IV Brazilian Workshop of Plant Disease Epidemiology

Contemporary statistical methods for the plant epidemiology research with R

Walmes Marques Zeviani
walmes@ufpr.br
(LEG/DEST/UFPR)

Louise Larissa May De Mio
maydemio@ufpr.br
(LEMID/DFF/UFPR)

Paulo Justiniano Ribeiro Jr
paulojus@c3sl.ufpr.br
(LEG/DEST/UFPR)

December 4th, 2013

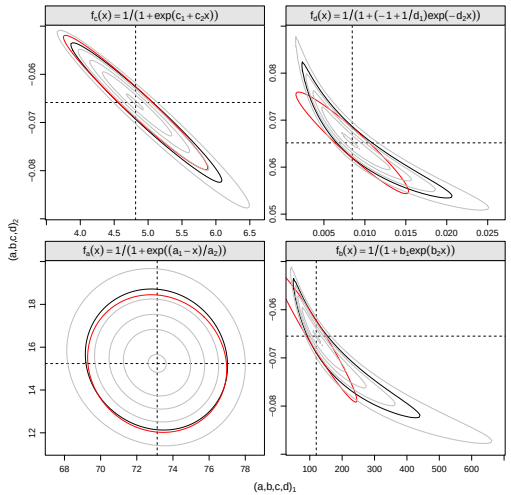
Outline

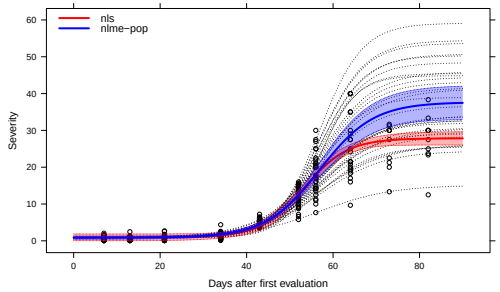
Issue - 1 Likelihood based inference vs frequentists for parametrizations of logistic model

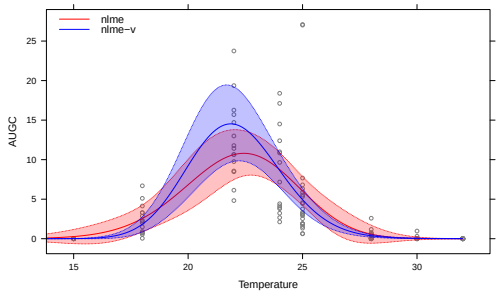
Issue - 2 Non linear mixed effects model for curve progress disease

Issue - 3 Modelling variance in non linear regression models for size-temperature data

Data, scripts and slides for download from (click on) www.leg.ufpr.br/~walmes/cursoR/wpde.zip



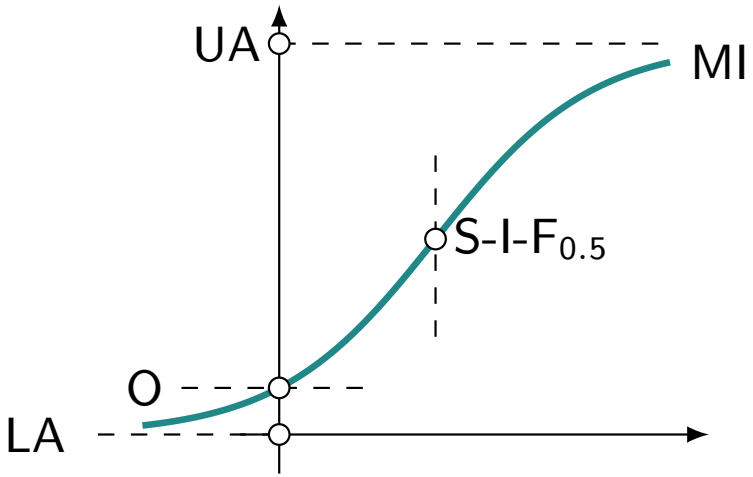




Issue 1

Likelihood based inference vs frequentists for parametrizations of logistic model

Logistic Model



- The most known "S" shaped model;
- The most used non linear model to describe disease progress curve;
- Extensive application contributed to the emergence of several parameterizations (2-4 parameters);
- Each parametrization has a particular interpretation;
- Is a function with several notable points that can be parameters;
- From statistical point of view, they have different aspects;
- What consider when select a parametrization?

Some parametrizations

General model

$$\eta(x, \theta) \propto \frac{1}{1 + \exp\{g(x, \theta)\}} \tag{1}$$

Parametrizations from Ratkowsky (1983) and Madden, Hughes e Bosch (2007)

$$\eta_a(x) = \frac{1}{1 + \exp\{-(x - a_1)/a_2\}} \tag{2}$$

$$\eta_b(x) = \frac{1}{1 + b_1 \exp\{b_2 x\}} \tag{3}$$

$$\eta_c(x) = \frac{1}{1 + \exp\{c_1 + c_2 x\}} \tag{4}$$

$$\eta_d(x) = \frac{1}{1 + \left(-1 + \frac{1}{d_1}\right) \exp\{-d_2 x\}} \tag{5}$$

Interpretation and relations

Parameter	Unit	Parametrization			
		A	B	C	D
a_1	x	-	$a_2 \log(b_1)$	$a_2 c_1$	$a_2 \log\left(-1 + \frac{1}{d_1}\right)$
a_2	x	-	$-\frac{1}{b_2}$	$-\frac{1}{c_2}$	$\frac{1}{d_2}$
b_1	$\varnothing$	$\exp\left\{\frac{a_1}{a_2}\right\}$	-	$\exp\{c_1\}$	$-1 + \frac{1}{d_1}$
b_2	x^{-1}	$-\frac{1}{a_2}$	-	c_2	$-d_2$
c_1	$\varnothing$	$\frac{a_1}{a_2}$	$\log(b_1)$	-	$\log\left(-1 + \frac{1}{d_1}\right)$
c_2	x^{-1}	$-\frac{1}{a_2}$	b_2	-	$-d_2$
d_1	$\varnothing$	$\frac{1}{1 + \exp\left\{\frac{a_1}{a_2}\right\}}$	$\frac{1}{1 + b_1}$	$\frac{1}{1 + \exp\{c_1\}}$	-
d_2	x^{-1}	$\frac{1}{a_2}$	$-b_2$	$-c_2$	-

Objectives

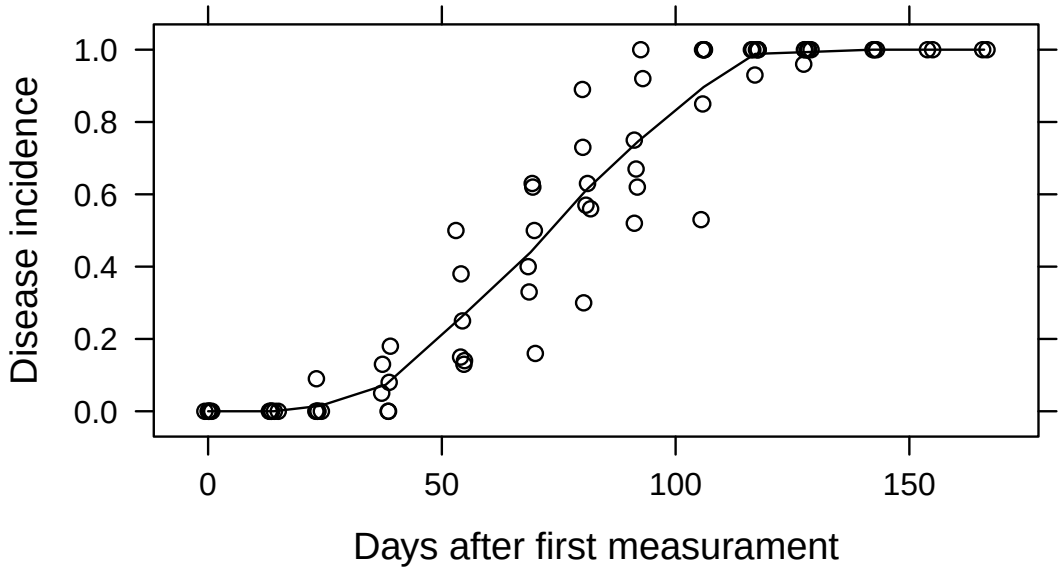
Topics

- Most of the software to non linear regression has standard output for inference based on frequentists approach;
- For linear models, frequentists and likelihood inference are equivalent once the likelihood is quadratic on the parameters;
- Non linear models has a non quadratic likelihood function on the parameters;
- For small sample sizes likelihood and frequentists approach can give different results.

Objectives

Study these parametrizations comparing the traditional methods for inference (frequentists) with the likelihood based inference;

Data set



- Incidence over time;
- Rust peach (*Tranzschelia discolor*);
- Cultivar Chimarrita;
- Six plants with (a non known number of) leaves evaluated at each time;

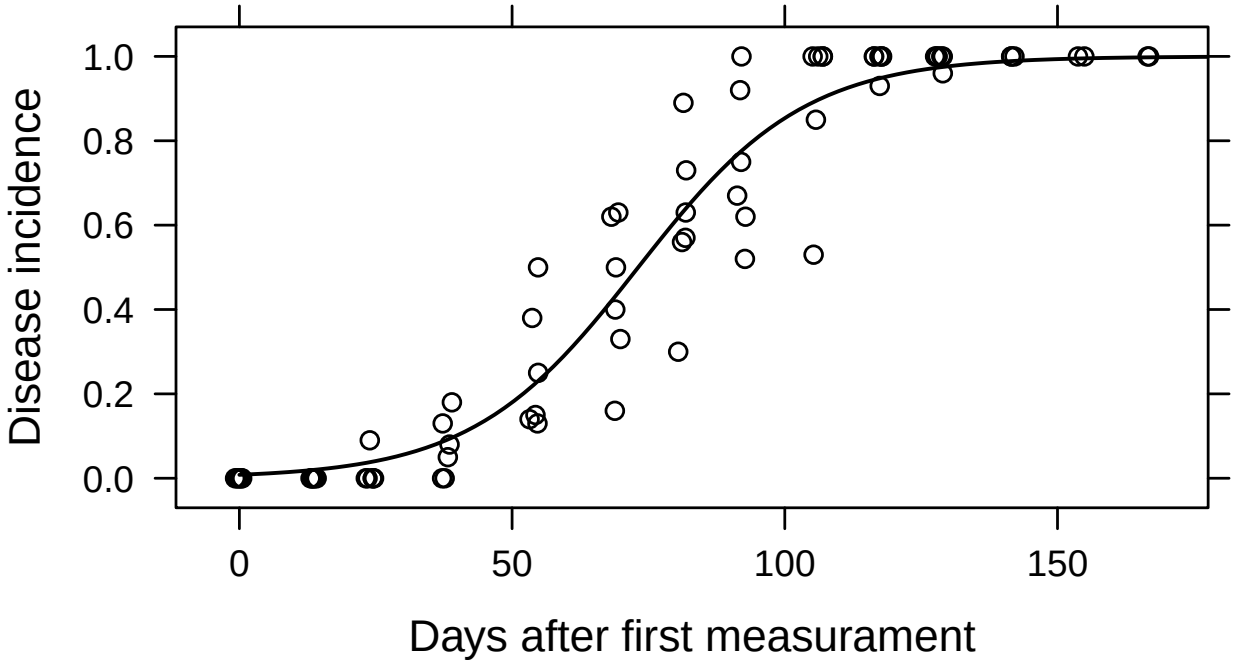
Likelihood based model inference with R

```
##-------------------------------------------------------------------------
## log-likelihood funtion
                                                                        ##
ll <- function(th, y, x, model, C=1){
  ex <- do.call(model, list(x=x, th=th))
  sd <- sqrt(crossprod(y-ex)/length(x))
  ll <- sum(dnorm(y, mean=ex, sd=sd, log=TRUE))
  ll*C
}
                                                                        ##
##-------------------------------------------------------------------------
## models written in vectorized form
                                                                        ##
f.a <- function(x, th){ 1/(1+exp((th[1]-x)/th[2])) }
f.b <- function(x, th){ 1/(1+th[1]*exp(th[2]*x)) }
f.c <- function(x, th){ 1/(1+exp(th[1]+th[2]*x)) }
f.d <- function(x, th){ 1/(1+(-1+1/th[1])*exp(-th[2]*x)) }
                                                                        ##
##-------------------------------------------------------------------------
```

