

# Ecological Statistics

## Introduction to non-linear models

### Non-linear module – part 2

C. Staudhammer – Eco Stats (fall 2008)

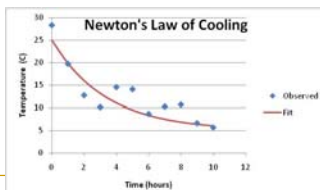
## Lecture outline

- Relevance of non-linear models in ecology and the applied sciences
- Relationship between non-linear and linear regression
- Estimation techniques for non-linear regression
- Fit statistics

C. Staudhammer – Eco Stats (fall 2008)

## Non-linear models in Applied Science

- Non-linear models give us the ability to represent biological processes as mathematical models
  - Models represent an understanding of the biological process
  - Often based on scientific theory and/or observation



C. Staudhammer – Eco Stats (fall 2008)

## What makes a model linear?

- In a linear model, the effects enter linearly
- Linear models include not only first-order models, but also more complex models, e.g.:
$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{2i}^2 + \beta_4 X_{1i} X_{2i} + \varepsilon_i$$
- Models with transformed variables that are *linear in the parameters* are also linear models, e.g.,
$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 \exp(X_{2i}) + \beta_4 X_{1i} + \varepsilon_i$$
- In general, a linear model is of the form:
$$Y_i = f(\mathbf{X}_i, \boldsymbol{\beta}) + \varepsilon_i$$
 where:  $f$  is an additive function,  $\mathbf{X}_i$  is a vector of observations on the predictor variables for the  $i$ th observation,  $\boldsymbol{\beta}$  is a vector of regression coefficients

C. Staudhammer – Eco Stats (fall 2008)

## What makes a model non-linear?

- In general, a non-linear model is of the form:
$$Y_i = f(\mathbf{X}_i, \boldsymbol{\gamma}) + \varepsilon_i$$
 where:  $\mathbf{X}_i$  is a vector of observations on the predictor variables for the  $i$ th observation,  $\boldsymbol{\gamma}$  is a vector of regression coefficients, and  $f$  denotes a *non-linear* function.
- As in linear regression, errors are assumed to have expectation zero, constant variance, and be uncorrelated

C. Staudhammer – Eco Stats (fall 2008)

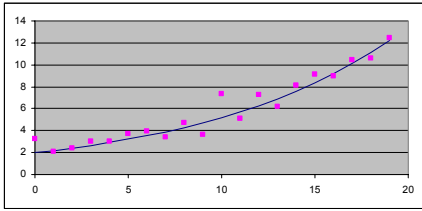
## Examples of non-linear models of growth

- Exponential regression model:
$$Y_i = \gamma_0 \exp(\gamma_1 X_i) + \varepsilon_i$$
  - rate of growth at time  $X$  is proportional to the amount remaining;  $\gamma_0$  is maximum growth when  $\gamma_1 < 0$
- von Bertalanffy growth function
$$L(t) = L_\infty (1 - \exp(k(t - t_0)))$$
  - $t_0$  is the time when  $L=0$ ;  $L_\infty$  represents the maximum size
- Chapman-Richards growth function
$$Y(t) = \beta_0 (1 - \exp(\beta_1 - \beta_2 t))^{\beta_3} + \varepsilon$$
  - $\beta_0$  is maximum growth,  $\beta_1$  governs the rate at which the maximum is approached

C. Staudhammer – Eco Stats (fall 2008)

## Why not linearize?

For example, what if you have weight-length data that looks like this?



We could fit the model:

$$y_i = \gamma_0 \gamma_1^{x_i}$$

which linearizes as:

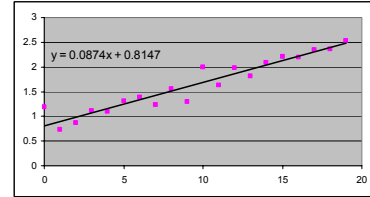
$$\ln(y_i) = \ln(\gamma_0) + \ln(\gamma_1)x_i$$

C. Staudhammer – Eco Stats (fall 2008)

## Options

(1) Transform the data (take the log of  $y$ ) and fit the model

- This "straightens" the curvilinear relation between weight and length, which allows for estimation of  $\gamma_0$  and  $\gamma_1$  via linear regression

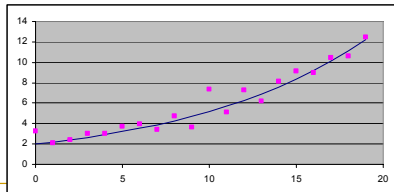


C. Staudhammer – Eco Stats (fall 2008)

## Options - continued

(2) Fit the model direction, retaining the logarithmic relationship

- Use Excel Solver (later), or
- Use SAS procedure PROC NLIN



C. Staudhammer – Eco Stats (fall 2008)

## Which option should you use?

- |                        |                             |
|------------------------|-----------------------------|
| ■ Transformed (linear) | Un-transformed (non-linear) |
| $g_0 = 2.25851$        | $g_0 = 2.23272$             |
| $g_1 = 1.09136$        | $g_1 = 1.09335$             |

Shouldn't these be the same?

