

## Ecological Statistics

Choosing, evaluating, and  
comparing non-linear models

Non-linear module – part 3

C. Staudhammer - Eco Stats (fall 2008)

## Lecture outline

- The variety of non-linear model forms
- Comparing the fit of two (or more) models
- Evaluating assumptions
- What to do when you cannot meet the assumptions

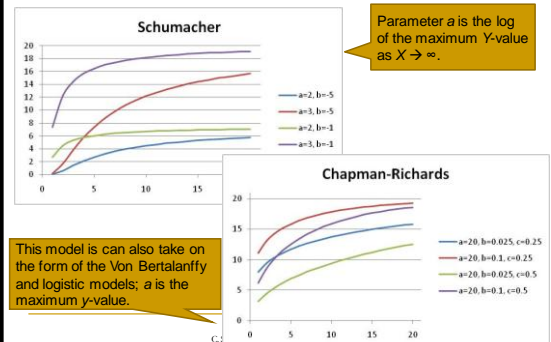
C. Staudhammer - Eco Stats (fall 2008)

## Choosing a non-linear model

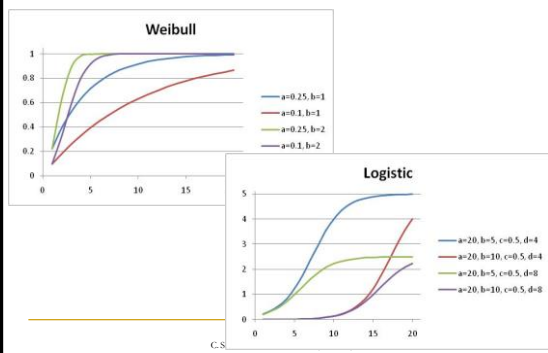
- Schumacher's Equation:  $Y = e^{a+b/X}$
  - Weibull:  $Y = 1 - e^{-aX^b}$
  - Chapman-Richard:  $Y = a \left( 1 - e^{-bX^c} \right)$
  - Logistic:  $Y = \frac{a}{d + e^{-b-cX}}$
- "Catalog of curves", e.g., see:  
<http://www.for.gov.bc.ca/hfd/pubs/docs/Bio/Bio04.pdf>

C. Staudhammer - Eco Stats (fall 2008)

## Examples of non-linear models



## Examples of non-linear models - 2



## Biologically-based choices

- Process models attempt to use theoretical biological relationships to describe observed data, e.g.:
  - If pipe model for water uptake by trees is correct, then water flow through trees should be strongly related to diameter-squared
  - VonBertalanffy equation is often used to describe fish size, as the equation is bounded from below and above

C. Staudhammer - Eco Stats (fall 2008)

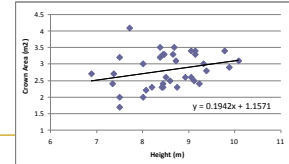
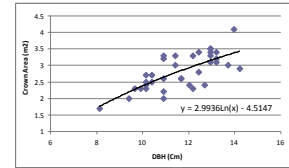
## Using graphs

- Univariate graphing
  - Works best when independent variables are independent of each other
- Example: we collect tree data, including diameter (dbh), height, crown length, etc. in order to predict crown area
  - Make univariate graphs of crown area versus each variable and look for patterns

C. Staudhammer - Eco Stats (fall 2008)

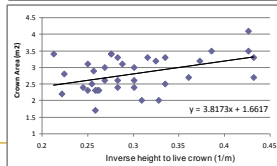
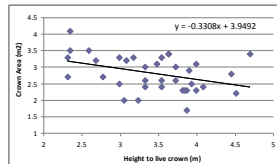
## Example graphs

- There may be a relationship between crown area and log(dbh)
  - $y = 2.9936\ln(x) - 4.5147$
- There may be a (weak) linear relationship between crown area and height
  - $y = 0.1942x + 1.1571$



## Example graphs - 2

- There may be a (weak) negative linear relationship between crown area and height to live crown
  - $y = -0.3308x + 3.9492$
- There may be a (stronger) relationship between crown area and the inverse of height to live crown
  - $y = 3.8173x + 1.6617$



