Experimental Design and Statistical Methods

Workshop

CURVILINEAR REGRESSION

Jesús Piedrafita Arilla

jesus.piedrafita@uab.cat

Departament de Ciència Animal i dels Aliments

UAB

Items

- Curvilinear regression
  - Polynomial regression
  - Segmented regression
  - Non linear functions

- Basic commands
  - legend
  - lines
  - nls

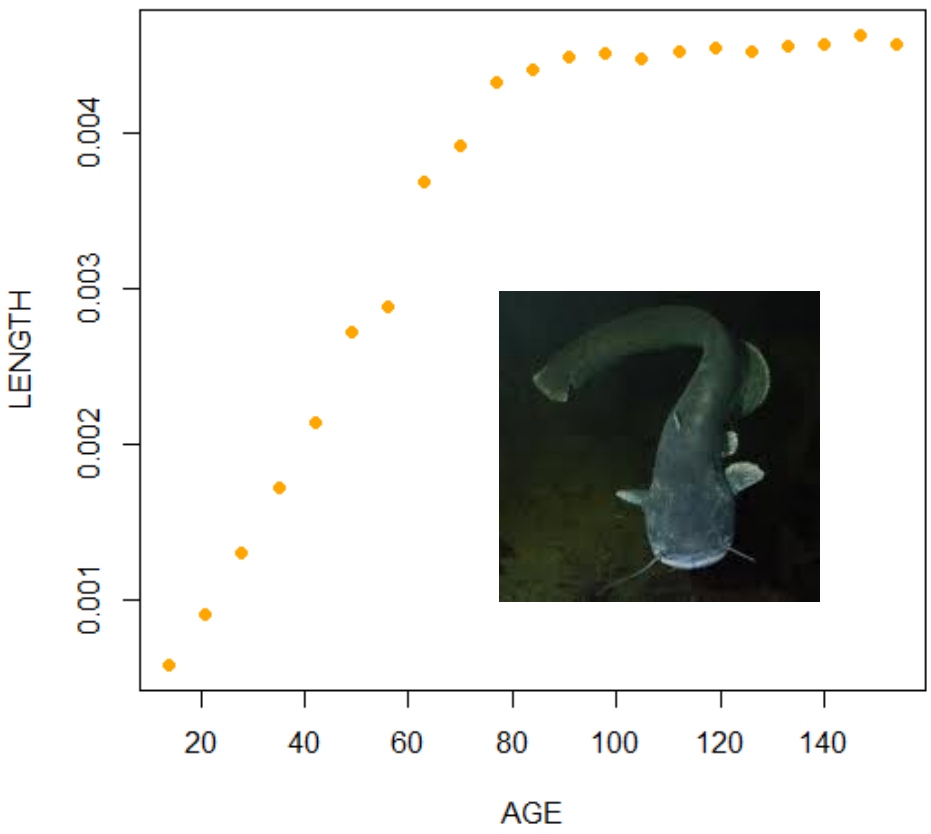
- Libraries
  - segmented

Growth data in fish

Length of a fish species recorded weekly (data from Freund and Littell, 1991):

Age (days)	Length (cm)
14	5.90
21	9.10
28	13.05
35	17.30
42	21.40
49	27.25
56	28.90
63	36.85
70	39.20
77	43.25
84	44.10
91	44.85
98	45.15
105	44.80
112	45.20
119	45.45
126	45.25
133	45.60
140	45.65
147	46.26
154	45.66

```
> plot(AGE, LENGTH, pch=19, col="orange")
```



No linear influence of independent on dependent variable

Polynomial regression

A one-variable polynomial model is defined as follows:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \ldots + \beta_m x_i^m + \varepsilon_i$$

This model is linear in the parameters (i.e., the first partial derivative for each parameter does not contain any of the parameters in the model), despite the relationship between *x* and *y* is not linear.

$\Rightarrow$ All parameters and statistics have the same connotation as in any linear regression analysis.