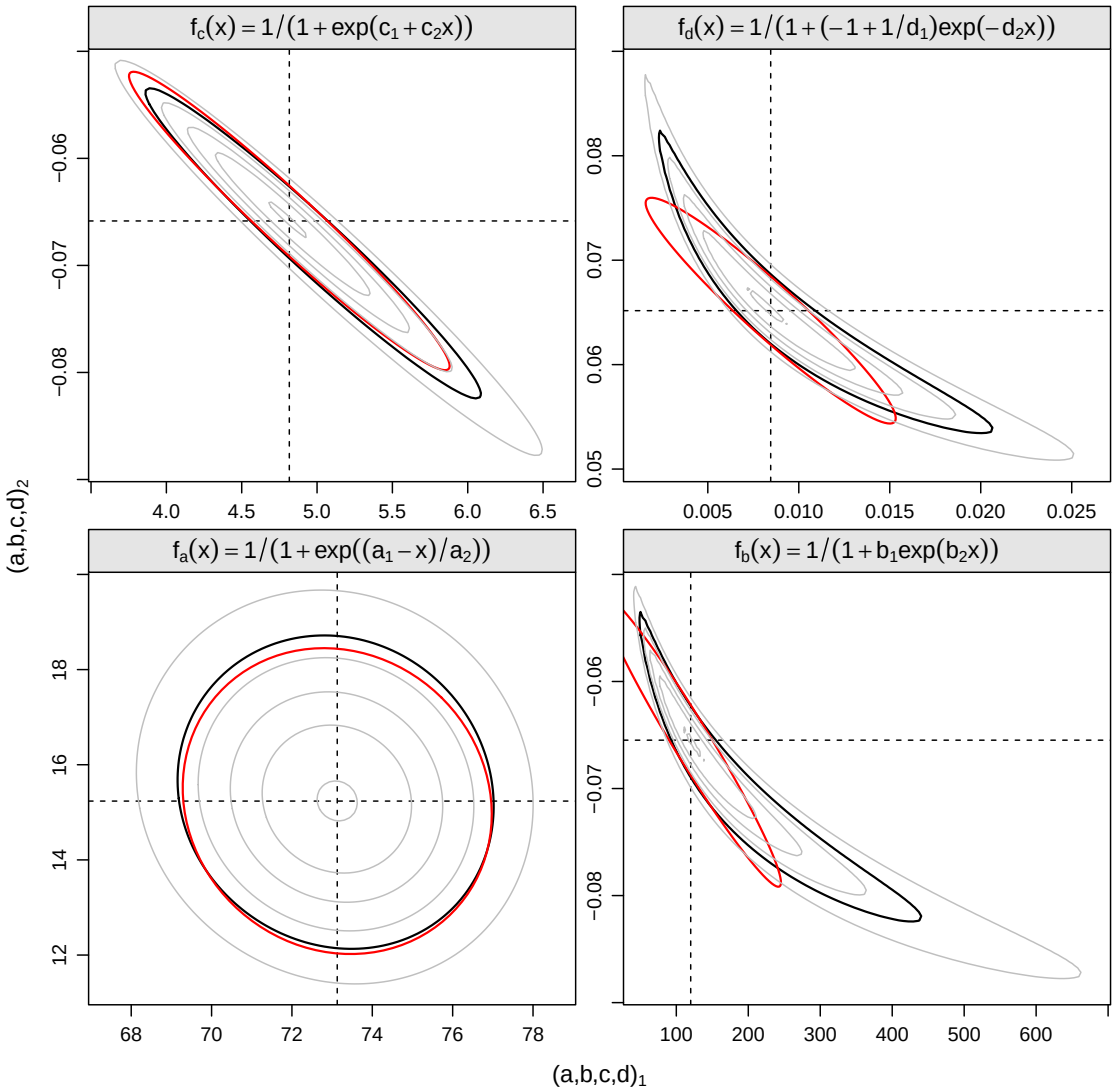
Results and discussion: *Invariant results*

```
load("./scripts/parametrizations.RData")
## parameter estimates
pars <- sapply(op.all, "[[", "par"); pars


##           A         B        C        D
## [1,] 73.13 120.01001  4.81593 0.008456
## [2,] 15.24  -0.06548 -0.06585 0.065167


## log-likelihood
llo <- sapply(op.all, "[[", "value"); llo


##     A     B     C     D
## 57.14 57.14 57.14 57.13
```

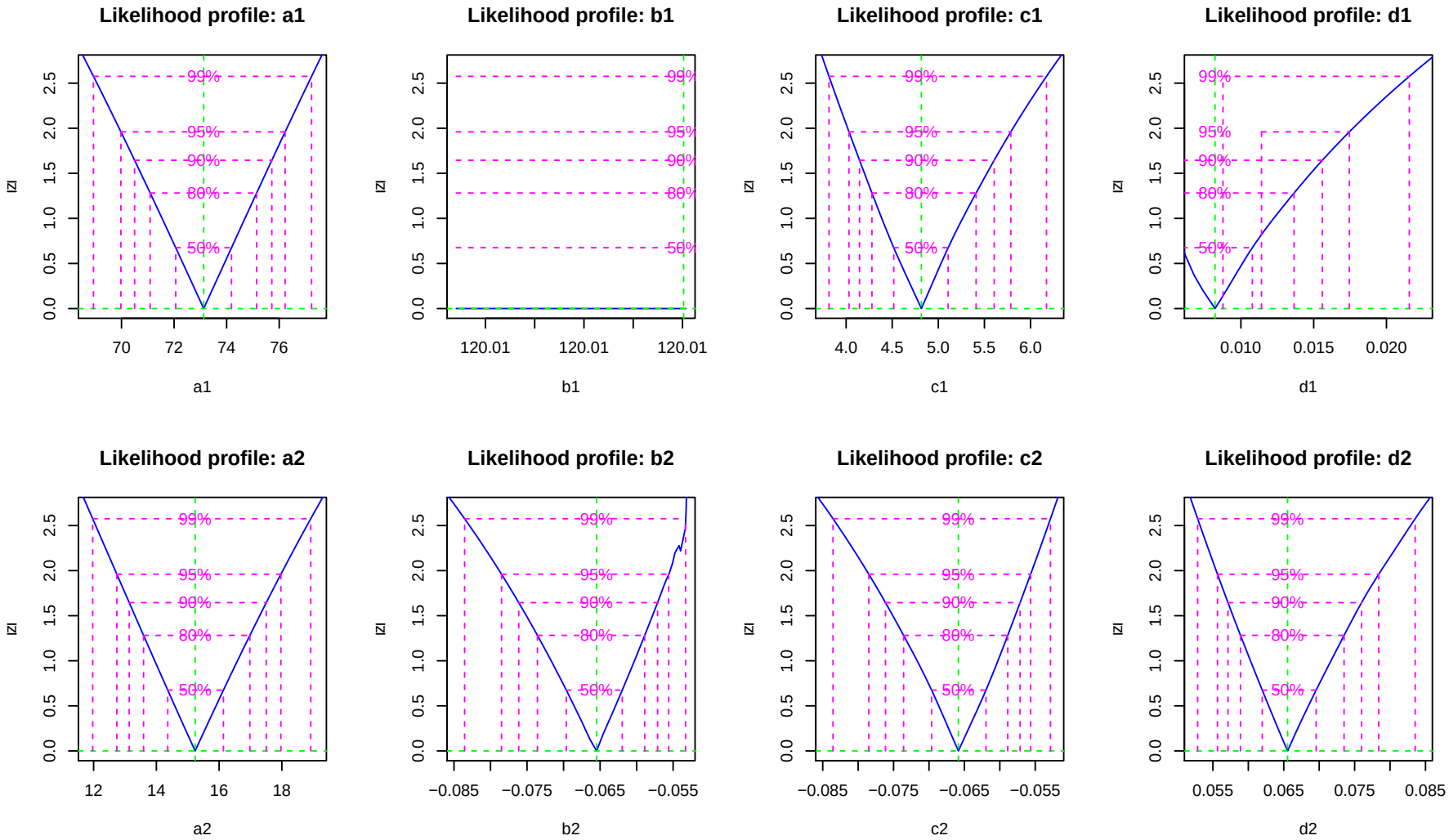
Results and discussion: *Invariant results*



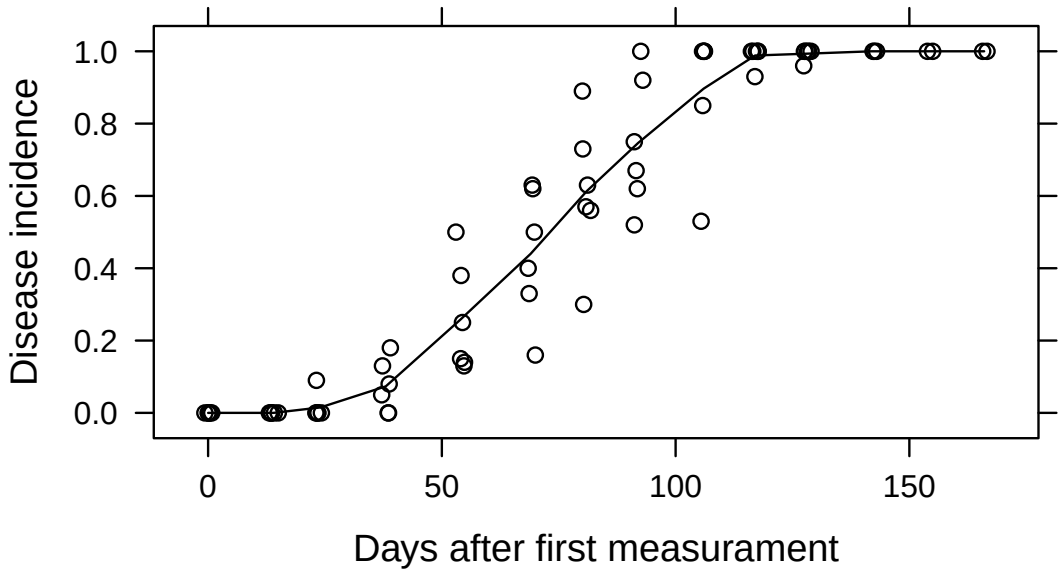
Results and discussion: *Parametrization dependent results*