C. Staudhammer - Eco Stats (fall 2008)

## Comparing the fit of two (or more) models

- When can you directly compare models?
  - Direct comparisons can be made when the same data were used to fit the models, *and* the fit statistics are expressed on the same y-unit basis (e.g., do not compare statistics computed for  $\ln(y)$  versus those of  $y$ )
- How do you compare models?
  - Logic
  - Fit statistics
  - Information criteria
  - Likelihood ratio test

C. Staudhammer - Eco Stats (fall 2008)

## Steps in comparing models

1. Examine the fitted curves of each model
  - Are they biologically reasonable?
  - Do the data systematically deviate from the model?
    - can be tested with the non-parametric *Runs Test*
  - Are the residuals  $\sim N$ ?
    - can be tested with Kolmogorov-Smirnov, etc...

C. Staudhammer - Eco Stats (fall 2008)

## Steps in comparing models - 2

2. Examine the CI's around the curves
  - Are they wide/narrow?

→ If the fitted curves do not make sense, or if the CI's are unreasonably wide, then reject that model - *you don't need statistical tests to reject a model for biological reasons*
3. If both models "make sense", compare goodness-of-fit statistics and use statistical tests

C. Staudhammer - Eco Stats (fall 2008)

### Comparing nested models with an F-test

- F-test can be used *only* if models are nested (one is a simplified version of another)
  - The *extra sum of squares test* can be used to compare any two nested models:
 
$$F_{(df_{full}-df_{alt}, df_{alt})} = \frac{(SS_{full} - SS_{alt}) / (df_{full} - df_{alt})}{SS_{alt} / df_{alt}}$$
  - When the alternate model has only one additional independent variable, this test simplifies to a Partial F
    - A partial F test answers the question: *Given the other variables are already in the model, does the additional variable explain a significant amount of variation in y?*

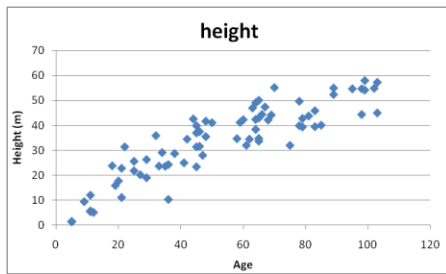
C. Staudhammer - Eco Stats (fall 2008)

### Comparing nested models with an F-test - example

- We want to describe the relationship between tree height and diameter.
- We have collected data on 75 trees.
- What do we do first?

C. Staudhammer - Eco Stats (fall 2008)

### Comparing nested models with an F-test – example (2)



C. Staudhammer - Eco Stats (fall 2008)

### Comparing nested models with an F-test – example (3)

- We decide that the Chapman-Richards equation describes the data well

$$Y = \text{maxht} (1 - e^{-bx})^c$$

- We fit the full model (with 3 parameters)
- We also fit an alternative (with 2 parameters, c=1)

C. Staudhammer - Eco Stats (fall 2008)

### Comparing nested models with an F-test – example (4)

- Results from fitting 3-parameter model

Source	DF	Sum of Squares
Model	3	100339
Error	72	2630.8
Uncorrected Total	75	102970

Parameter	Estimate	Approx Std Error	Approximate 95% Confidence Limits	
maxht	63.14	10.3385	42.5306	83.7494
b	0.0172	0.00796	0.0013	0.033
c	1.0493	0.2468	0.5574	1.5412

C. Staudhammer - Eco Stats (fall 2008)

### Comparing nested models with an F-test – example (5)

- Results from fitting 2-parameter model

Source	DF	Sum of Squares
Model	2	100337
Error	73	2632.3
Uncorrected Total	75	102970

Parameter	Estimate	Approx Std Error	Approximate 95% Confidence Limits	
maxht	65.183	6.0578	53.1096	77.2561
b	0.0156	0.00253	0.0106	0.0206

C. Staudhammer - Eco Stats (fall 2008)

### Comparing nested models with an F-test – example (6)

- Extra sums of squares test

Full model with three parameters

Source	DF	Sum of Squares	Mean Square	F Value	Approx Pr > F
Model	3	100339	33446.3	915.37	
Error	72	2630.8	36.5387		
Corrected Total	75	102970			

Alternative model with two parameters

Source	DF	Sum of Squares	Mean Square	F Value	Approx Pr > F
Model	2	100337	50168.7	1391.28	
Error	73	2632.3	36.0595		
Corrected Total	75	102970			

F-test

$$F(1, 2) = \frac{(100339 - 100337) / (3 - 2)}{100337 / 2} = 0.00004$$

Conclusion? Do not reject H<sub>0</sub>... But what is H<sub>0</sub>? H<sub>0</sub>: The full model is not significantly better than the alternative. Decision: do not reject H<sub>0</sub>; the full model is not significantly better than the alternative. What is another way I could have reached the same conclusion?