Polynomial models are used to fit a relatively smooth curve to a set of data.

It is customary to build an appropriate polynomial model by sequentially fitting equations with higher order terms.

Polynomial regression in matrix terms

Suppose we have a fourth order polynomial model:

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \dots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 & x_1^2 & x_1^3 & x_1^4 \\ 1 & x_2 & x_2^2 & x_2^3 & x_2^4 \\ 1 & x_3 & x_3^2 & x_3^3 & x_3^4 \\ \dots & \dots & \dots & \dots & \dots \\ 1 & x_n & x_n^2 & x_n^3 & x_n^4 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \dots \\ \varepsilon_n \end{bmatrix}$$

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

The normal equations to obtain the β estimates are the usual:

$$\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y} \quad \Rightarrow \quad \hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

Polynomial regression in R – first order -

```
> L.POL1<-lm(LENGTH ~ AGE)
> summary(L.POL1)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 11.43959    3.17371   3.604  0.00189 **
AGE          0.28341    0.03373   8.402 8.03e-08 ***

Residual standard error: 6.552 on 19 degrees of freedom
Multiple R-squared: 0.7879,     Adjusted R-squared: 0.7768
F-statistic:  70.6 on 1 and 19 DF,  p-value: 8.029e-08

> anova(L.POL1)

Analysis of Variance Table
Response: LENGTH
          Df  Sum Sq Mean Sq F value    Pr(>F)
AGE        1 3030.56 3030.56  70.595 8.029e-08 ***
Residuals 19  815.64
```

Polynomial regression in R – second order, quadratic -

```
> AGE2=AGE^2
> L.POL2<-lm(LENGTH ~ AGE + AGE2)
> summary(L.POL2)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -8.1664197  1.6706215  -4.888 0.000118 ***
AGE          0.9096909  0.0453168  20.074 9.04e-14 ***
AGE2        -0.0037279  0.0002632 -14.165 3.35e-11 ***

Residual standard error: 1.931 on 18 degrees of freedom
Multiple R-squared: 0.9825,     Adjusted R-squared: 0.9806
F-statistic: 506.5 on 2 and 18 DF,  p-value: < 2.2e-16

> anova(L.POL2)

Analysis of Variance Table
Response: LENGTH
          Df  Sum Sq Mean Sq F value    Pr(>F)
AGE        1 3030.56 3030.56  812.41 < 2.2e-16 ***
AGE2       1  748.50  748.50  200.65 3.347e-11 ***
Residuals 18   67.15    3.73
```

Polynomial regression in R – third order, cubic -

```
> AGE3=AGE^3
> L.POL3<-lm(LENGTH ~ AGE + AGE2 + AGE3)
> summary(L.POL3)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.037e+01  2.846e+00  -3.645  0.00200 **
AGE          1.033e+00  1.361e-01   7.589 7.43e-07 ***
AGE2        -5.457e-03  1.821e-03  -2.997  0.00811 **
AGE3         6.860e-06  7.150e-06   0.960  0.35075

Residual standard error: 1.936 on 17 degrees of freedom
Multiple R-squared: 0.9834,     Adjusted R-squared: 0.9805
F-statistic: 336.5 on 3 and 17 DF,  p-value: 2.481e-15

> anova(L.POL3)

Analysis of Variance Table
Response: LENGTH
          Df  Sum Sq Mean Sq  F value    Pr(>F)
AGE        1 3030.56 3030.56 808.8346 8.904e-16 ***
AGE2       1  748.50  748.50 199.7685 7.925e-11 ***
AGE3       1    3.45    3.45   0.9207    0.3507
Residuals 17   63.70    3.75
```