Results and discussion: *Parametrization dependent results*

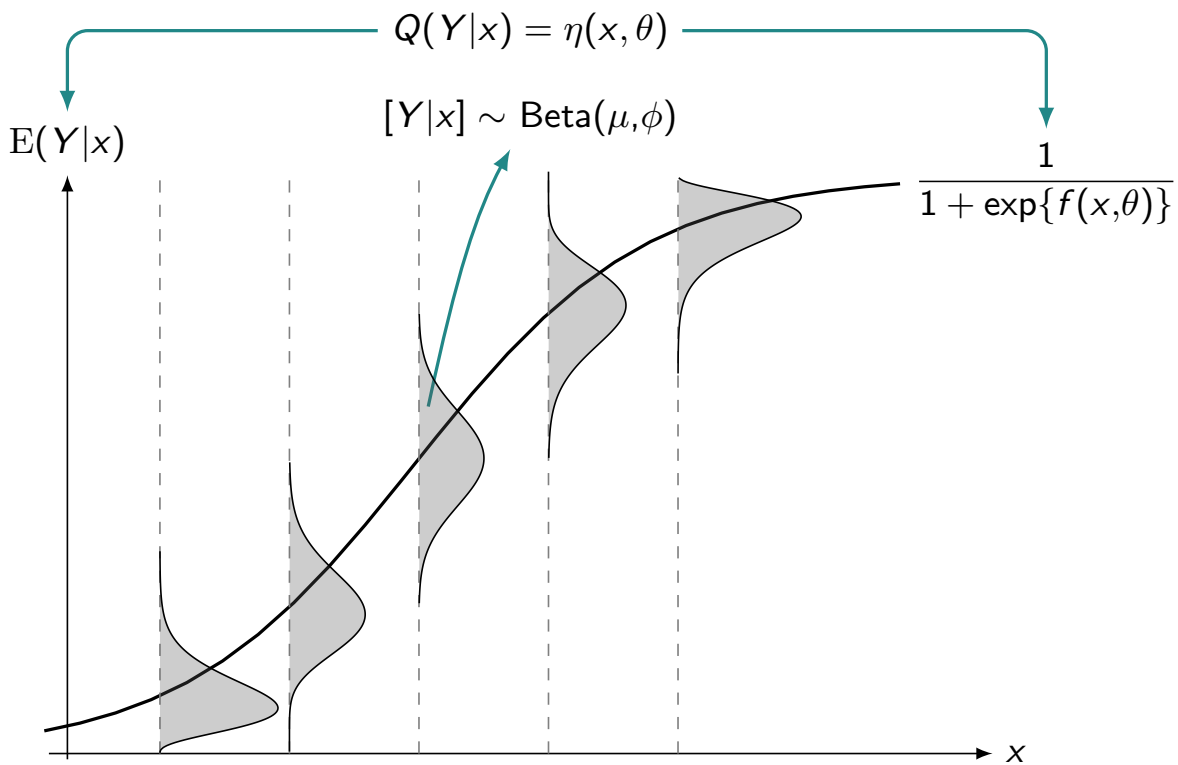


A closer examination



- Response limited to unit interval;
- Variance mean relation;
- Skewness.

Further improvement



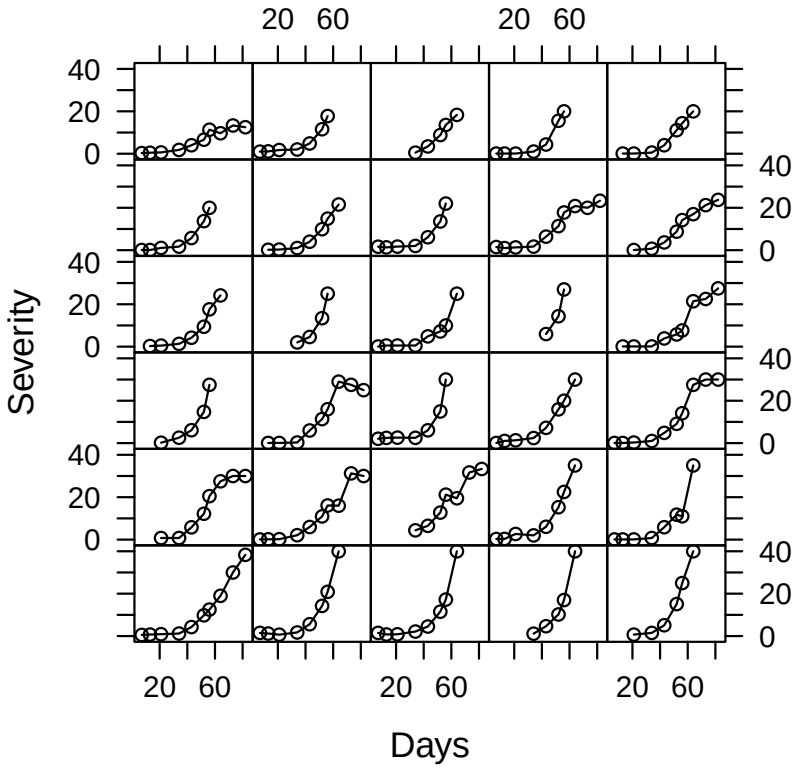
Conclusions

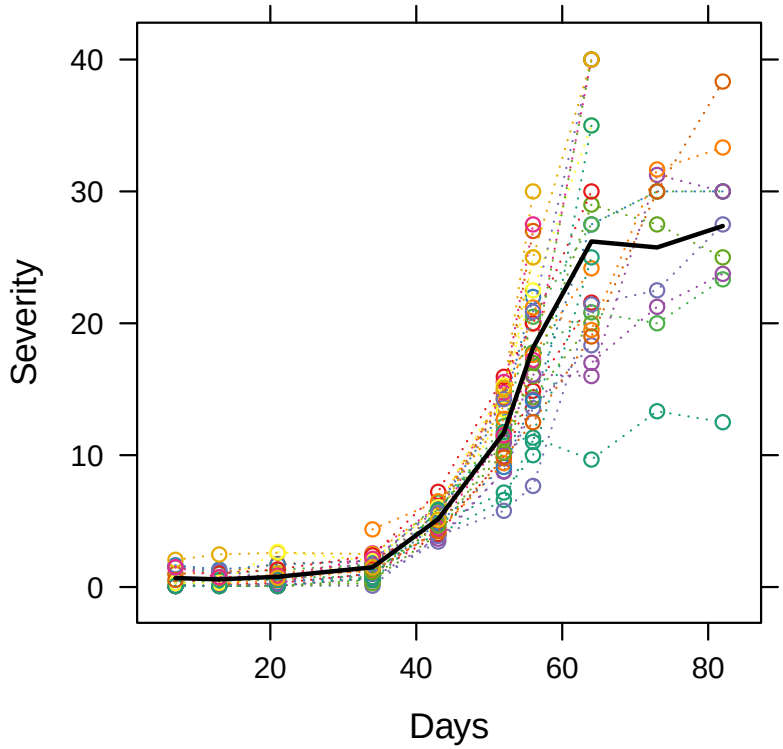
- Inference based on approximation (frequentists) for nonlinear models should be conducted carefully;
- A proper parametrization should be used frequentists inference;
- Delta method could be used to go from a paramtrization to functions of model parameters, but is still based on linear approximations;
- Orthogonal and symmetric likelihood function are desirable properties for numerical methods used for estimation and inference;
- It is possible take advantage of a parametrization for estimation and use another for inference.

Issue 2

Non linear mixed effects model for curve progress disease

Data set


Severity
Days
20 60
40
20
0


Severity
Days
0
10
20
30
40
20
40
60
80

Data set

- Longitudinal study;
- Experimental unit are plants with marked leaves measured over time;
- The between plant effect could not be assumed to be null;
- Leaves are subject to disease occurrence;
- The leaves do not fall randomly, its fall is possibly due to the level of disease;
- So the data is not missing at random;
- Do not take those characteristics into account compromises inferences;

Modelling approaches

It will be used the four parameter logistic centre-scale model.

- Ordinary non linear regression (ONLR);

$$\eta(x, \theta) = \theta_l + \frac{\theta_u}{1 + \exp\{-(x - \theta_i)/\theta_s\}}.$$

- Nonlinear random effects model (NREM);

$$\eta(x, \theta, b_i) = \theta_l + \frac{(\theta_u + b_{1i})}{1 + \exp\{-(x - \theta_i)/(\theta_s + b_{2i})\}}.$$

- Estimation by maximum likelihood;

Likelihood functions *(don't panic!)*

- Ordinary nonlinear regression

$$\mathrm{L}(\theta,\sigma^2) = \prod_{i=1}^{I}\prod_{j=1}^{n_i} \phi(y_{ij},\eta(x_{ij},\theta),\sigma^2); \tag{6}$$

  - $y_{ij}|x_{ij} \sim \text{Gaussian}(\eta(x_{ij},\theta),\sigma^2)$;
- Nonlinear random effects model

$$\mathrm{L}(\theta,\Psi,\sigma^2) = \prod_{i=1}^{I}\int\prod_{j=1}^{n_i} \phi(y_{ij},\eta(x_{ij},\theta,b_i),\sigma^2) \cdot \phi(b_i,0,\Psi)\,\mathrm{d}b_i; \tag{7}$$

  - $y_{ij}|x_{ij}, b_i \sim \text{Gaussian}(\eta(x_{ij},\theta, b_i),\sigma^2)$;
  - $b_i \sim \text{Gaussian}_k(0,\Psi)$;
  - $y_{ij}|x_{ij}, b_i$ e b_i são independentes.
- In simple words, these likelihood functions are very close to the **completely randomized design** and **complete randomized block design** models used to analyse experiments.