C. Staudhammer - Eco Stats (fall 2008)

### Comparing models with AIC

- AIC can be used even if models are not nested
  - Does not rely on hypothesis testing principles, but rather computes how likely the data were to arise from your model

$$AIC = -2\ell + 2(p+1)$$

where:  $\ell$  is the marginal log likelihood

- If  $Y|X$ -Normal:

$$AIC = n \ln \left( \frac{SS_{RES}}{n} \right) + 2(p+1) + n \ln(2\pi) + n$$

Residual variance is added to the # of parameters and # of observations

C. Staudhammer - Eco Stats (fall 2008)

### Comparing models with AIC - continued

- A single AIC computed from one model is not terribly meaningful; the change in AICs shows the change in fit vs. number of parameters, e.g.:

$$\Delta AIC = AIC_B - AIC_A = n \left[ \ln \left( \frac{SS_B}{n} \right) - \ln \left( \frac{SS_A}{n} \right) \right] + 2(p_B + 1) - (p_A + 1)$$

$$= n \ln \left( \frac{SS_B}{SS_A} \right) + 2(p_B - p_A)$$

A smaller AIC indicates a "better" model. Reduced SS<sub>RES</sub> is balanced against the change in # of parameters.

C. Staudhammer - Eco Stats (fall 2008)

### Comparing non-nested models with AIC – example

- We decide to fit two non-nested models
  - We fit the Chapman-Richards (with 3 parameters)
  - We also fit the Schumacher equation (2 param)

$$Y = e^{a+b/X}$$

C. Staudhammer - Eco Stats (fall 2008)

### Comparing non-nested models with AIC – example (2)

- Results from fitting the Chapman-Richards model

Source	DF	Sum of Squares
Model	3	100339
Error	72	2630.8
Uncorrected Total	75	102970

$$\rightarrow AIC = 2^*(p+1) + n (\ln(2\pi SS_{res}/n) + 1)$$

$$= 2^*4 + 75 (\ln(2\pi^* 2630.8 / 75) + 1)$$

$$= 487.7$$

C. Staudhammer - Eco Stats (fall 2008)

### Comparing non-nested models with AIC – example (3)

- Results from fitting the Schumacher model

Source	DF	Sum of Squares
Model	2	100134
Error	73	2835.2
Uncorrected Total	75	102970

$$\rightarrow AIC = 2^*(p+1) + n (\ln(2\pi SS_{res}/n) + 1)$$

$$= 2^*3 + 75 (\ln(2\pi^* 2835.2 / 75) + 1)$$

$$= 491.3$$

First model has lower AIC – the data were more likely to arise from the CR model

C. Staudhammer - Eco Stats (fall 2008)

## The Corrected AIC (AICC)

- When  $n$  is small relative to  $p$ , the AIC is biased too low
  - The corrected AIC adjusts for this difference

$$AICC = AIC + \frac{2(p+1)(p+2)}{n-p-2}$$

- When  $n > p + 30$ , this difference is trivial; nonetheless, some researchers recommend using AICC in all cases

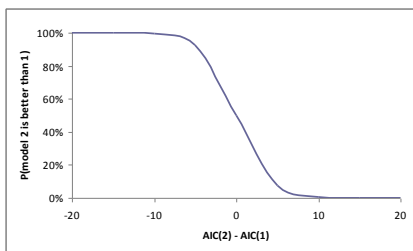
C. Staudhammer - Eco Stats (fall 2008)

