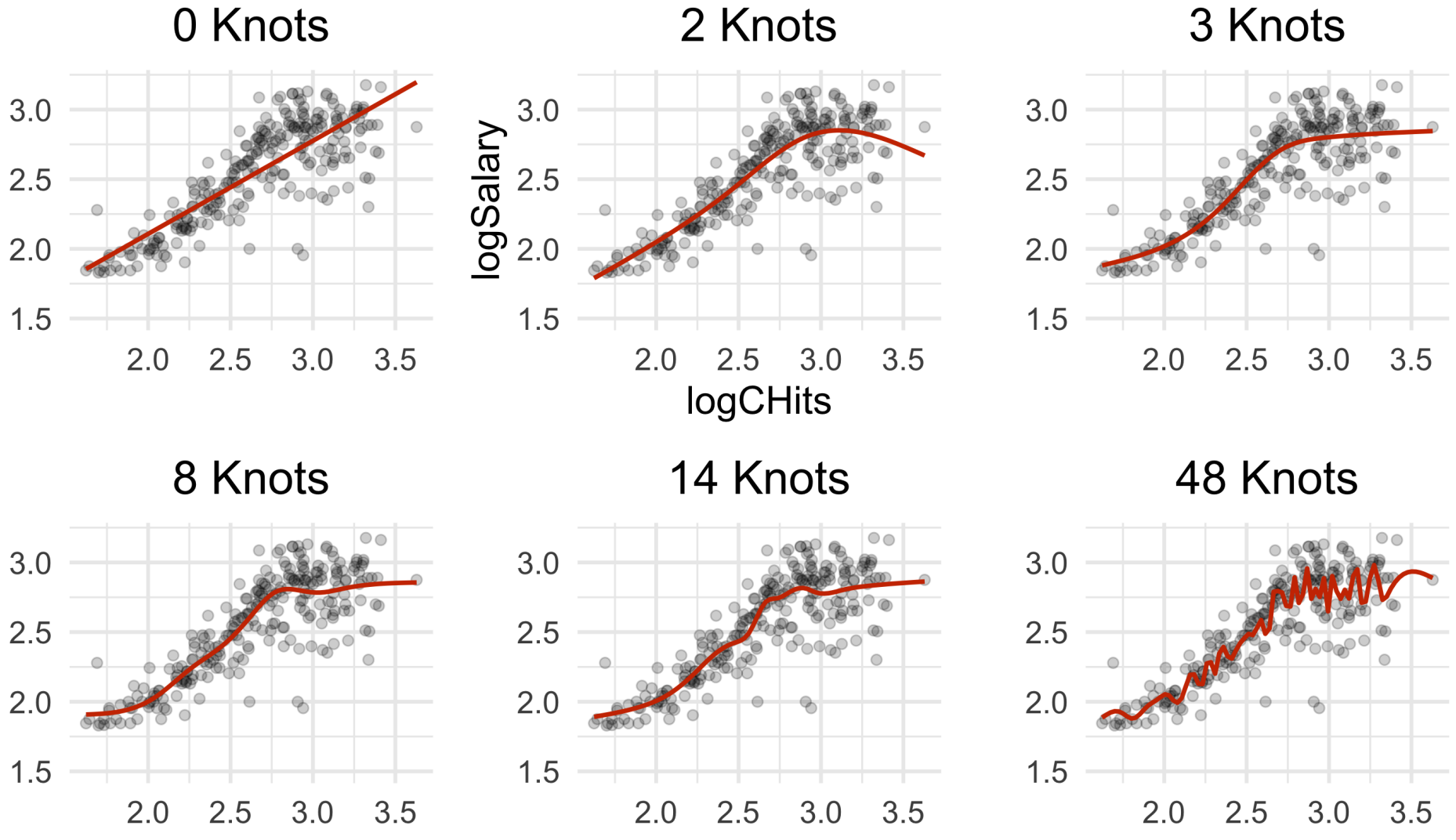
Natural splines

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

Something like this illustration.