Polynomial regression in R – fourth order, quartic -

```
> AGE4=AGE^4
> L.POL4<-lm(LENGTH ~ AGE + AGE2 + AGE3 + AGE4)
> summary(L.POL4)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.802e+00  2.870e+00   0.976 0.343501
AGE         -2.803e-02  2.035e-01  -0.138 0.892189
AGE2         1.956e-02  4.537e-03   4.312 0.000537 ***
AGE3        -2.158e-04  3.945e-05  -5.472 5.12e-05 ***
AGE4         6.628e-07  1.167e-07   5.679 3.42e-05 ***

Residual standard error: 1.149 on 16 degrees of freedom
Multiple R-squared: 0.9945,     Adjusted R-squared: 0.9931
F-statistic: 724.3 on 4 and 16 DF,  p-value: < 2.2e-16

> anova(L.POL4)

Analysis of Variance Table
Response: LENGTH
          Df  Sum Sq Mean Sq   F value    Pr(>F)
AGE        1 3030.56 3030.56 2295.4596 < 2.2e-16 ***
AGE2       1  748.50  748.50  566.9399 6.404e-14 ***
AGE3       1    3.45    3.45    2.6129    0.1255
AGE4       1   42.57   42.57   32.2457 3.424e-05 ***
Residuals 16   21.12    1.32
```

Polynomial regression in R – summary -

The addition of the different terms is justified, although the cubic term only approached statistical significance. Polynomials beyond the fourth degree are not often used.

With the coefficients obtained we can construct the following regression equations:

Mean: $\hat{y}_i = 35.25$

Simple: $\hat{y}_i = 11.43959 + 0.283412x_i$

Quadratic: $\hat{y}_i = -8.16642 + 0.90969x_i - 0.00373x_i^2$

Cubic: $\hat{y}_i = -10.37000 + 1.03300x_i - 0.00546x_i^2 + 0.000007x_i^3$

Quartic: $\hat{y}_i = 2.80200 - 0.02803x_i + 0.01956x_i^2 - 0.000216x_i^3 + 0.00000066x_i^4$

The polynomial model is only used to approximate a curve, and the **coefficients have no practical interpretation**.

Polynomial regression in R – program for graphics -

```
> plot(AGE, LENGTH, pch=19, col="orange")
> lines(AGE, predict(L.POL1), type="l", col="maroon")
> lines(AGE, predict(L.POL2), type="l", col="red")
> lines(AGE, predict(L.POL3), type="l", col="darkblue")
> lines(AGE, predict(L.POL4), type="l", col="darkgreen")
> legend(100, 30, c("Linear", "Quadratic", "Cubic", "Quartic"),
pch=151, col=c("maroon", "red", "darkblue", "darkgreen"))
```

pch indicates the character to be printed (see character tables in Dalgaar, for example) and **col** the colour.

