Dealing with non-linearity

(Fractional) Polynomial Regression and Regression Splines

Matthias Mittner

Institute for Psychology, University of Tromsø, Norway

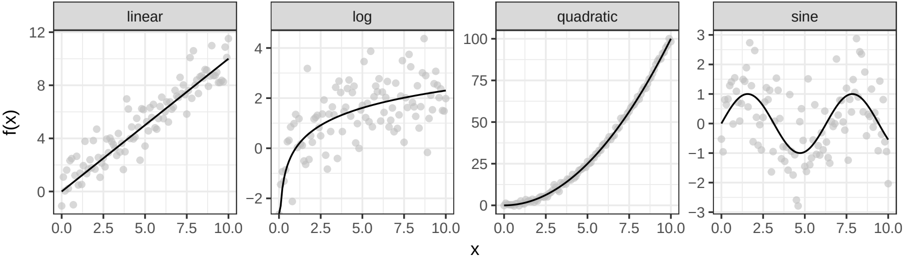
12.05.2017

UNIVERSITETET I TROMSØ UiT

Overview

1. Nonlinear relationships
2. Polynomial Regression
3. Bias-Variance tradeoff
    - predictive accuracy
    - model-selection
4. Fractional Polynomial Regression
5. Regression Splines

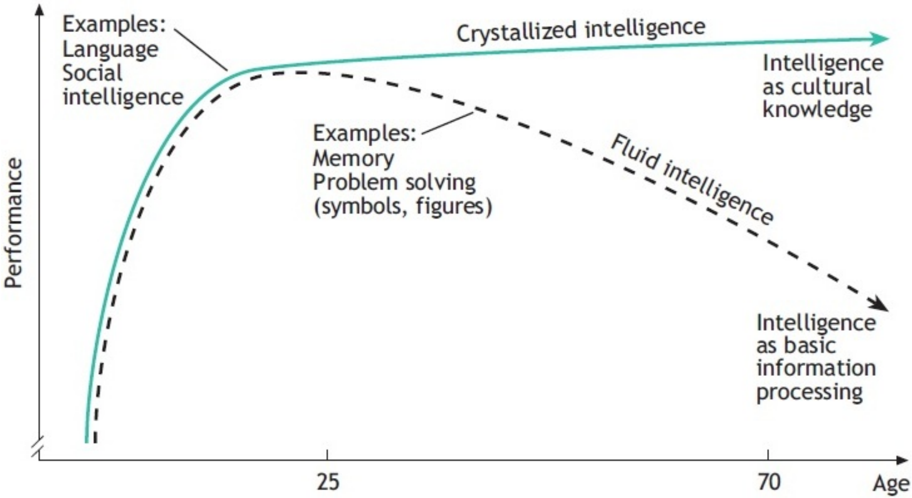
Nonlinear Relationships



- so far: *linear* regression only
- what if relationship between variables is not linear?
- **can you think of examples of non-linear relationships?**

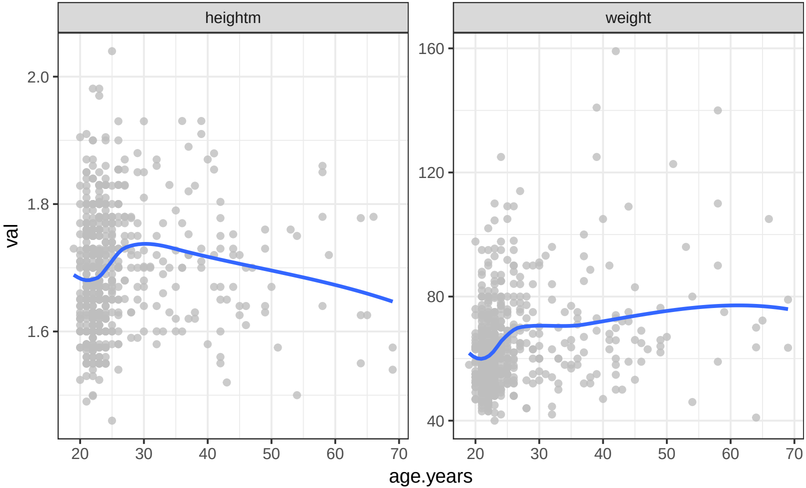
Examples of non-linearities due to growth