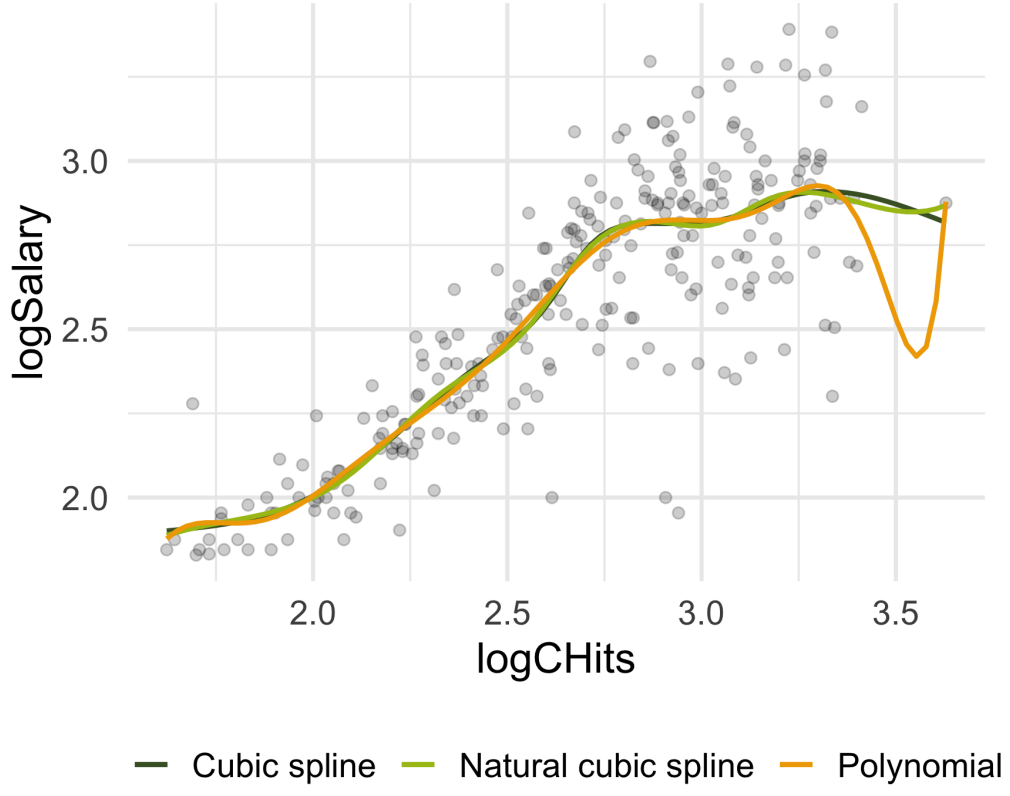
Natural cubic splines with differing number of knots



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.

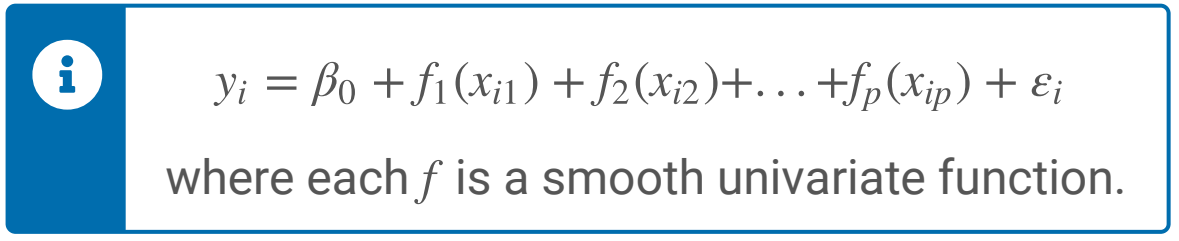


Generalised additive models (GAMs)