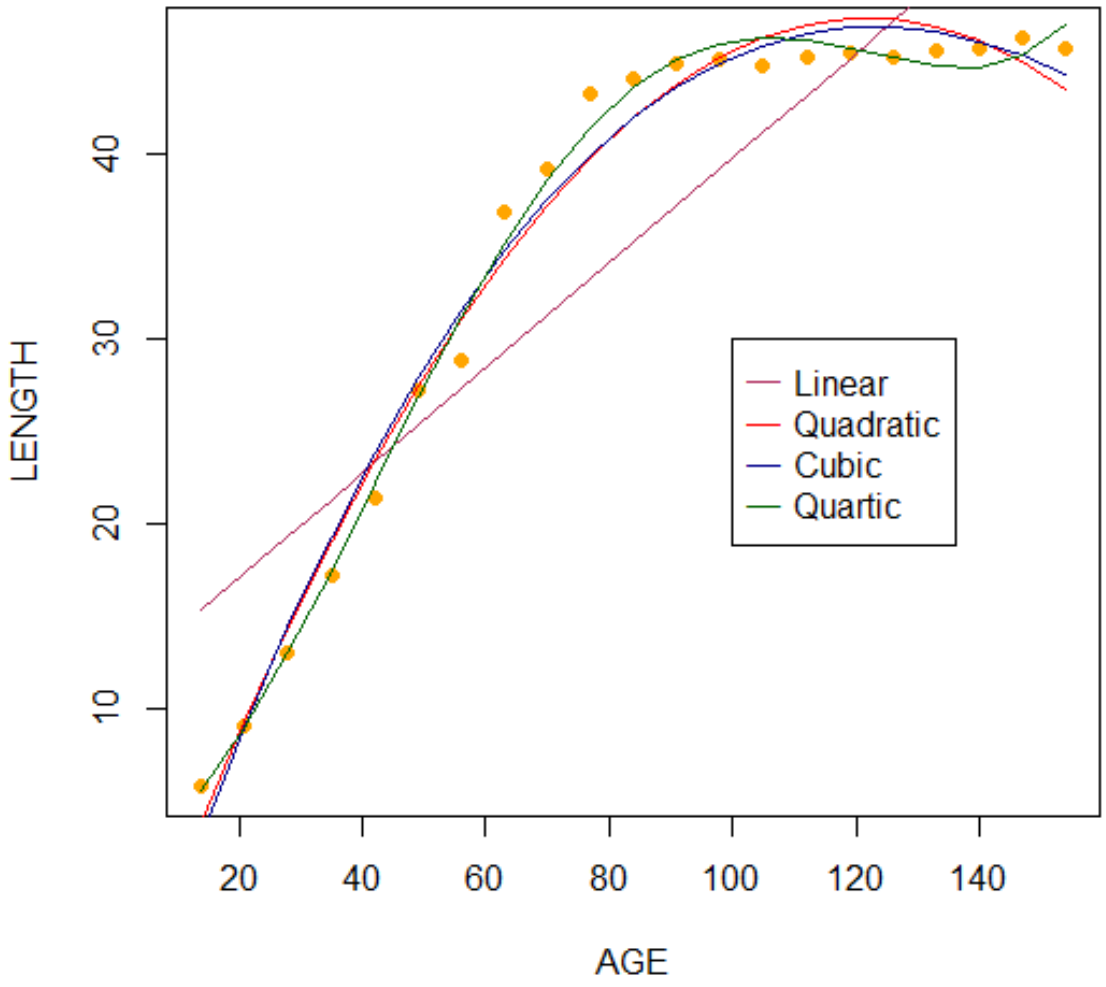
Additional graphics are added to the original plot with the command **lines**.

Predicted values in each model are obtained with **predict** applied to a lm object.

A line is specified with **type=l**.

legend(100, 30, ...) includes a legend at position x = 100 and y = 30 (upper left corner) with the specified content.