Age and IQ



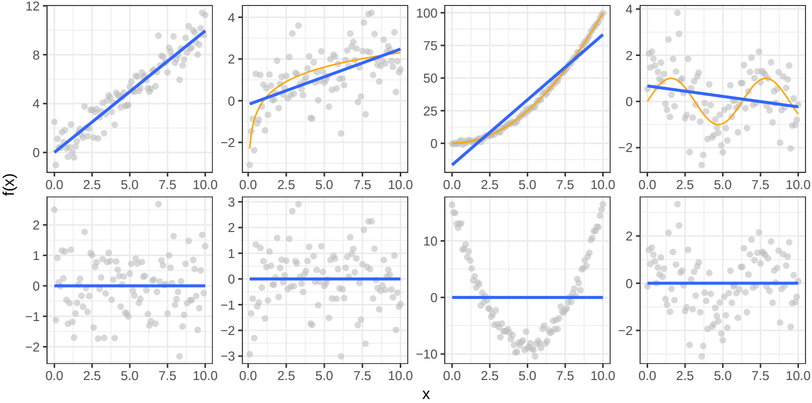
Examples of non-linearities due to growth

Development of body height and weight with age



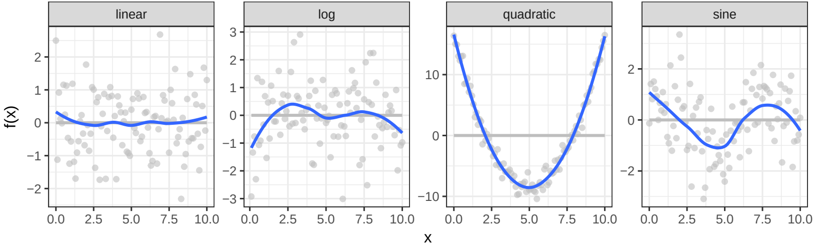
Data from `https://osf.io/2rm5b/`

How do we detect non-linearities?



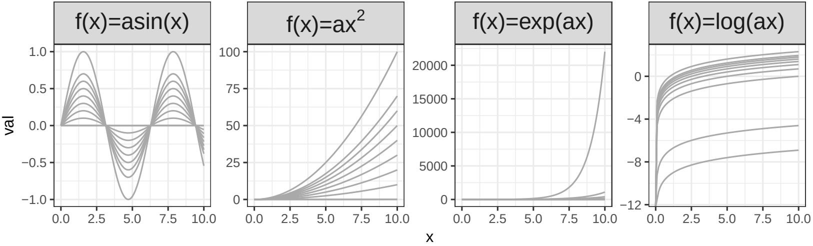
- look a the regression residuals (lower plot)
- is there any structure in the residuals?

How do we detect non-linearities?



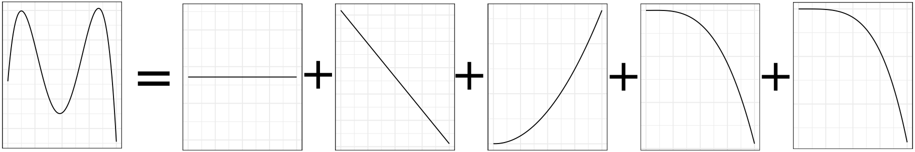
- adding a smoother to the plot can help to detect non-linearities
- when non-linearity is suspected, fit a non-linear model and compare it to the linear one (model-selection)

Nonlinear Regression



- in principle, we can assume any (parametrized) curve-shape and fit it to data
- in these example, we could "tweak" the parameter a to best account for the data
- this is called "Nonlinear regression"
- linear regression: $y = b_0 + b_1 x + \epsilon$
- non-linear regression: $y = f(x; \theta) + \epsilon$