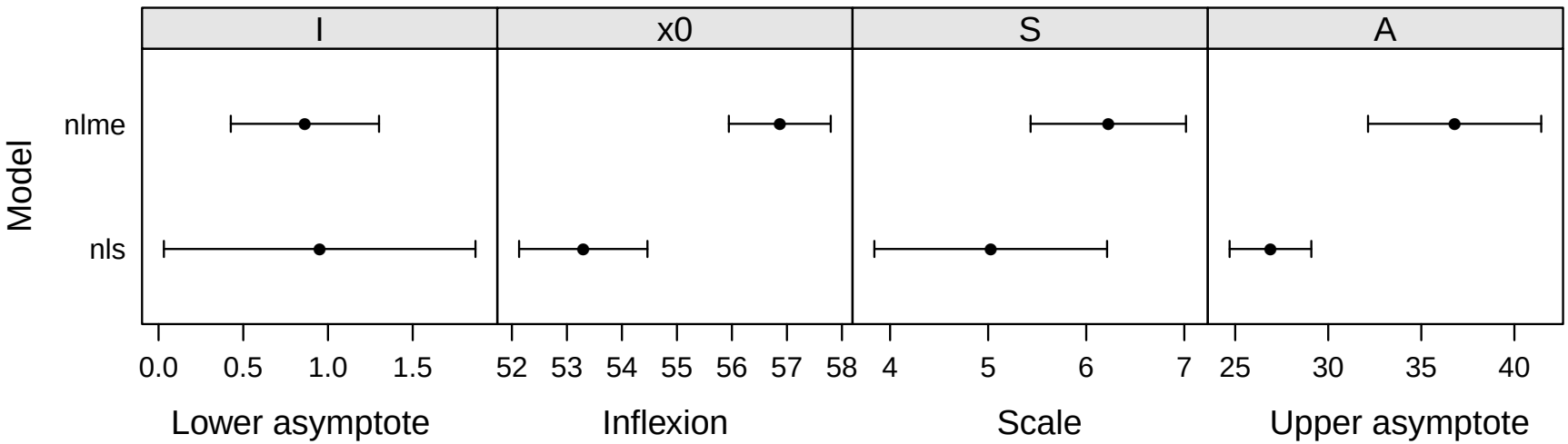
Using R to fit these models

```
##-------------------------------------------------------------------------
                                                                        ##
plot(y~x, da)
mystart <- list(I=1, A=30, x0=55, S=7)
with(mystart, curve(I+(A-I)/(1+exp(-(x-x0)/S)), add=TRUE, col=2))
                                                                        ##
##-------------------------------------------------------------------------
## Fitting the naive model.
##
                                                                        ##
n0 <- nls(y~I+A/(1+exp(-(x-x0)/S)), data=da,
          start=mystart, trace=TRUE)
                                                                        ##
##-------------------------------------------------------------------------
## Fitting the non linear random effects model.
                                                                        ##
dd <- groupedData(y~x|ar, data=da, order.groups=FALSE)
                                                                        ##
n00 <- nlme(y~I+A/(1+exp(-(x-x0)/S)),
            start=coef(n0), fixed=I+A+x0+S~1,
            random=A+S~1, data=dd) # converge
                                                                        ##
##-------------------------------------------------------------------------
```

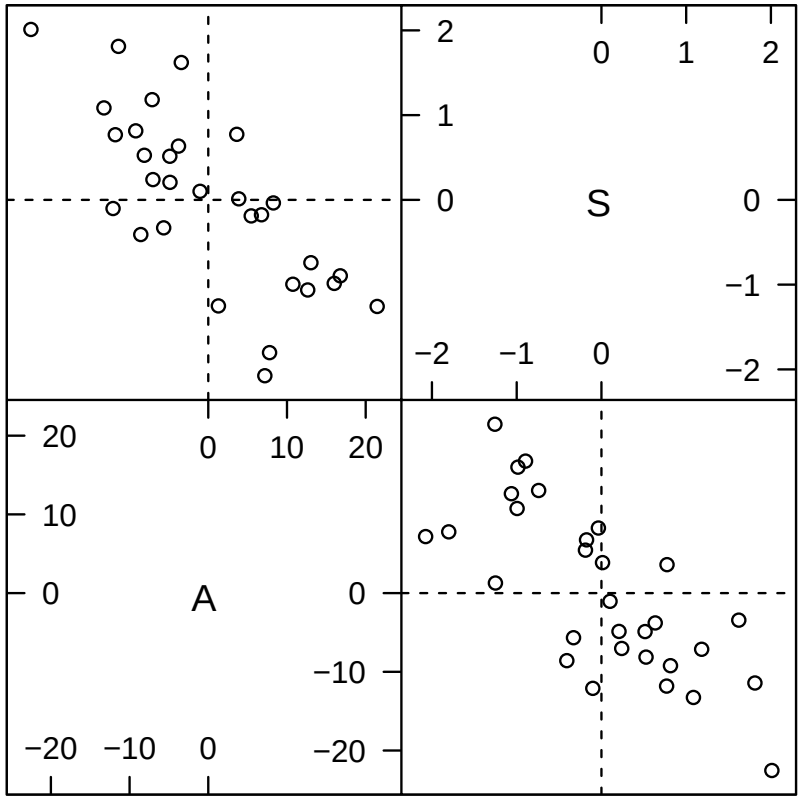
Results: *Standard output for nlme model*

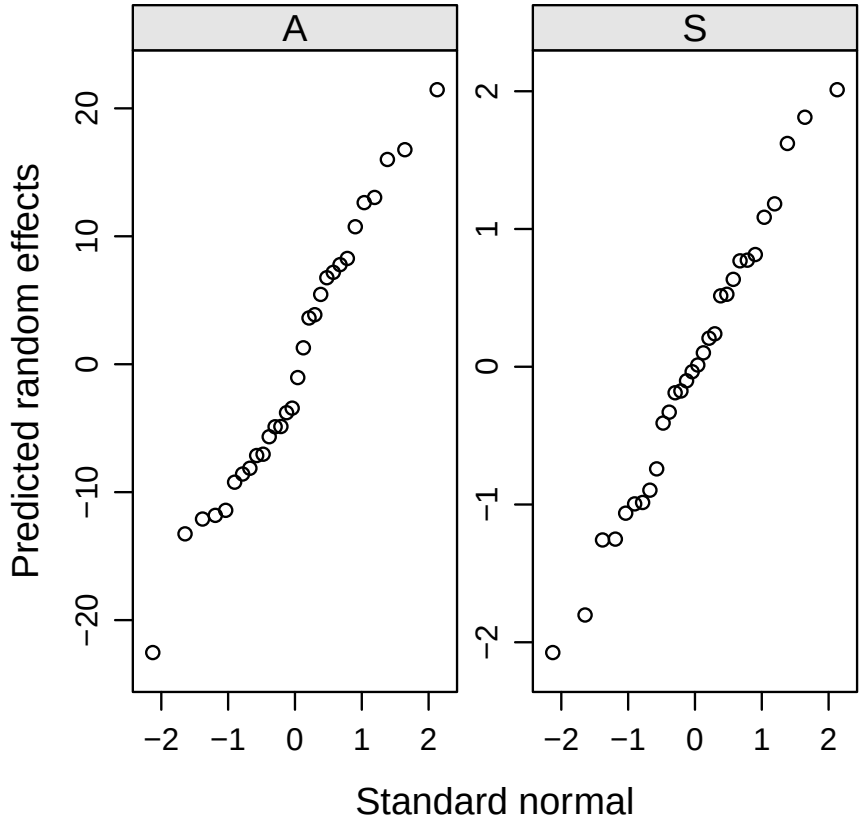
```
Nonlinear mixed-effects model fit by maximum likelihood
  Model: y ~ I + A/(1 + exp(-(x - x0)/S))
 Data: dd
       AIC      BIC    logLik
  1068.555 1095.813 -526.2777

Random effects:
 Formula: list(A ~ 1, S ~ 1)
 Level: ar
 Structure: General positive-definite, Log-Cholesky parametrization
         StdDev    Corr
A        10.712344 A
S         1.261347 -0.614
Residual  1.948165

Fixed effects: I + A + x0 + S ~ 1
      Value Std.Error  DF   t-value p-value
I   0.86381 0.2237778 190   3.86012   2e-04
A  36.79136 2.3798492 190  15.45953   0e+00
x0 56.87014 0.4738078 190 120.02787   0e+00
S   6.22496 0.4053498 190  15.35701   0e+00

Number of Observations: 223
Number of Groups: 30
```