## Comparing models with AIC and Relative Likelihood

- The probability that model B is better than model A is computed as:  $p = \frac{e^{-0.5\Delta AIC}}{1 + e^{-0.5\Delta AIC}}$  where:  $\Delta AIC = AIC_B - AIC_A$ 
  - These are relative probabilities, also called Akaike Weights
- The relative likelihood that model B is better than model A is computed as:  $\text{relative likelihood} = \frac{1}{e^{-0.5\Delta AIC}}$
- NOTE: because these concepts are not based on hypothesis testing, do not use the terms "significant" or "reject" in presenting their results*

C. Staudhammer - Eco Stats (fall 2008)

## Akaike weights



C. Staudhammer - Eco Stats (fall 2008)

## Comparing models with AIC – back to our example

- The probability that Chapman-Richards model is better than model Schumacher is computed as:

$$p = \frac{e^{-0.5\Delta AIC}}{1 + e^{-0.5\Delta AIC}}$$

$$\text{where: } \Delta AIC = AIC_{CR} - AIC_{SCH}$$

$$\Delta AIC = 487.7 - 491.3 = -3.6$$

$$p = \frac{e^{-0.5(-3.6)}}{1 + e^{-0.5(-3.6)}} = 0.859$$

Conclusion?

There is an 86% chance that the CR model is better than the Schumacher model

C. Staudhammer - Eco Stats (fall 2008)

## What method of comparison should I use?

- F-test
  - Can be used with nested models only
  - P-value tells you the probability that you would observe a difference between models as large as yours or larger if the experiment were repeated, assuming the null hypothesis is correct
- AIC
  - Can be used with non-nested models
  - Tells you the probability that a particular model is more likely than another, given your data

C. Staudhammer - Eco Stats (fall 2008)

## Evaluating assumptions

- We assume:
  - The independent variable(s) are fixed or measured without error
  - $Y$  at a particular  $X \sim N$
  - The data are homoscedastic
  - The observations are independent

C. Staudhammer - Eco Stats (fall 2008)

## Evaluating assumptions with graphs

- Scatter plots of  $Y$ 
  - Versus each independent variable
  - Provides support for model choice, can show heteroscedasticity
- Normal probability plots of  $Y|X$ 
  - Data points deviate substantially from a  $45^\circ$  line indicate a non-normal distribution

C. Staudhammer - Eco Stats (fall 2008)

## Evaluating assumptions with graphs -2

- Scatter plots of the residuals
  - Versus the independent variable(s) or versus the predicted  $Y$  values
  - A pattern indicates lack of fit; an increase in variability indicates heteroscedasticity
- Plots of standardized or studentized residuals
  - Large values indicated outliers and data points with substantial leverage

C. Staudhammer - Eco Stats (fall 2008)

## Violating assumptions

- Non-normal data can distort relationships and significance tests
  - This assumption refers to the dependent variable *only*
  - *Transformations can mitigate non-normality issues*
- Slight heteroscedasticity has little effect on significance tests; however, marked heteroscedasticity can be serious, increasing the possibility of a Type I error
  - *Transformations can mitigate homoscedasticity issues*

C. Staudhammer - Eco Stats (fall 2008)

## Statistical test of assumptions

- Normality of dependent variable
  - can be tested with Kolmogorov-Smirnov, Shapiro-Wilk, Anderson-Darling, or Cramer-von Mises tests in PROC UNIVARIATE (with NORMAL option)
- Homoscedasticity
  - Can be tested with model specification test (PROC REG, MODEL statement, SPEC option – White (1980))

C. Staudhammer - Eco Stats (fall 2008)

## What if you do not meet the assumptions?

- Be “honest”
  - Always describe how assumptions are violated when presenting results
- Investigate methods for non-linear data, or non-parametric analysis methods
  - More on this in future lectures
- **Note: there are some assumptions that cannot be mitigated, e.g., random selection of observations- No statistical method can overcome poor data collection methods**

C. Staudhammer - Eco Stats (fall 2008)

## Take home messages

- There are a wide variety of non-linear model forms
- Selecting an appropriate form should be based on:
  - Statistical considerations, such as goodness-of-fit tests, tests of assumptions, and comparison tests
  - Biological considerations and logic
- If you cannot meet the assumptions, other methods of analysis may be more appropriate

C. Staudhammer - Eco Stats (fall 2008)