Linearization



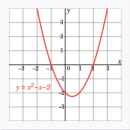
- in practice: general nonlinear regression can be hard (fitting the function can be difficult)
- smart to stick to functions that can be linearized $\rightarrow$ least-squares fitting from linear regression can be used!
- polynomials are useful because they can be decomposed linearly

Polynomials

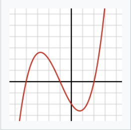
Definition

$$f(x) = a_0 + a_1 x + a_2 x^2 + \ldots + a_m x^m$$

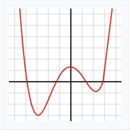
- the highest power m in the polynomial is called the "degree" or "order" of the polynomial
- some coefficients can be zero $a_i = 0$, then the term is left out of the equation
- the constant function $f(x) = a_0$ is a polynomial (degree 0)
- the linear function $f(x) = a_0 + a_1 x$ is a polynomial (degree 1)



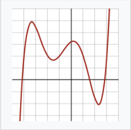
Polynomial of degree 2:
$f(x) = x^2 - x - 2$
$= (x + 1)(x - 2)$



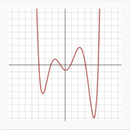
Polynomial of degree 3:
$f(x) = x^3/4 + 3x^2/4 - 3x/2 - 2$
$= 1/4\ (x + 4)(x + 1)(x - 2)$



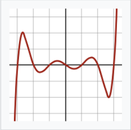
Polynomial of degree 4:
$f(x) = 1/14\ (x + 4)(x + 1)(x - 1)(x - 3)$
$+ 0.5$



Polynomial of degree 5:
$f(x) = 1/20\ (x + 4)(x + 2)(x + 1)(x - 1)$
$(x - 3) + 2$

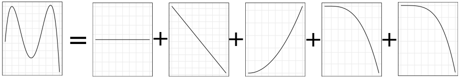


Polynomial of degree 6:
$f(x) = 1/100\ (x^6 - 2x^5 - 26x^4 + 28x^3$
$+ 145x^2 - 26x - 80)$



Polynomial of degree 7:
$f(x) = (x - 3)(x - 2)(x - 1)(x)(x + 1)(x + 2)$
$(x + 3)$

Linearization and polynomial regression



Linearization:

$$y = f(x;\theta) + \epsilon = f_1(x;\theta_1) + f_2(x;\theta_2) + \ldots + f_m(x;\theta_m) + \epsilon$$

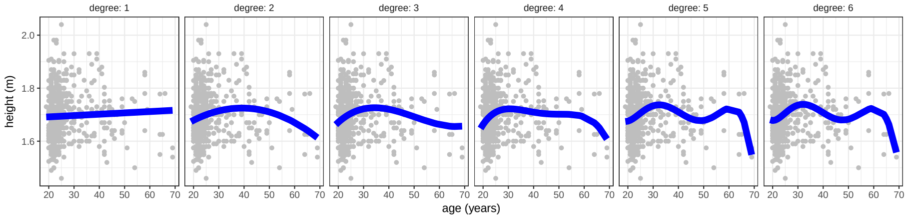
Polynomial regression

$$y = f(x; b_0, \ldots, b_m) + \epsilon = b_0 + b_1 x + b_2 x^2 + \ldots + b_m x^m + \epsilon$$

- polynomials can be linearized
- one predictor x is "spread out" over many variables $(x, x^2, x^3, \ldots)$
- this extended, multiple regression model can be fit as usual

Polynomial Regression

What is an appropriate degree for the polynomial?