Results: *Point estimates and confidence intervals*

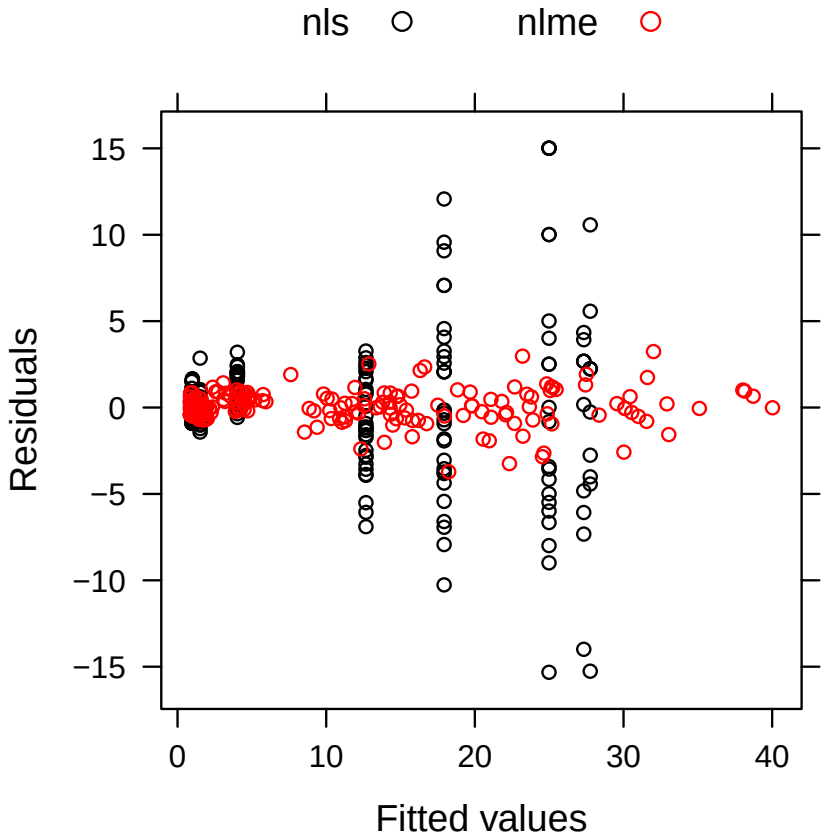


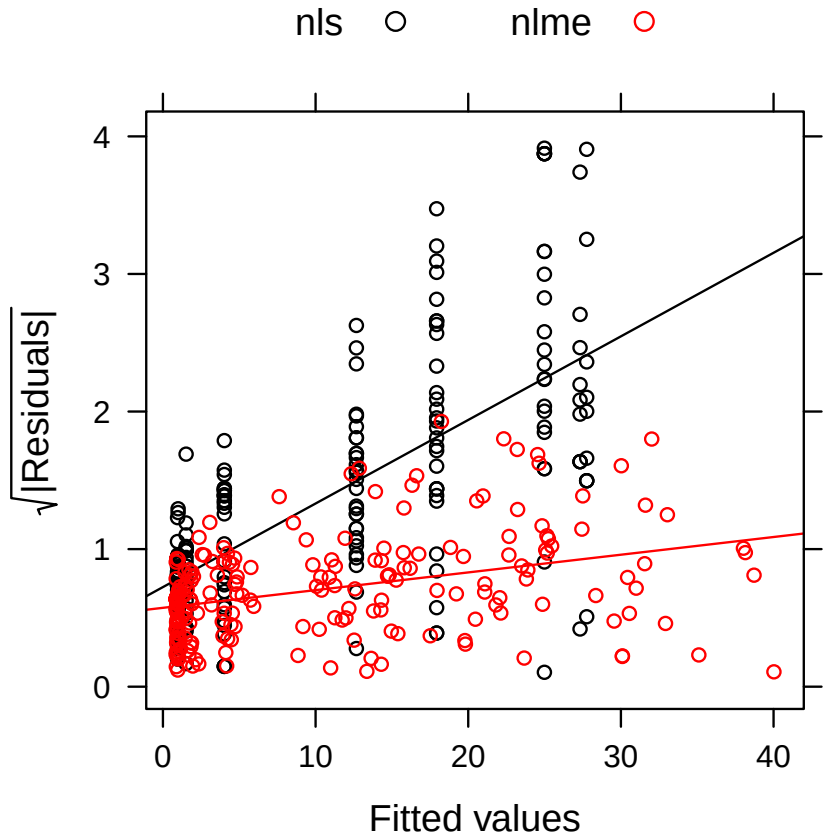
Results: *Random effects distribution*





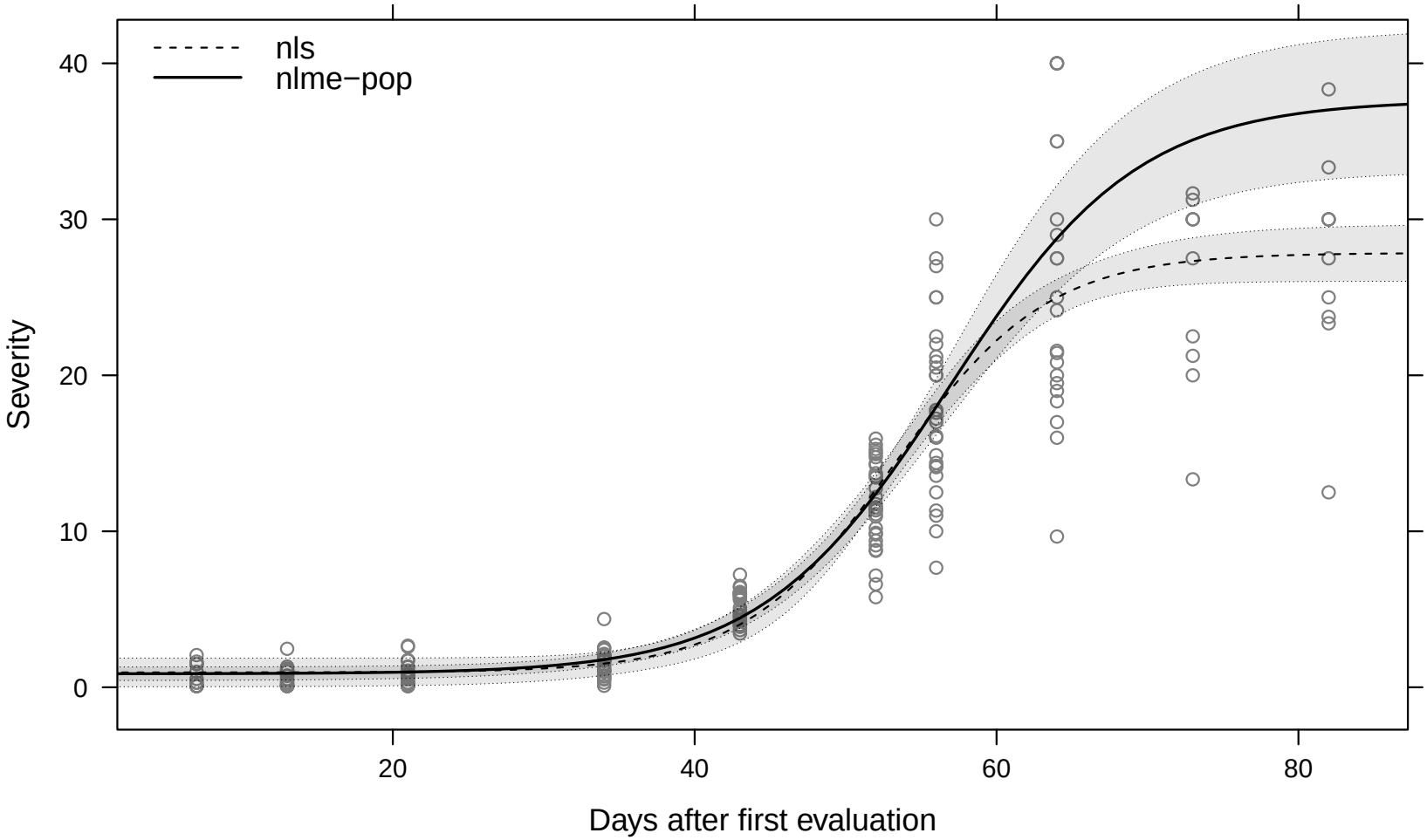
Results: *Residuals diagnostics*





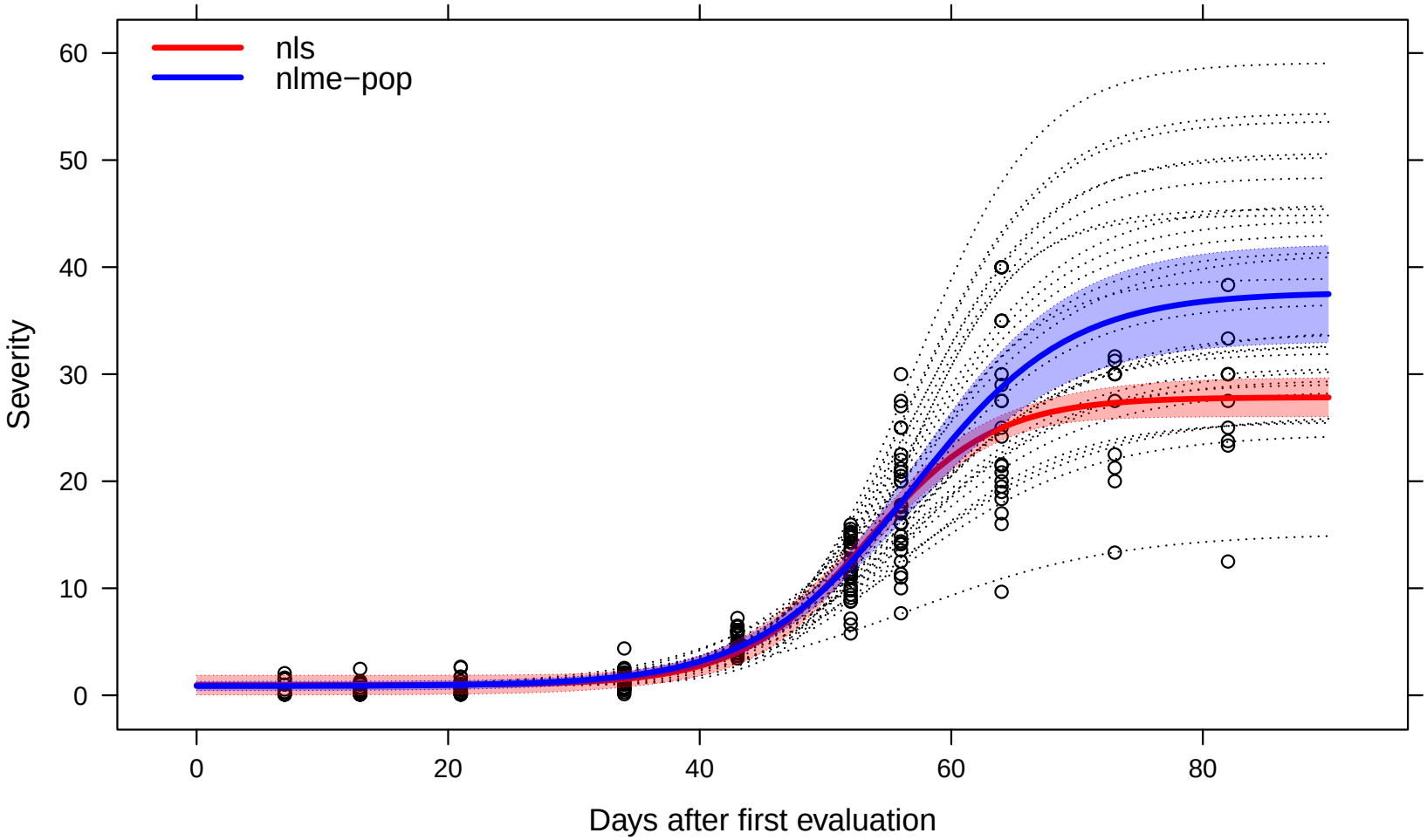
Results: *Observed vs. fitted values*

Does the fit was wrong?