It's 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



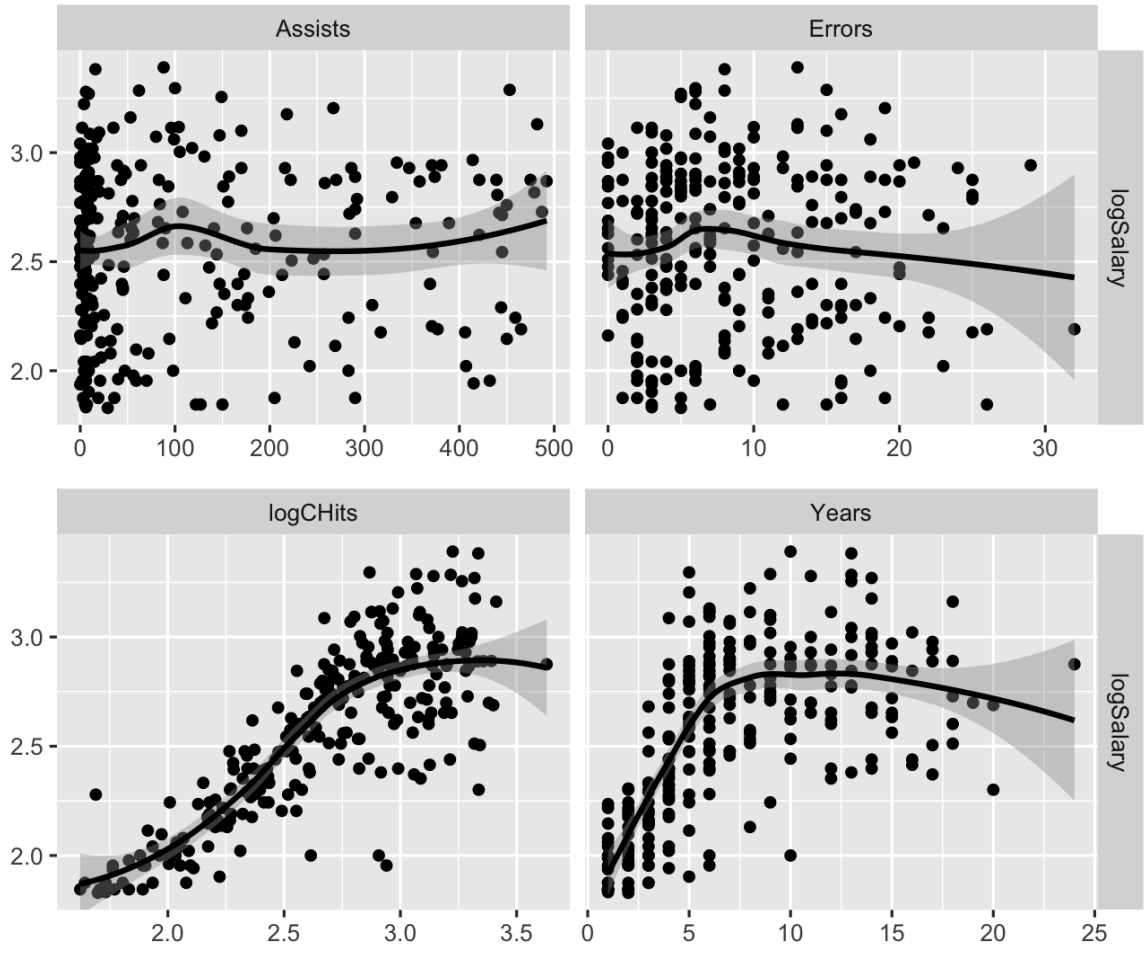
$$y_i = \beta_0 + f_1(x_{i1}) + f_2(x_{i2}) + \ldots + f_p(x_{ip}) + \varepsilon_i$$

where each f is a smooth univariate function.

Example: Baseball

Scatterplots of `logSalary` vs predictors, with a loess smoother overlaid.