- a polynomial of degree m can only have $m - 1$ turning points
- it is not always obvious from the data what an appropriate degree is
- for additional degree, we add an additional variable to the regression model

Polynomial regression: Problems

- bad behaviour at the extremes of the predictor variable
- very bad out-of-sample behaviour (go off to infinity)
- coefficients become increasingly difficult to interpret
- easy to “overfit”

Overfitting

Bias-Variance tradeoff

Predictive accuracy

Within-sample vs. out-of sample prediction

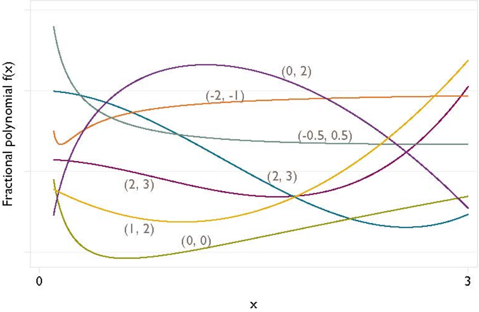
Which graph best predicts the datapoints?

- what is best?
- `https://ipsuit.shinyapps.io/splinedemo/`

out-of-sample prediction

- calculate error
- leave-one-out cross-validation

Model-selection



- FPs allow a large class of candidate models
- each of these models is fitted to produce the best parameters for this model
- how can we distinguish which of the many models is most appropriate?

Likelihood-ratio test

Likelihood

The "likelihood", $p(x|\theta)$ is the conditional probability that the data x will be observed given a model structure and a set of parameters θ.

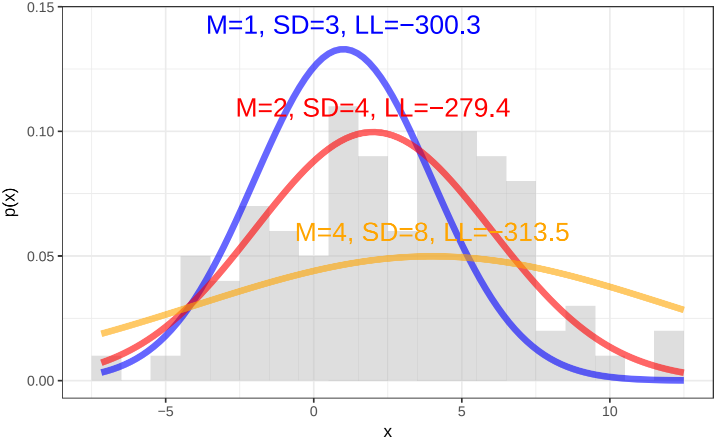
- usually, the logarithm is used and expressed as a function of the parameters

  $$L(\theta) = \log p(x|\theta)$$

  and we want to find the parameters that maximize this likelihood (maximum-likelihood)

  $$\hat{\theta} = \mathrm{argmax}_\theta L(\theta).$$

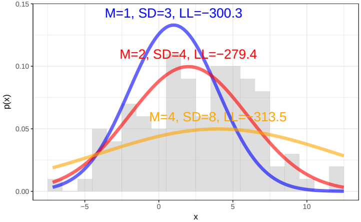
Likelihood



Likelihood

The “likelihood”, $p(x|\theta)$ is the conditional probability that the data x will be observed given a model structure and a set of parameters θ.

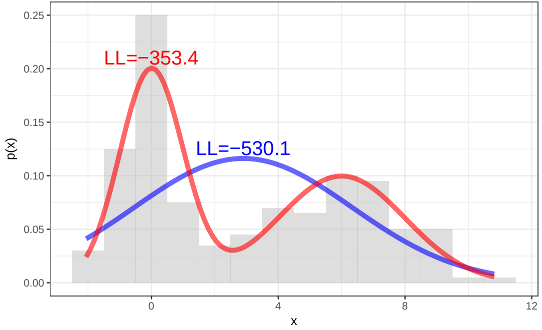
Likelihood



Examples:

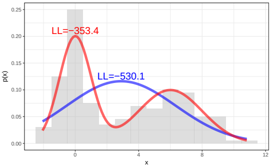
- calculating the mean and standard deviation of a sample is a maximum-likelihood estimation (we find $\hat{\theta} = (\mu, \sigma)$ that are most likely to underly the data)
- fitting a simple linear regression model is maximum-likelihood estimation, $\hat{\theta} = (b_0, b_1)$
- most other models are fit using ML estimation