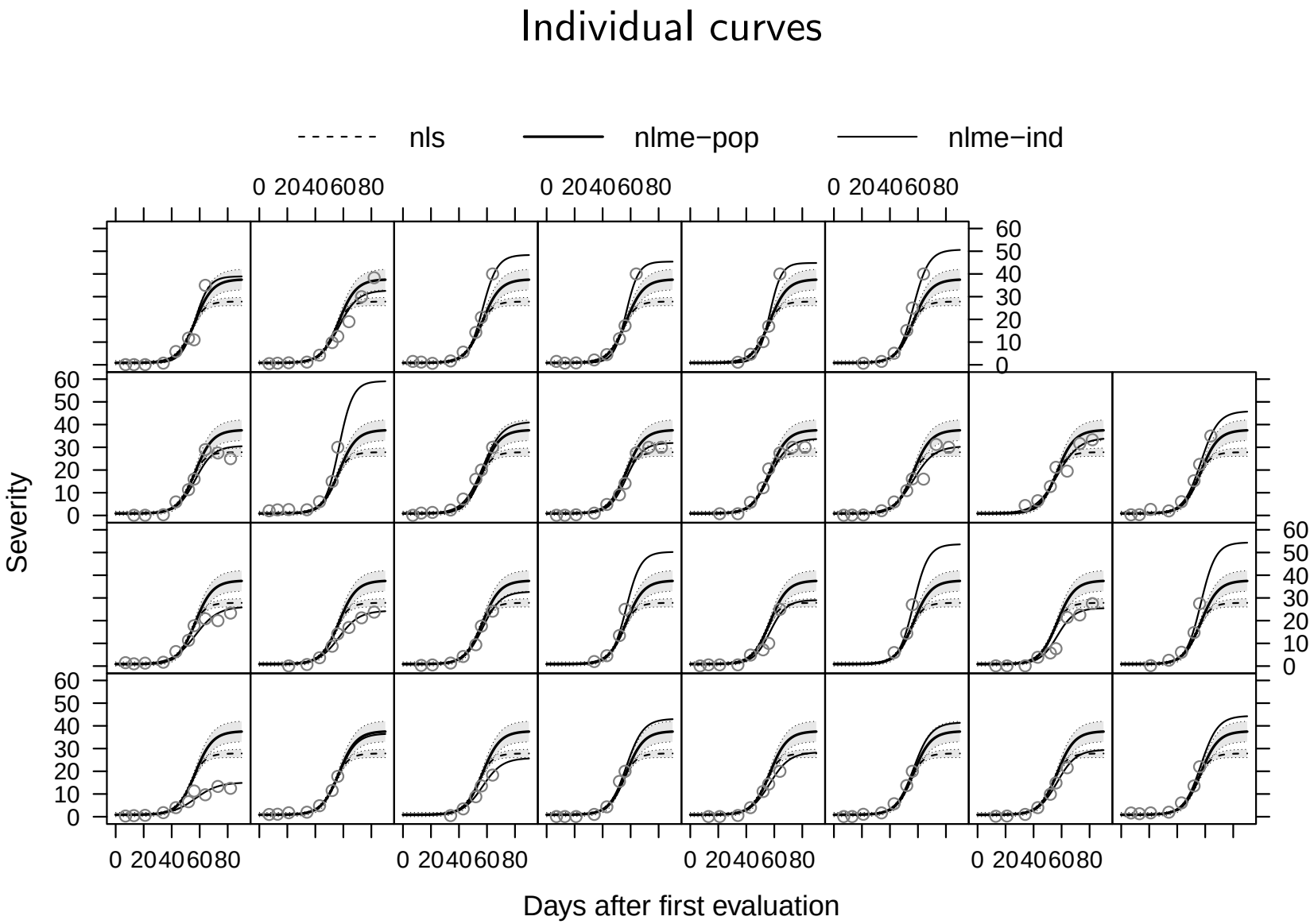


Results: *Observed vs. fitted values*

Curve averaging observations *vs.* curve averaging curves



Results: *Observed vs. fitted values*



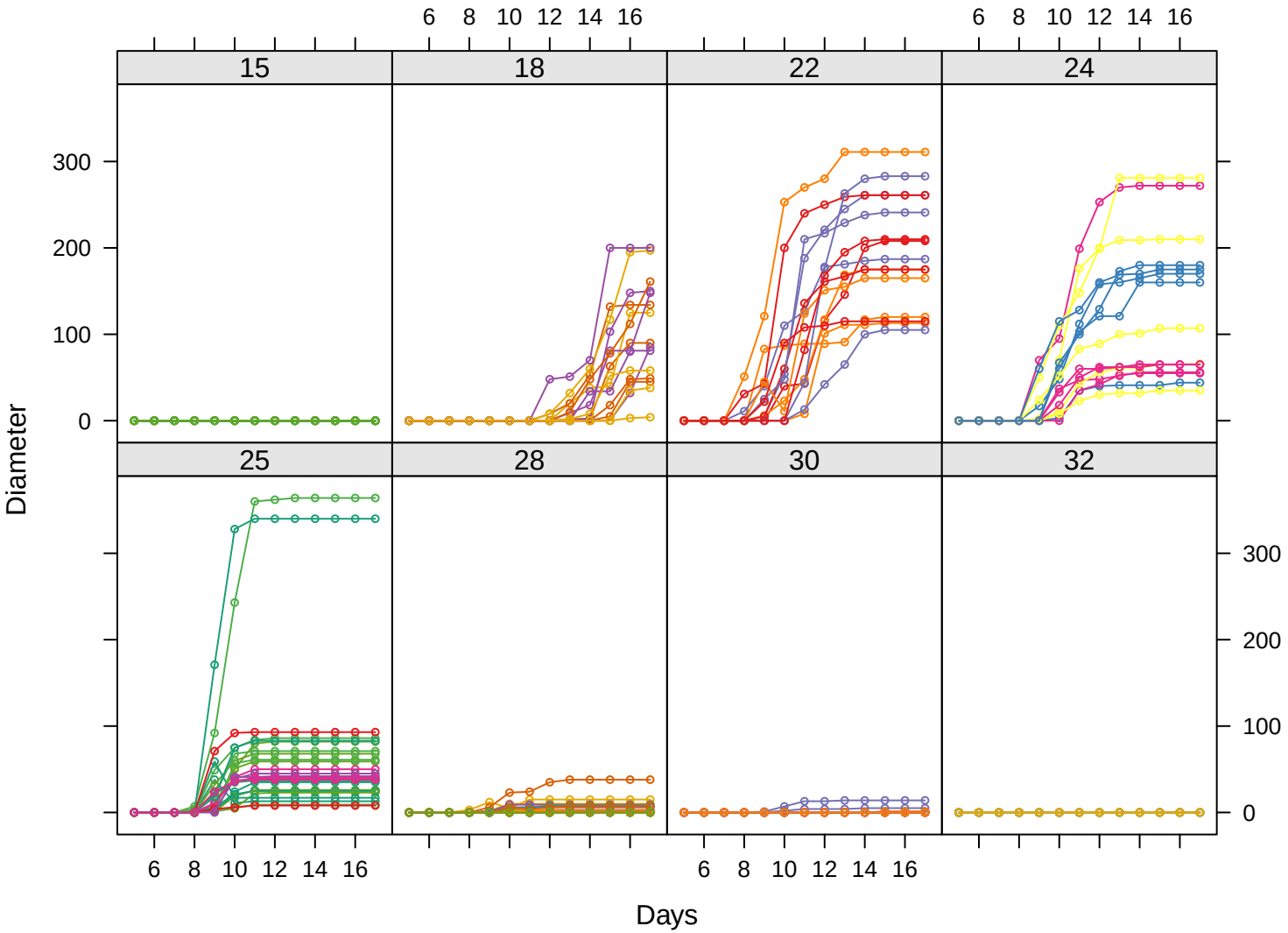
Conclusions

- The ONLR does not take into account the effect of experimental units or the fact that the leaves are not missing randomly;
- The NLRE accounts the effect of experimental units;
- By the principle of "averaging curves", the populational curve is not brought down due the lower severity of remaining leaves;
- In simple words, the ONLR is like to analyze as completely randomized design an experiment that was done in a randomized complete block;
- The impacts are not only the precision of the estimates (and curve) but also in the point estimates (bias).

Issue 2

Modelling variance in non linear regression models for size-temperature data

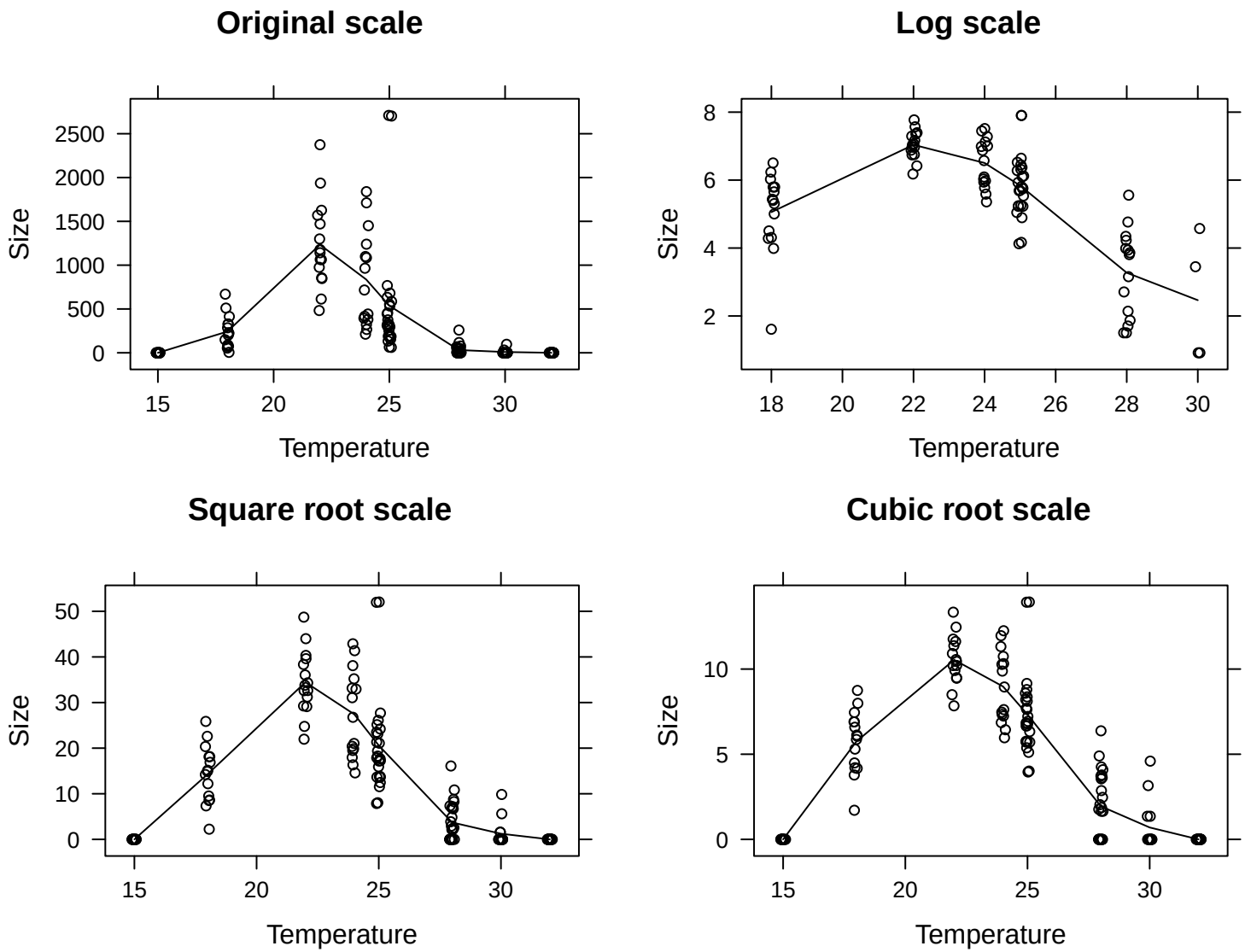
Data set



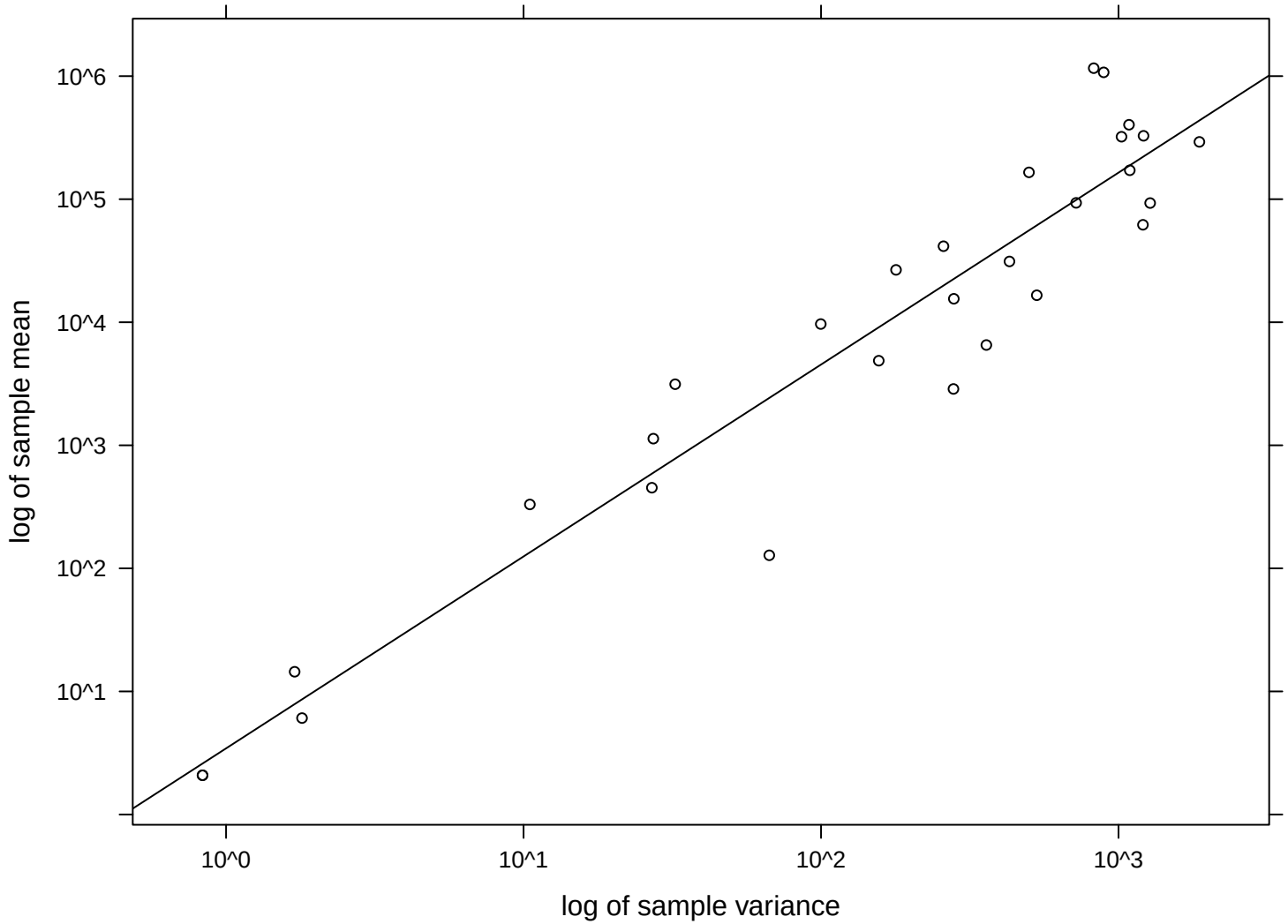
Data set

- Its a growth curve as a function of temperature;
- Area under the growth curve will be used;
- Objectives are optimum temperature and temperature range to positive development;