Polynomial regression in R – graphic -



Comments in the next slide

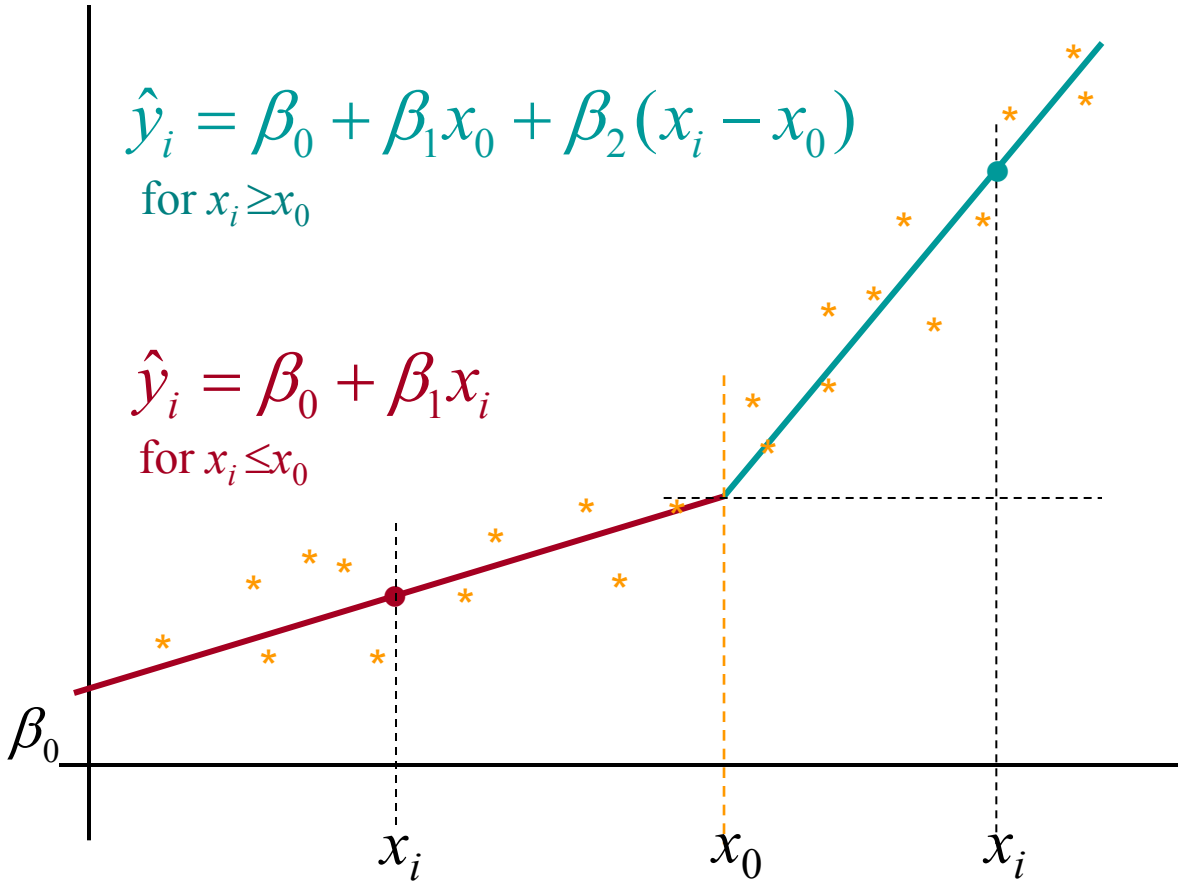
Polynomial regression in R – comments on the graphic -

- The plot clearly shows the limited improvement due to the cubic and fourth degree terms.
- The four degree curve shows a particular hook a the upper end that is not typical of growth curves.
- The quadratic curve shows negative growth in this region, which is unsatisfactory.
- In summary, the polynomial model could be unsatisfactory for this particular dataset.

$\Rightarrow$ Other approaches, such as **segmented regression** or **nonlinear regression** (different growth curves) must be explored.

Segmented (spline) regression

Some distributions of data can be fitted by several regressions, simple or polynomial. Let's see a simple example with a knot (x_0):



Splines are piecewise polynomials of degree n whose function values and first n-1 derivatives agree at points where they join.

Segmented regression in R – program -

We will explore now a spline model with a unknown knot.

```
> library(segmented)
> L.SEG<- segmented(lm(LENGTH~AGE), seg.Z=~AGE, psi=c(80))
> summary(L.SEG)
> anova(L.SEG)
> intercept(L.SEG)
> slope(L.SEG)
> confint(L.SEG)
```

Note that **segmented** works on a **lm** object.

seq.Z equals the model without the dependent variable.

psi specifies a guess of the value of the knot or knots we assume that can be the breakpoints of two or more regression lines. When the specified psi is close to the true knot, less iterations are needed. The approximated value of the knot can be guessed by inspection of the graphic.