In this case the methods produces estimates that are about the same... but why are they different?

$$y_i = \gamma_0 \gamma_1^{x_i} + \varepsilon_i \rightarrow \ln(y_i) = \ln(\gamma_0) + \ln(\gamma_1)x_i + ? \xi_i$$

Linearizing the model cannot linearize the error, and therefore errors can not be assumed  $\sim N(0, \sigma^2)$

C. Staudhammer – Eco Stats (fall 2008)

## Assumptions of nonlinear regression

- The model is correctly specified, e.g., you are not trying to fit a simple exponential function when the "real" relationship is Chapman-Richards
- The dependent variable is normally distributed
- The dependent variable is *homoscedastic*, i.e., the variability in  $y$  is approximately constant over all values of  $x$
- The values of the independent variable are known or measured without error
- The observations are independent

C. Staudhammer – Eco Stats (fall 2008)

## Estimation of non-linear regression parameters with Least Squares

- In Least Squares, we minimize the sum of the squared errors,  $Q$ :

$$Q = \sum_{i=1}^n (Y_i - f(\mathbf{X}_i, \boldsymbol{\gamma}))^2$$

- In this case,

$$f(\mathbf{X}_i, \boldsymbol{\gamma}) = \gamma_0 \gamma_1^{x_i}$$

- So, we minimize:

$$Q = \sum_{i=1}^n (Y_i - \gamma_0 \gamma_1^{x_i})^2$$

- To obtain the normal equations, take partial derivatives with respect to  $\gamma_0$  and  $\gamma_1$ , and set them equal to zero

Unlike linear regression, finding this solution requires iteration

C. Staudhammer – Eco Stats (fall 2008)

## Estimation of simple linear regression parameters with Maximum Likelihood

- In general, the probability density of an observation,  $Y_i$ , for the normal error simple linear regression model is:

$$f_i = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{Y_i - \beta_0 - \beta_1 X_i}{\sigma}\right)^2\right]$$

- Note that:  $E[Y_i] = \beta_0 + \beta_1 X_i$  and  $\sigma^2[Y_i] = \sigma^2$
- The likelihood function for  $n$  observations,  $Y_1, Y_2, \dots, Y_n$ , is the product of  $n$  densities:

$$L(\beta_0, \beta_1, \sigma) = \prod_{i=1}^n \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left[-\frac{1}{2}\left(\frac{Y_i - \beta_0 - \beta_1 X_i}{\sigma}\right)^2\right]$$

C. Staudhammer – Eco Stats (fall 2008)

## Estimation of non-linear regression parameters with Maximum Likelihood

- In a simple linear regression, the likelihood function reduces to:

$$L(\beta_0, \beta_1, \sigma) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_i)^2\right]$$

- It can be shown that the maximum likelihood estimates of  $\beta_0$  and  $\beta_1$  are the same as the least squares estimates
- For our non-linear example, the likelihood function is:

$$L(\gamma_0, \gamma_1, \sigma) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^n (Y_i - \gamma_0 \gamma_1^{X_i})^2\right]$$

- To obtain the maximum likelihood estimates, take partial derivatives with respect to  $\gamma_0$  and  $\gamma_1$ , and set them equal to zero

Estimates are the same as those obtained by least squares

C. Staudhammer – Eco Stats (fall 2008)

## Estimation of non-linear regr. parameters

- Regardless of the method, there is inevitably no closed-form solution to the minimization/maximization problem
- It is usually more practical to find solutions with direct numerical search procedures
  - Gauss-Newton method (linearization method) - Uses Taylor series expansion to approximate the model, the uses ordinary least squares to estimate parameters
  - Method of Steepest Descent - Searches for the minimum least squares criterion by iteratively determining the direction in which the regression coefficients should be changed
  - Marquardt algorithm - Uses the best features of the Gauss-Newton method and method of steepest descent
- Note: starting values must be given for all methods
- A global maximum/minimum is never assured