Comparing Likelihoods across models



- assume two types of model, here:
  - a single normal distribution (blue) $\rightarrow$ parameters μ, σ
  - mixture of two normal distributions (red) $\rightarrow$ parameters $\mu_1, \sigma_1, \mu_2, \sigma_2$
- get ML estimate for each of the two model-types, LL_1, LL_2
- we can compare the likelihoods of those fits
- likelihood-ratio: $\frac{LL_1}{LL_2}$ quantifies difference

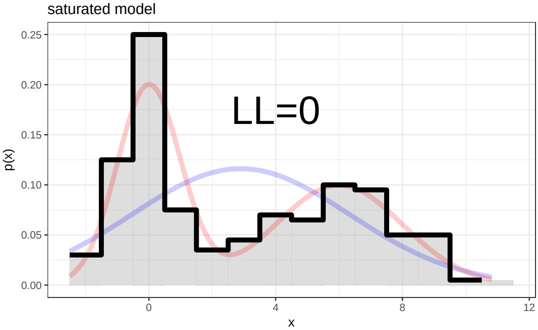
Likelihood-ratio test



Problem:

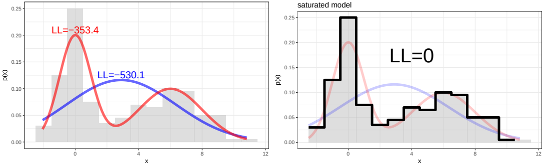
- if fit with ML, a model with more parameters is *guaranteed* to have higher LL
- choosing always the model with higher LL $\rightarrow$ always choose more complicated model
- results in always choosing a "saturated model"

Likelihood-ratio test



- model that predicts each point perfectly always has highest LL
- however, this model needs N parameters (one for each datapoint)
- maybe we want something simpler?

Likelihood-ratio test



Logic:

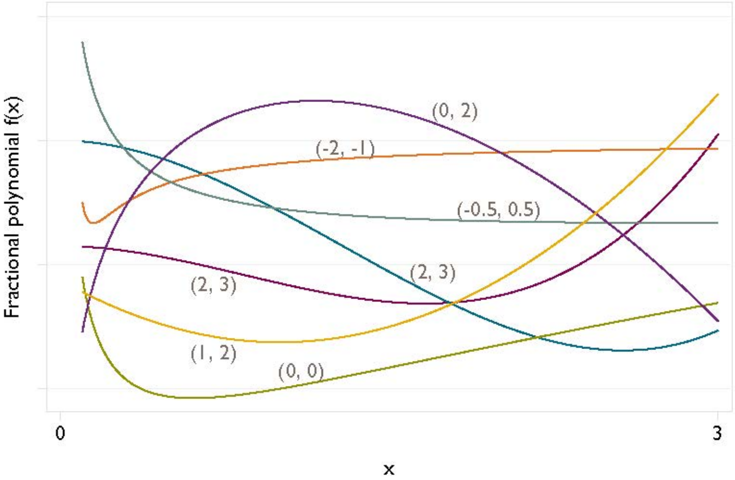
- adding more parameters always results in higher LL
- so $\frac{LL_2}{LL_1} > 1$ when model 2 has more parameters than model 1
- How much increase in LL would be expected *given that the real model is the simpler model*?
- the likelihood-ratio test, tests whether the increase in LL is significantly stronger than that

Fractional Polynomial Regression

Royston et al. 1994

- extends the idea of polynomial regression
- basic procedure restricts powers to a subset $-2, -1, -0.5, 0, 0.5, 1, 2, 3$

Fractional Polynomials



Summary: Fractional Polynomial Regression

- simultaneous selection of variables and transformations
- sometimes more parsimoneous:
  - variables that might be included to account for non-linearity can be dropped
- conservative test of non-linearity (can be emphasized by `select`-parameter)

References I

References I

Royston, Patrick and Douglas G Altman (1994). "Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling." In: *Applied statistics*, pp. 429–467.