Segmented regression in R – results 1 -

```
Estimated Break-Point(s):
   Est. St.Err
78.490  1.144
```

knot estimate

```
Meaningful coefficients of the linear terms:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.61182    0.60411  -5.979  1.5e-05 ***
AGE          0.61169    0.01214  50.368  < 2e-16 ***
U1.AGE      -0.59014    0.01607 -36.734       NA

Residual standard error: 0.7721 on 17 degrees of freedom
Multiple R-Squared: 0.9974,  Adjusted R-squared: 0.9969

Convergence attained in 2 iterations with relative change 2.082593e-15

Analysis of Variance Table
Response: LENGTH
          Df  Sum Sq Mean Sq F value Pr(>F)
AGE        1 3030.56 3030.56  5083.1 <2e-16 ***
U1.AGE     1  805.51  805.51  1351.1 <2e-16 ***
psi1.AGE   1    0.00    0.00     0.0      1
Residuals 17   10.14
```

Segmented regression in R – results 2 -

Parameter estimates of β's and the knot (with their CI):

```
$AGE
              Est.
intercept1 -3.612
intercept2 42.710
```

A negative length at time 0 (**intercept1**) is a nonsense

```
$AGE
          Est. St.Err. t value CI(95%).l CI(95%).u
slope1 0.61170 0.01214  50.370  0.586100   0.63730
slope2 0.02155 0.01052   2.049 -0.000644   0.04373
```

```
$AGE
  Est. CI(95%).l CI(95%).u
 78.49     76.07      80.9
```

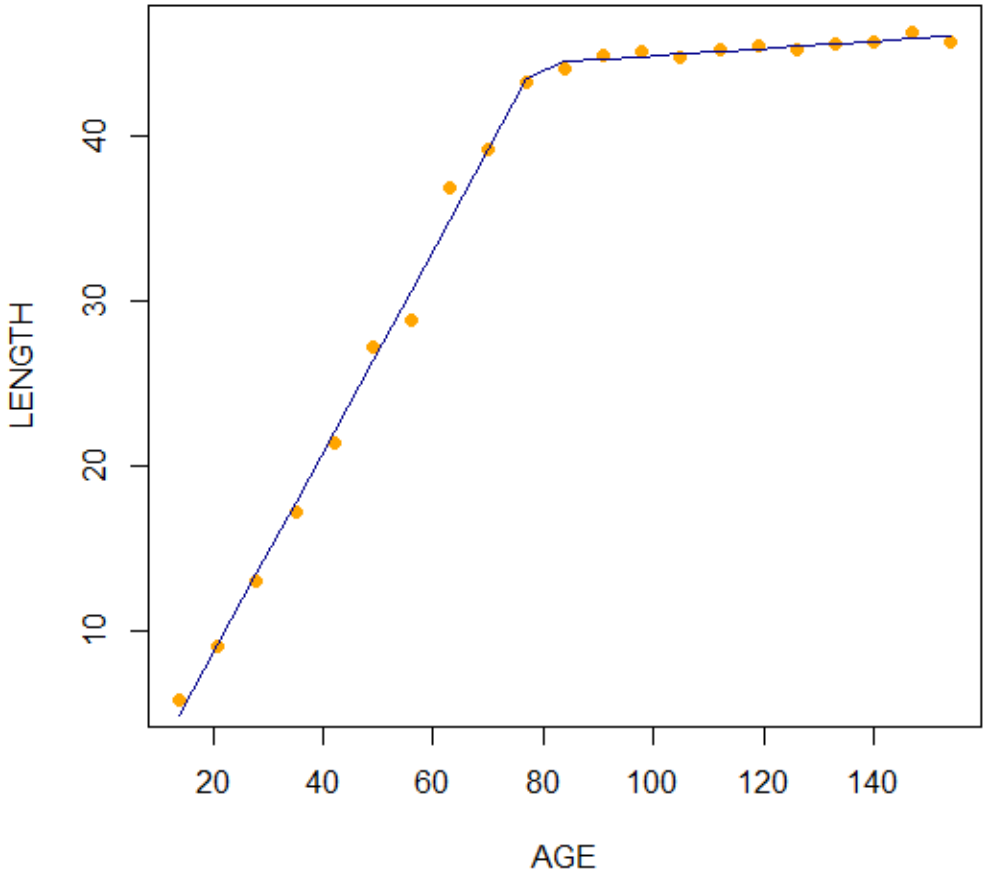
Note that **slope2 = AGE+U1.AGE** in the previous slide.
Its CI includes the 0 (not significant).