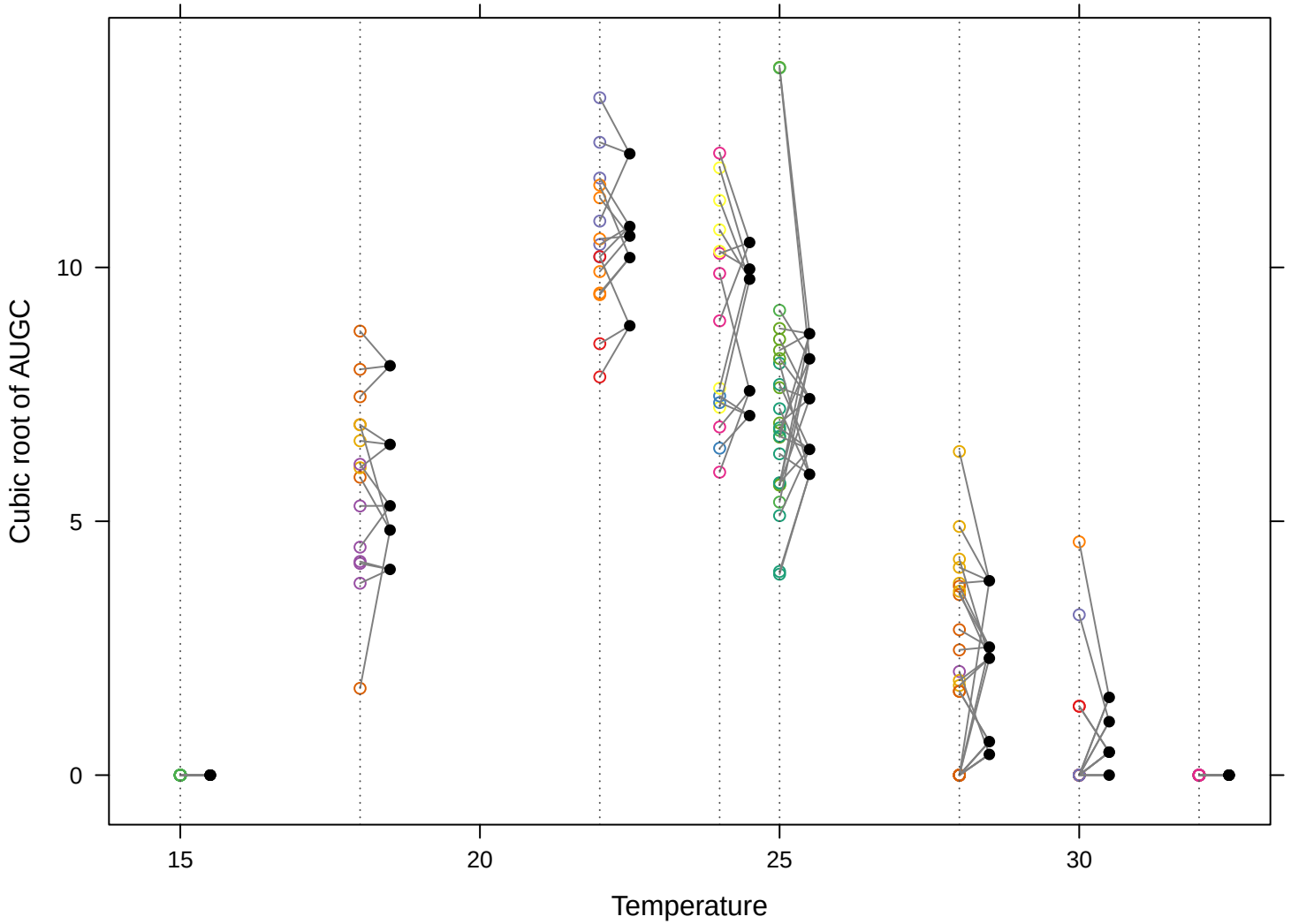
Exploratory analysis



Non null variance and mean relation



Also effect of replicates



Models

- Ordinary nonlinear regression;

$$\eta_\mu(x, \theta).$$

- Generalized nonlinear regression (with variance modelling);

$$\eta_\mu(x, \theta) \qquad \eta_{\sigma^2}(z, \varphi).$$

Likelihood functions

- Ordinary nonlinear regression

$$\ell(\theta, \sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \eta(x_i, \theta))^2; \tag{8}$$

- Nonlinear regression with variance modelling

$$\ell(\theta, \varphi) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{n}\left\{\log(\eta_{\sigma^2}(z, \varphi)) + \frac{(y_i - \eta_\mu(x_i, \theta))^2}{\eta_{\sigma^2}(z, \varphi)}\right\}; \tag{9}$$

- Random effects associated with θ_y to represent replicate effect.

Non linear models

- Generalized beta regression model is common in applications;
- Despite its five parameters, in general, two of them need be fixed by the user;
- A less restrictive and still interpretable model will was developed to be used

$$\eta_\mu(x,\theta) = \theta_y \exp\{\theta_q(x-\theta_x)^2 + \theta_c(x-\theta_x)^3\}, \tag{10}$$

  - θ_y is the size at optimum temperature;
  - θ_x is the optimum temperature;
  - θ_q is curvature around the optimum;
  - θ_c is related to the skewness around the optimum;
- Variance function

$$\eta_{\sigma^2}(z,\varphi) = \sigma_1^2|z|^{2\varphi};$$

- $z = \eta_\mu(x,\theta)$ was used.

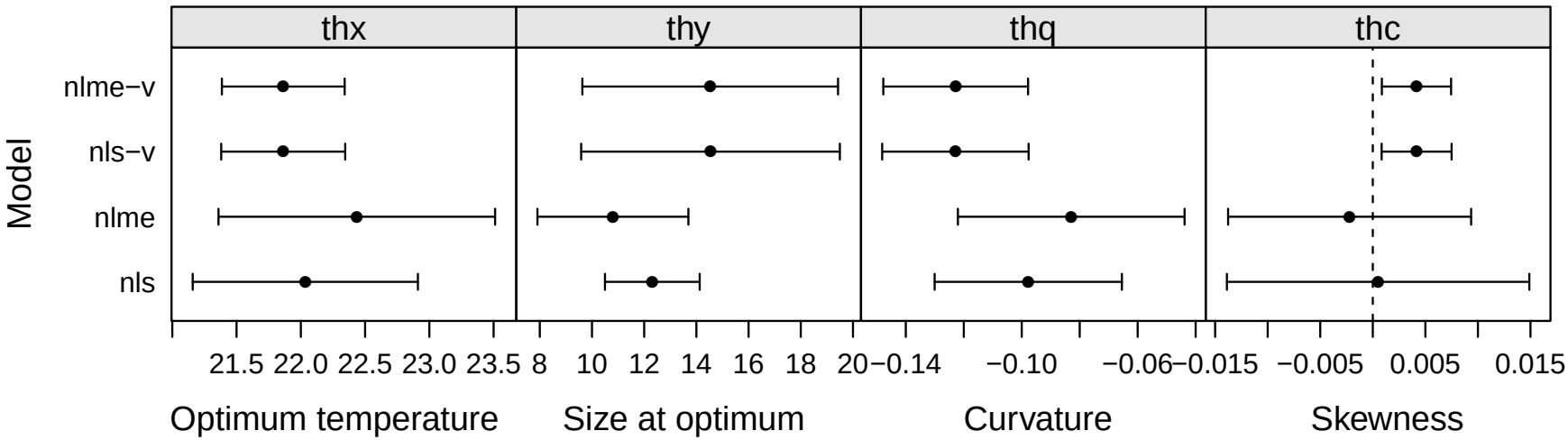
Fitted models

- NLS (ordinary);
- NLME (random effects);
- GNLS (variance modelling);
- GNLME (random effects and variance modelling).