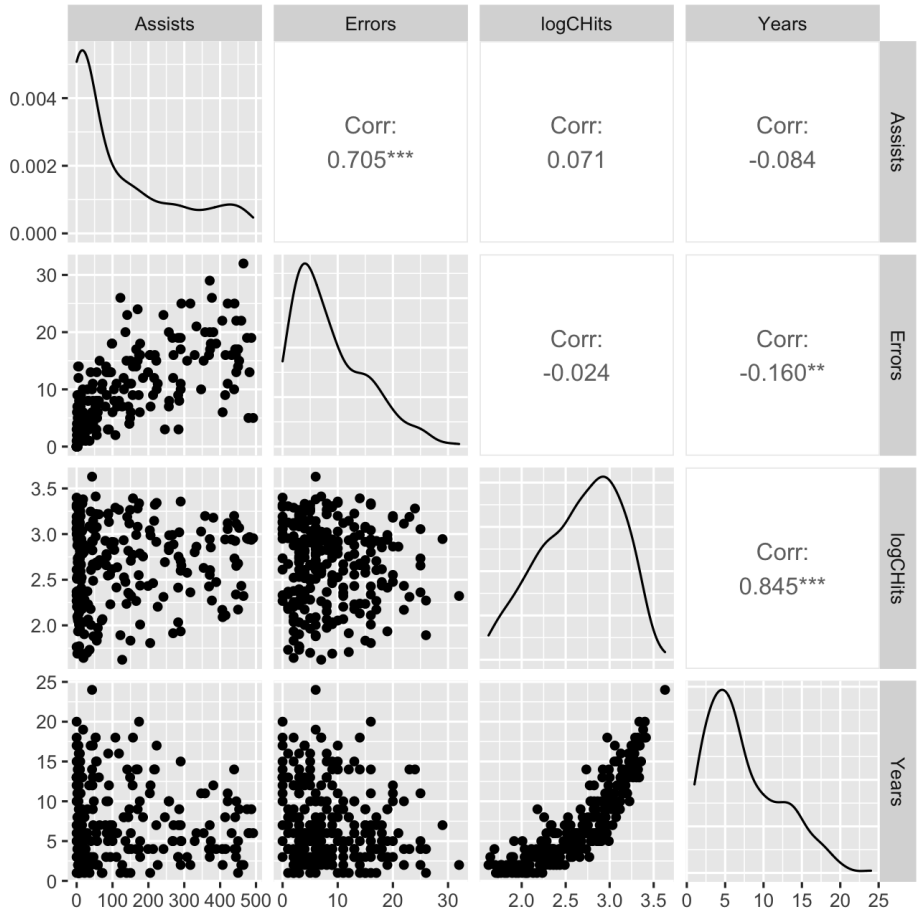
Strong nonlinear relationships with `logCHits` and moderate relationship with `Years`. No relationship with `Assists` and `Errors`.



Example: Baseball

Examine the predictors. There should be no strong associations, or outliers or clusters.

Unfortunately, `logCHits` and `Years` are collinear.