Beforetheknot: $\hat{y}_i = -3.612 + 0.6117 x_i$

Aftertheknot: $\hat{y}_i = 42.710 + 0.02155 x_i$

Segmented regression in R – graphic -

```
> plot(AGE, LENGTH, pch=19, col="orange")
> lines(AGE, predict(L.SEG), type="l", col="darkblue")
```



At first sight we have obtained a good and simple fit to our data, BUT …

Segmented regression in R – comments -

Some of the results may not be valid for models like the one fitted here, because the partial derivative of the model with respect to one of the parameters (knot) is not continuous (the curve has a discontinuity in the knot). In particular, the confidence intervals and standard errors may not be all correct.

In general, this discontinuity can disturb convergence of the iterative process and result in different final estimates depending on modest differences in specified starting values. An strategy could be to run the program with different starting points and see how parameter estimates vary.

Logistic growth curve

A **growth curve** usually shows a rapid initial growth that gradually becomes slower and may eventually cease.

Several **nonlinear** functions are used to describe growth trajectories in animals, among them, exponential (decay), Brody, Gompertz, … and the logistic growth curve:

$$y_i = \frac{k}{1+((k-y_0)/y_0)e^{-rt}} + \varepsilon_i$$

y_0 : expected value of y at time t = 0

k : height of the horizontal asymptote (the expected value of y for very large t)

r : measure of growth rate

ε_i : random error

In many nonlinear models the parameters represent meaningful quantities of the process described by the model.

The solutions to the equations are obtained by means of iterative processes.

Logistic growth curve in R – program -

```
> L.LOG<-nls(LENGTH ~ k/(1+((k-y0)/y0)*exp(-r*AGE)),
+            start=list(k = 40, y0 = 1, r = .05),
+            algorithm = "port")
> summary(L.LOG)
> confint(L.LOG)
```

nls allows to estimate parameters of some particular formula if we know some variables (response and independent variables).

Starting values are needed (**start=list(k = 40, y0 = 1, r = .05**). They are an arbitrary guess of the parameters or estimates of a recent run that had arbitrary starting values .

There are several algorithms available and the most appropriate one for our data set must be chosen.