C. Staudhammer – Eco Stats (fall 2008)

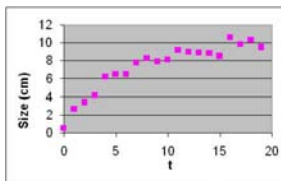
## Procedures for non-linear regression

- Using Excel
  - Solver can be used to find the solutions to maximization/minimization problems via the Generalized Reduced Gradient Algorithm
  - Only estimates are output; CI's can only be obtained via bootstrapping/Monte Carlo methods
- Using SAS procedure PROC NLIN
  - Uses one of 5 iterative methods
  - Outputs CI's and an array of fit statistics

C. Staudhammer – Eco Stats (fall 2008)

## Estimation of non-linear regression parameters with Excel Solver

Example: We have data on a the height of a certain ground cover species over time



t	Size (cm)
0	0.55
1	2.68
2	3.41
3	4.15
4	6.21
5	6.53
6	6.49
7	7.80
8	8.30
9	7.92
10	8.13
11	9.21
12	8.99
13	8.90
14	8.87
15	8.54
16	10.62
17	9.82
18	10.37
19	9.52

C. Staudhammer – Eco Stats (fall 2008)

## Estimation of non-linear regression parameters with Excel Solver - 2

- We think that size follows a von Bertalanffy model
- Recall the model form is:
 
$$L(t) = L_{\infty} * (1 - \exp(k*(t-t_0)))$$
- We approximate that the maximum height is 12 cm, and guess  $k=0.5$
- Then we compute the predicted values using this equation, as well as the residuals, and the sum of squared residuals

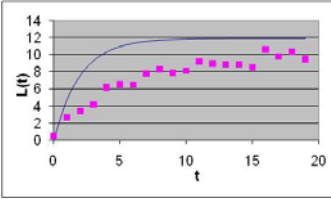
C. Staudhammer – Eco Stats (fall 2008)

## Estimation of non-linear regression parameters with Excel Solver - 3

$$L(t) = L_{\infty}(1 - \exp(-k(t-t_0)))$$

$L_{\infty}$	GUESS
$k$	0.5

Graphing the predicted vs. observed, we know this is not a great first estimate, but that is ok!

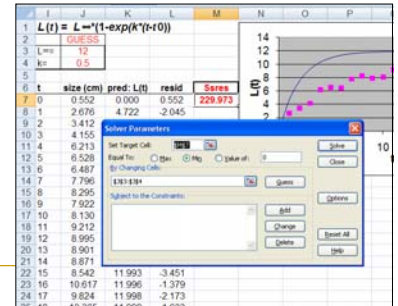


t	size (cm)	pred: L(t)	resid
0	0.552	0.000	0.552
1	2.676	4.722	-2.045
2	3.412	7.585	-4.173
3	4.155	9.322	-5.168
4	6.213	10.376	-4.163
5	6.528	11.015	-4.487
6	6.487	11.403	-4.915
7	7.796	11.638	-3.842
8	8.295	11.780	-3.485
9	7.922	11.867	-3.945
10	8.130	11.919	-3.789
11	9.212	11.951	-2.739
12	8.995	11.970	-2.975
13	8.901	11.982	-3.081
14	8.871	11.989	-3.118
15	8.542	11.993	-3.451
16	10.617	11.996	-1.379
17	9.824	11.998	-2.173
18	10.365	11.999	-1.633
19	9.921	11.999	-2.078

Seres  
229.9731

## Estimation of non-linear regression parameters with Excel Solver - 4

- Use Solver (Tools menu) to minimize SSres by changing the cells for  $L_{\infty}$  and  $k$

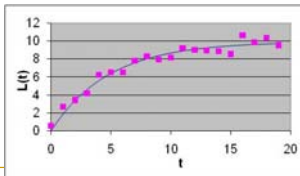


## Estimation of non-linear regression parameters with Excel Solver - 5

- The resulting parameter estimates minimize the residual sums of squares
- No CI's are given, nor fit statistics
  - These can be obtained via bootstrapping with PopTools

$$L(t) = L_{\infty}(1 - \exp(-k(t-t_0)))$$

$L_{\infty}$	GUESS
$k$	0.20763



t	size (cm)	pred: L(t)	resid
0	0.552	0.000	0.552
1	2.676	1.854	0.822
2	3.412	3.360	0.052
3	4.155	4.584	-0.429
4	6.213	5.579	0.634
5	6.528	6.387	0.141
6	6.487	7.043	-0.556
7	7.796	7.577	0.220

Seres  
5.528485

C. Staudhammer - Eco Stats (fall 2008)

## Estimation with SAS procedure PROC NLIN

- Produces least squares (or weighted least squares) estimates of the parameters of a nonlinear model
- Instead of simply listing regressor variables, you must:
  - write the regression expression
  - declare parameter names
  - supply initial parameter values
- Some models are difficult to fit
  - there is no guarantee that the procedure can fit the model successfully!