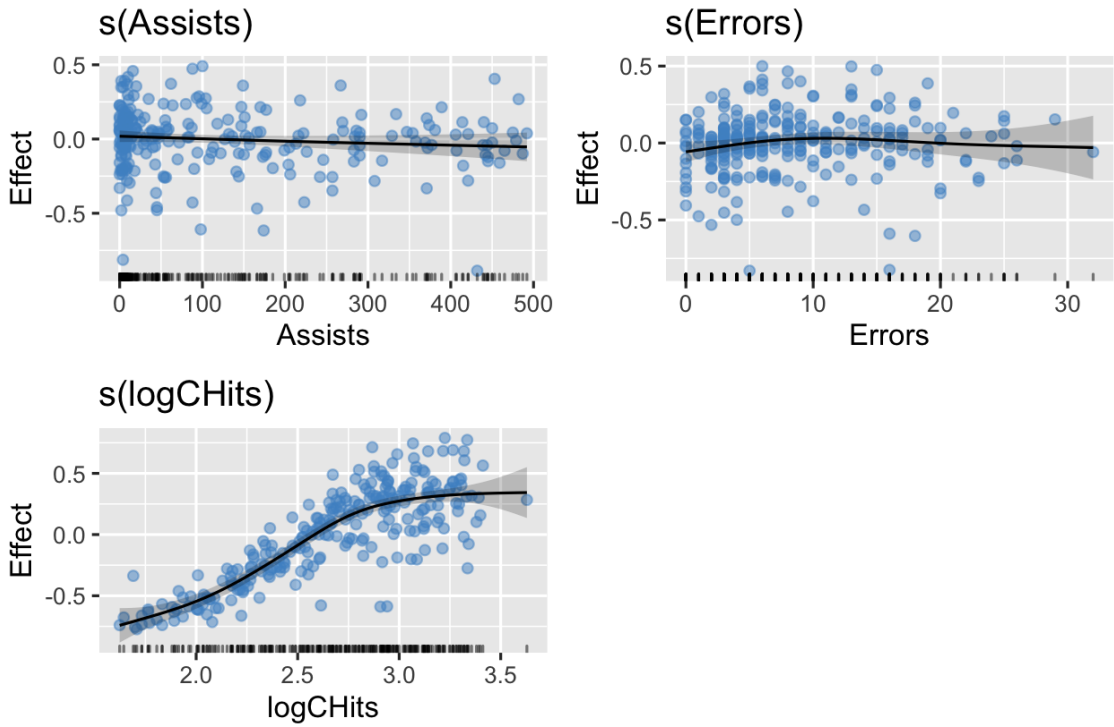


Example: Baseball

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

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

```
gratia::draw(hits_gam, residuals=TRUE)
```



Estimated smooths from fitted model. (See Gavin Simpson's explanations.)