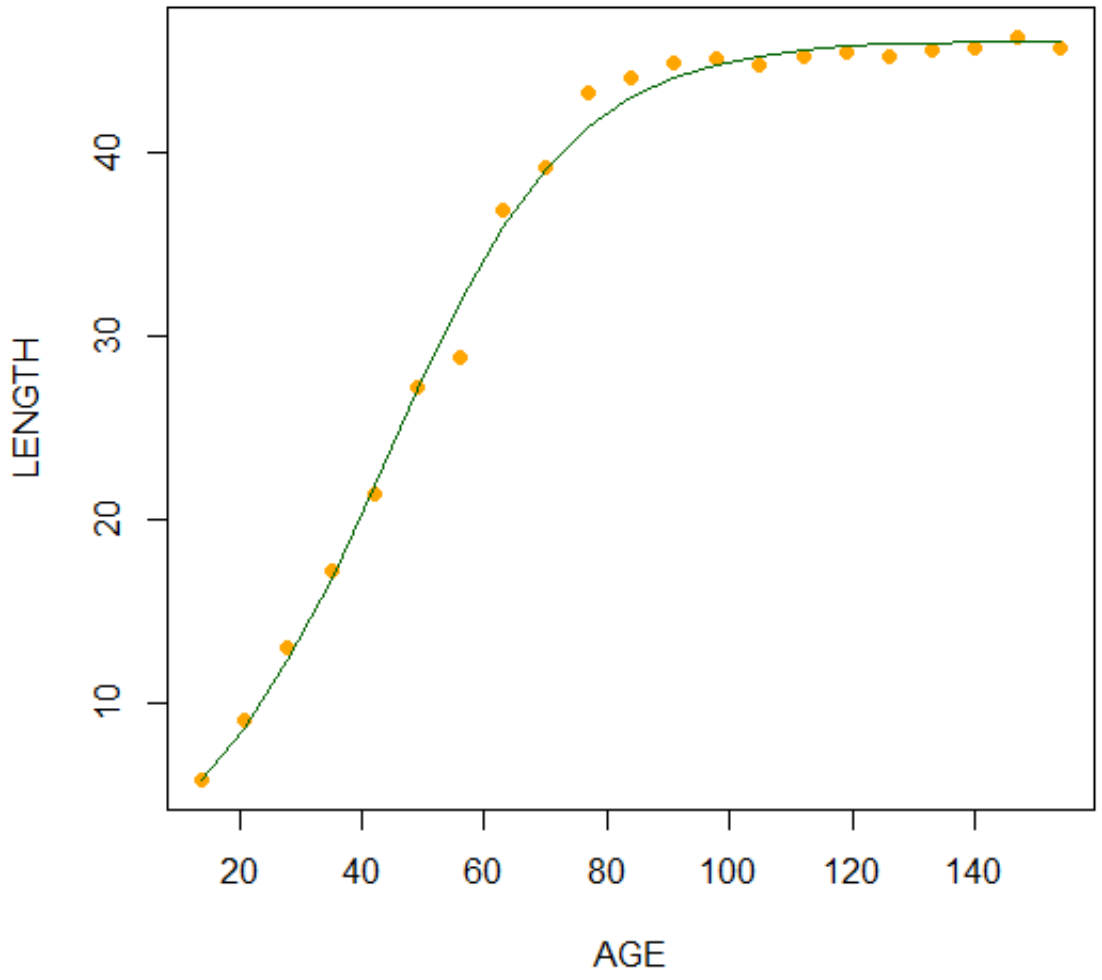
Logistic growth curve in R – results -

```
Parameters:
    Estimate Std. Error t value Pr(>|t|)
k  46.101579   0.351096 131.308  < 2e-16 ***
y0  2.571762   0.281782   9.127 3.57e-08 ***
r   0.065008   0.002615  24.863 2.18e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9941 on 18 degrees of freedom

Algorithm "port", convergence message: relative convergence (4)

          2.5%       97.5%
k  45.38498233 46.83617481
y0  2.03107612  3.19114640
r   0.05993243  0.07053084
```

Asymptotic standard errors and corresponding confidence intervals for the parameters show that they are estimated with useful precision.
The logistic growth curve fits quite well our data set.

Logistic growth curve in R – graphic -

```
> plot(AGE, LENGTH, pch=19, col="orange")
> lines(AGE, predict(L.LOG), type="l", col="darkgreen")
```



Final comments

It is difficult to make a choice among several models. A family of polynomial models has been discarded to fit well this data set, although the corresponding R-square estimates were very high.

MODEL	R-square	Adj. R-square	MSE
Simple regression	0.7879	0.7768	42.93
Quadratic regression	0.9825	0.9806	3.73
Cubic regression	0.9834	0.9805	3.75
Fourth order regression	0.9945	0.9931	1.32
Spline regression	0.9974	0.9969	0.6
Logistic growth curve			0.99

Another additional criterion is a lower MSE. This parameter clearly favours spline regression and logistic growth curve models. But remember that spline regression led to a nonsense estimate of β_0.

If we are interested in testing the predictive ability of our models, some additional analyses like Cross-validation would be recommended.

References

Freund R.J., Littell R.C. 1991. *SAS system for regression*, 2nd ed. SAS Institute, Cary, NC.