C. Staudhammer - Eco Stats (fall 2008)

## Estimation in PROC NLIN

- Estimation is an iterative process
  - NLIN procedure first examines the starting value specifications
  - Declared parameter values are used for the initial step of the iteration
- Iterative methods use derivatives or approximations to derivatives of the SSE with respect to the parameters to guide the search for the parameters producing the smallest SSE
  - Steepest-descent or gradient method
  - Newton method
  - Modified Gauss-Newton method
  - Marquardt method

C. Staudhammer - Eco Stats (fall 2008)

## PROC NLIN Syntax

```
PROC NLIN DATA=vonbert;
  PARMs L_0 = 9 k = 0.1;
  MODEL L_t = L_0 * (1 - exp(-k*t));
RUN;
```

This statement declares parameters to estimate and their starting values

C. Staudhammer - Eco Stats (fall 2008)

## PROC NLIN Output

```

The NLIN Procedure
Dependent Variable L_t
Method: Gauss-Newton

Iterative Phase
Iter   L_0      k      Sum of
      Squares
0      9.0000   0.1000   129.3
1      7.7928   0.2153   56.5141
2      9.8795   0.2062   5.5381
3      9.8892   0.2075   5.5277
4      9.8881   0.2076   5.5277
5      9.8880   0.2076   5.5277

Estimation Summary
Method      Gauss-Newton
Iterations      5
R            5.371E-6
PPC(k)       2.1E-6
RPC(k)       0.000031
Object       6.612E-9
Objective    5.527684
Observations Read      20
Observations Used      20
Observations Missing   0
    
```

NOTE: Convergence criterion met.

C. Staudhammer – Eco Stats (fall 2008)

## PROC NLIN Output - 2

```

The NLIN Procedure

Source      DF      Sum of Squares      Mean Square      F Value      Approx Pr > F
Model       2       1216.6              608.3           1980.75       <.0001
Error      18       5.5277              0.3071
Uncorrected Total  20      1222.1

Parameter      Estimate      Std Error      Approximate 95% Confidence Limits
L_0             9.8880       0.2816         9.2964      10.4797
k              0.2076       0.0191         0.1674      0.2478

Approximate Correlation Matrix
L_0      k
L_0      1.0000000      -0.8299280
k        -0.8299280      1.0000000
    
```

C. Staudhammer – Eco Stats (fall 2008)

## Convergence problems in NLIN

- Grid Search
  - Enables you to look for the best initial starting values
- Dependencies
  - Models with embedded dependencies will have a singular matrix of partial derivatives, resulting in non-identifiable estimates
- Non-convergence
  - Try a different METHOD= specification
- Divergence
  - Try including a BOUNDS statement for models where parameters are restricted (e.g., terms that involve SQRT, inverse, etc.)

C. Staudhammer – Eco Stats (fall 2008)

## Interpretation of parameters

- What is a parameter?
  - A parameter is “a quantitative property of a system that is assumed to remain constant over some defined time span of historical data and future prediction”
- Estimating parameters involves:
  - Finding values that provide the best fit between the model and the available data according to the criterion
  - Judging the goodness of fit of the data to the model and parameters
- Finding parameters that produce good fits *does not imply* that the model will make correct predictions or that there is *only one* combination of “good” parameters
  - Good fit can be obtained with incorrect parameters - model or data structure can be wrong

C. Staudhammer – Eco Stats (fall 2008)

## Is my nonlinear regression any good?

- Did the fitting algorithm converge?
- Is the curve close to the data (i.e., low MSE)?
- Is the equation biologically plausible?
  - Consider both the shape of the curve and the values of the estimated coefficients (i.e., intercepts)
- Is there any lack of fit?
  - Look for systematic deviations from the curve

C. Staudhammer – Eco Stats (fall 2008)

## Is my nonlinear regression any good? - 2

- Are the independent variables significant predictors of the dependent variable?
- Did you find the local or *global* minimum?
  - Try altering the starting values – do you get the same fit?
  - Are the parameters highly correlated?
    - When a model is mis-specified, or the estimation procedure gets “hung up” in a local minimum, the correlations between parameters may become very large.
    - This indicates that parameters are redundant, i.e., the effect of those two parameters on the function is very similar

C. Staudhammer – Eco Stats (fall 2008)

## What's ahead

- How do you choose a model?
- Evaluating and comparing the fit of multiple models
- Testing assumptions

## Tomorrow's lab

- SAS
  - PROC REG
  - PROC NLIN
- R
  - Requires additional LIBRARYs available from:  
<http://cran.r-project.org/> ← go to [Packages](#)
  - Before tomorrow, download two packages:  
Hmisc and Design