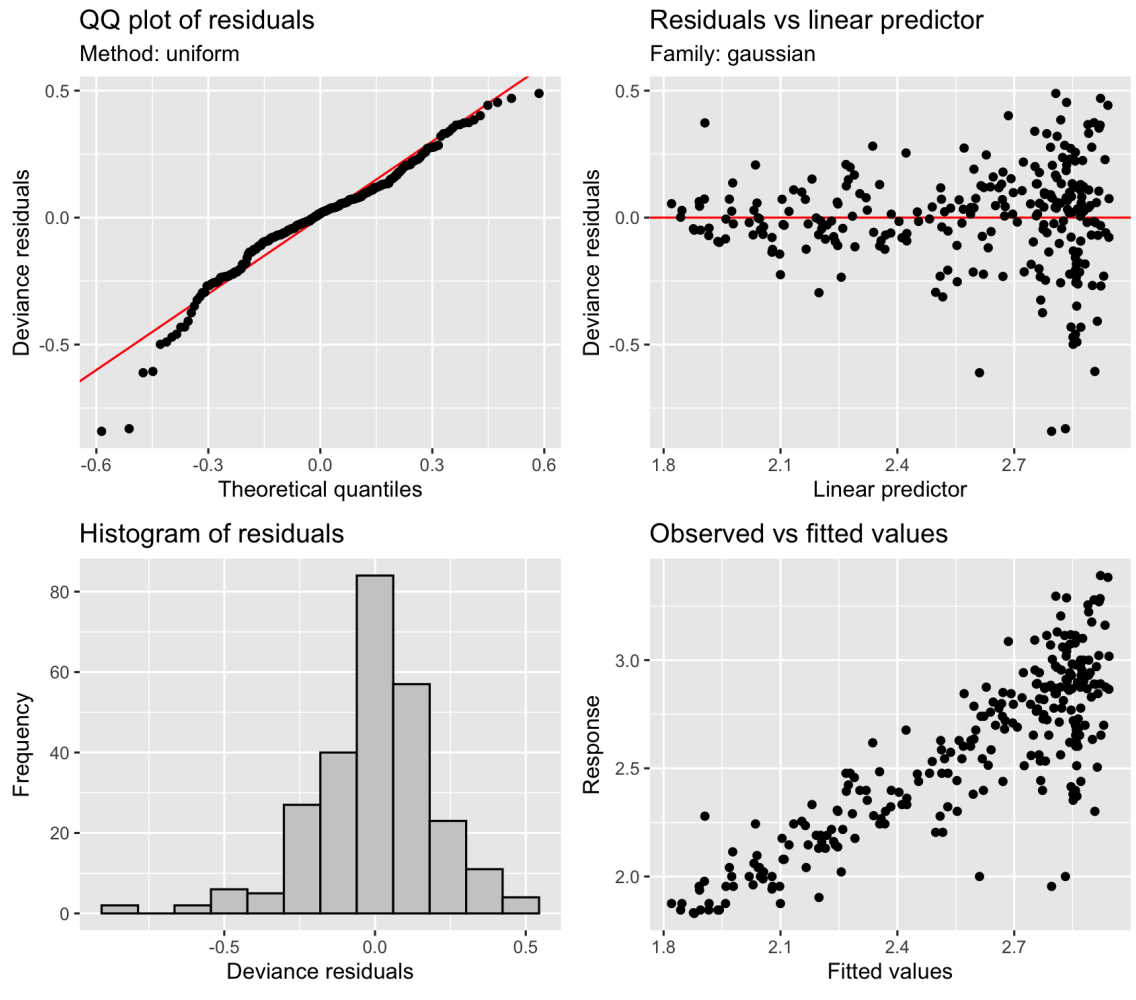
Summarising the model fit

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## logSalary ~ s(logCHits) + s(Errors) + s(Assists)
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.56978    0.01254   204.9   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                edf Ref.df       F p-value
## s(logCHits) 4.009  4.997 132.350  <2e-16 ***
## s(Errors)   2.360  2.983   1.813   0.171
```

Summarising the model fit

```
gratia::appraise(hits_gam)
```

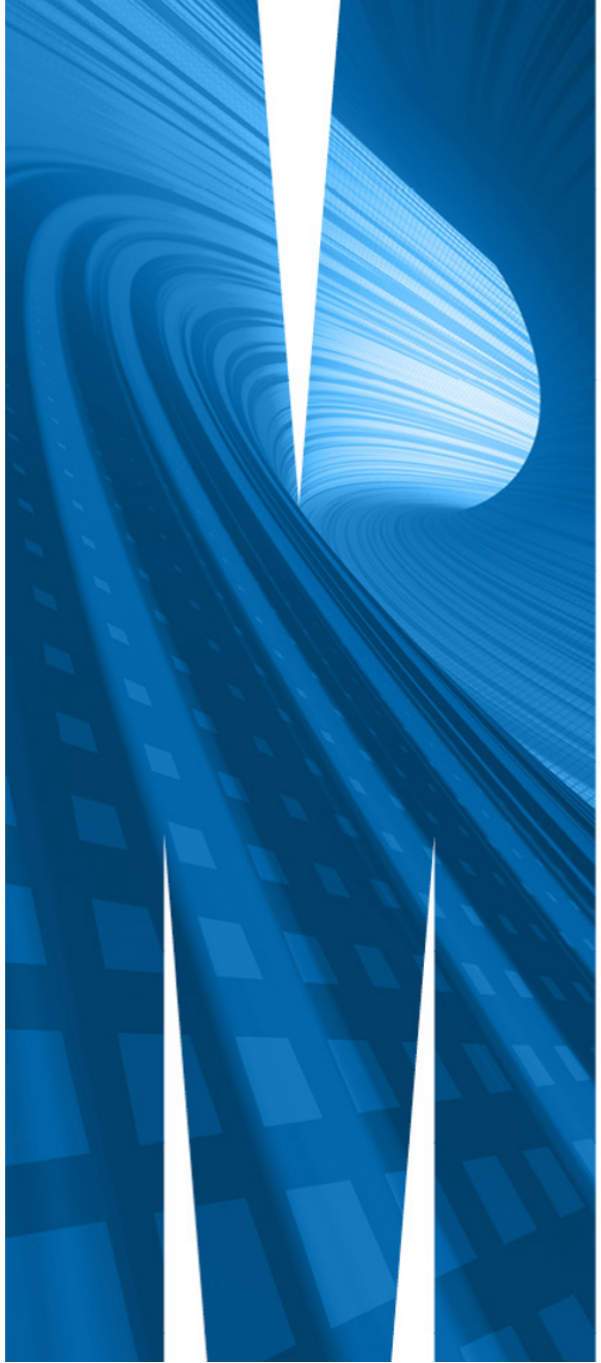
- Plot observed vs fitted: should be a strong association. (Mostly good, a few outliers.)
- Histogram of residuals: should be bell-shaped. (Slightly left-skewed, with some unusually small values.)
- Normal probability plot of residuals: if residuals are a sample from normal then these values form a straight line. (Good except for some low and high observations.)
- Residuals vs fitted: roughly even vertical spread for all x values. (Not good, spread is heteroskedastic.)



Summary

- A GAM is a fit to functions of each predictor, and can be manually fitted using natural splines, or other functions.
- Coefficients are generally not interesting, the fitted functions are.
- 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)`.

MONASH University



Lecturer: *Professor Di Cook*

Department of Econometrics and Business Statistics

ETC3250.Clayton-x@monash.edu

Week 2b