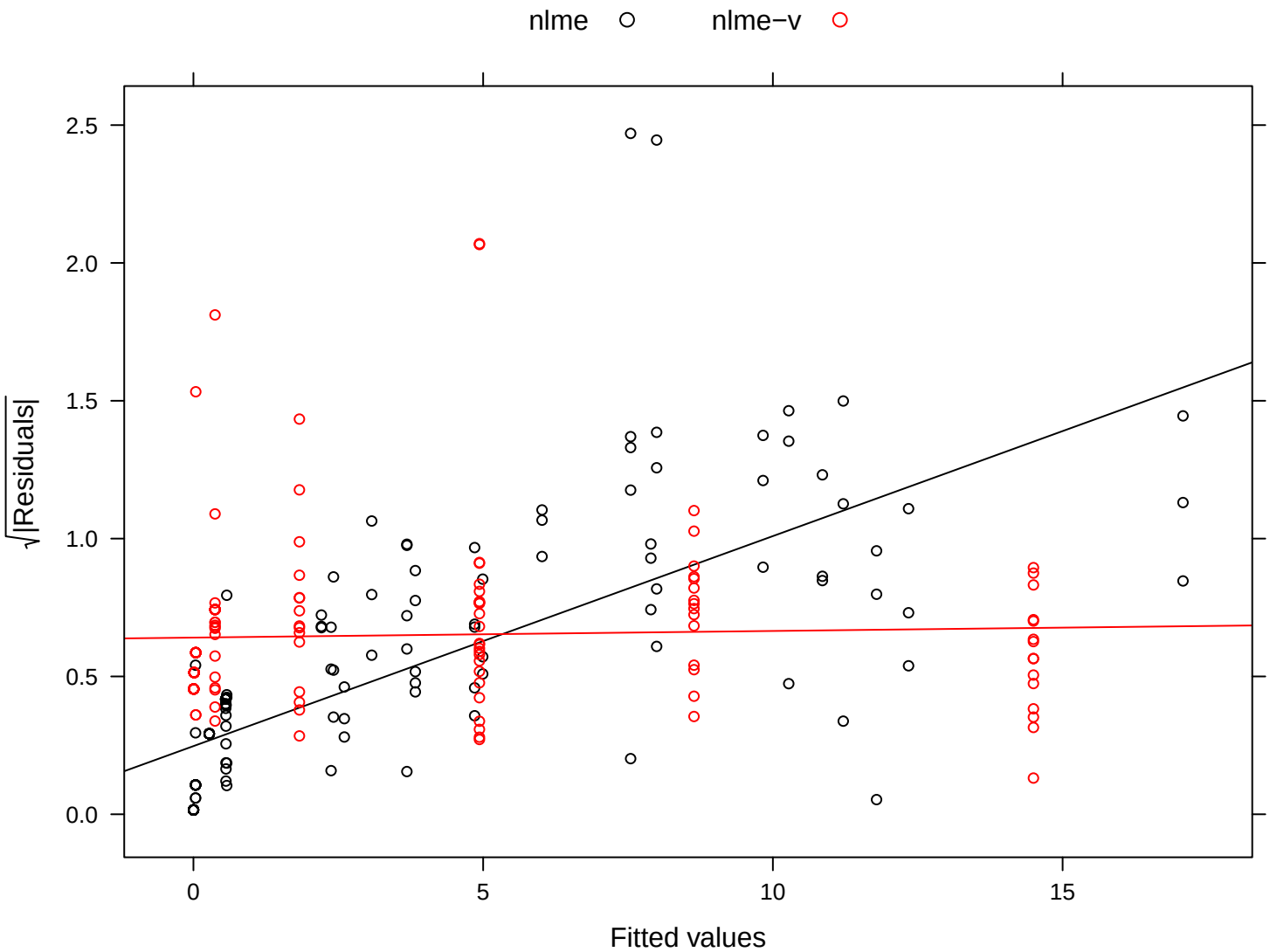
Results: *Standard output for the GNLS model*

```
Generalized nonlinear least squares fit
  Model: y ~ thy * exp(thq * (temp - thx)^2 + thc * (temp - thx)^3)
  Data: de
       AIC     BIC    logLik
  303.5402 321.604 -145.7701

Combination of variance functions:
 Structure: Power of variance covariate
 Formula: ~fitted(.)
 Parameter estimates:
    power
0.7877626

Coefficients:
        Value Std.Error  t-value p-value
thy 14.540902 2.5074295  5.79913  0.0000
thq -0.122891 0.0127825 -9.61399  0.0000
thx 21.863012 0.2440820 89.57240  0.0000
thc  0.004169 0.0016844  2.47532  0.0145

Residual standard error: 1.490477
Degrees of freedom: 150 total; 146 residual
```

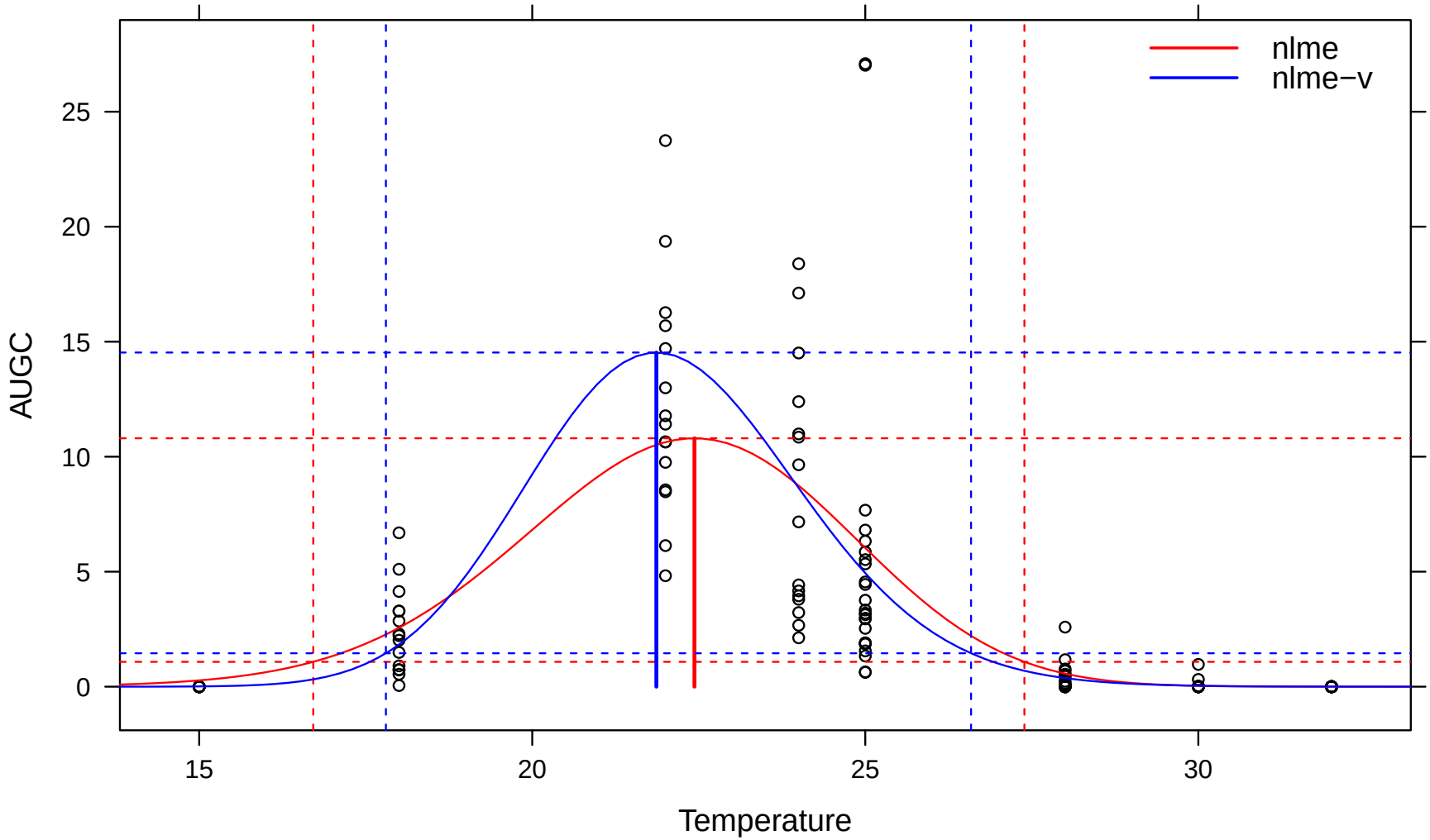
Results: *Point estimates and confidence intervals*



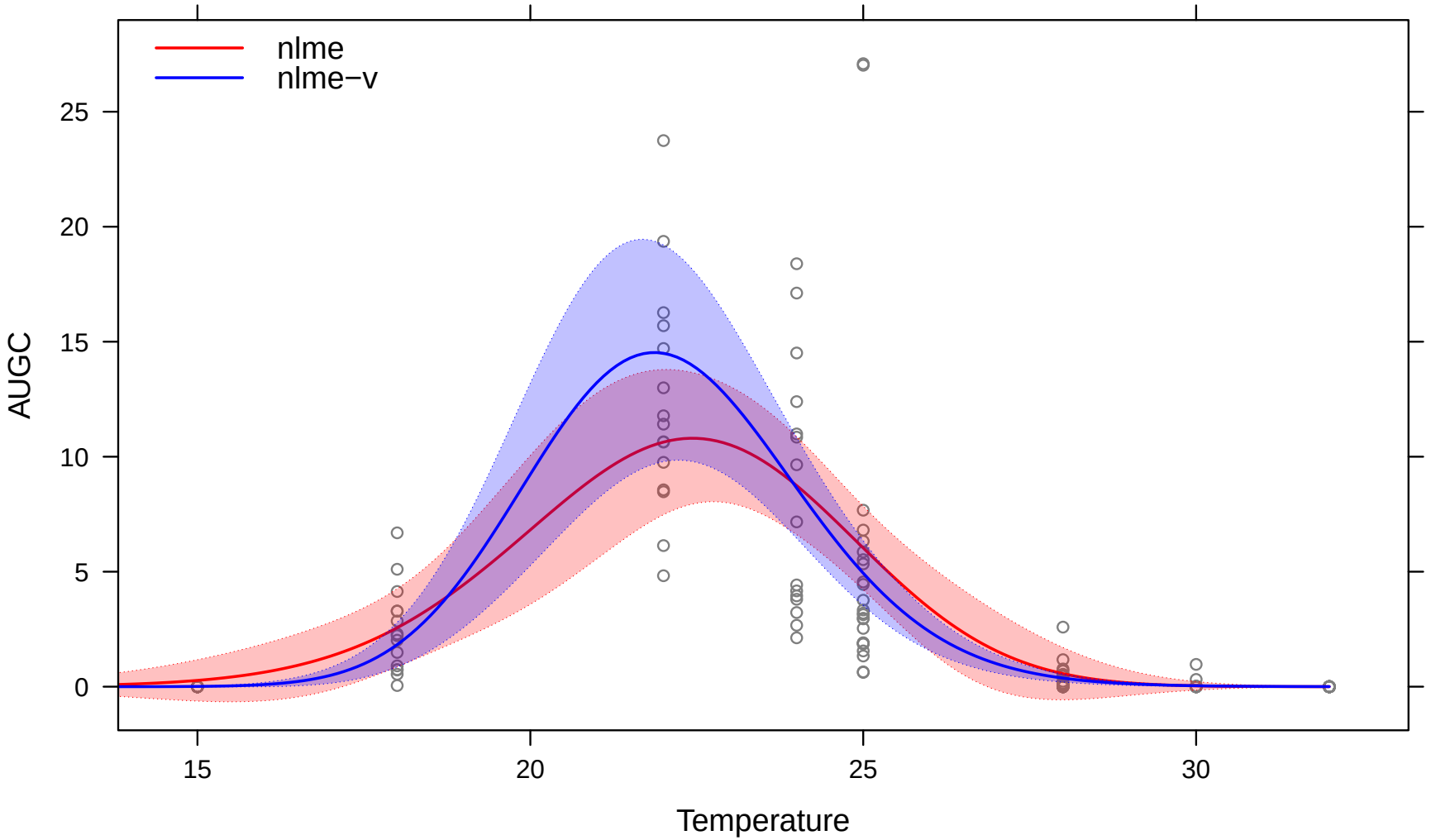
Results: *Residuals diagnostics*



Results: *Fitted values and limiting temperatures*



Results: *Fitted values with conficende bands*



Conclusions

- The NLS and NLME models do not indicate skewness;
- By modelling variance, skewness was found;
- Also, the null hypothesis of no random effects was not rejected;
- Using GNLS give more reliable confidence bands for the fitted curve;

General conclusions

- All analyzes were tailored to each data;
- In the past, to achieve this a large number of specialized software would be required;
- Today, R statical computing is free and open source software flexible and able to declare such models and conduct more appropriate analysis for the data better exploiting the information present there;
- Standard software is pretty available but its close flexibility often does not allow handle the current problems in plant epidemiology.
- Effort should be concentrated on disclosure of more contemporary methods of data analysis and software to their implementation.

References

MADDEN, L.; HUGHES, G.; BOSCH, F. van den. *The study of plant disease epidemics*. [S.l.]: American Phytopathological Society, 2007.

RATKOWSKY, D. A. *Nonlinear regression modeling*. New York: M. Dekker, 1983.