Model Selection Part 2 & Introduction to Nonlinear models

1. Quick review of model selection
2. Using AIC to compare models
3. Building nonlinear models
4. Comparing competing nonlinear models

1. Quick review of model selection

What is the best model when selecting among many X variables?

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \beta_3 X_{3,i} + \beta_n X_{n,i} \ldots + \varepsilon_i$$

1. Quick review of model selection

What is the best model when selecting among many X variables?

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \beta_3 X_{3,i} + \beta_n X_{n,i} \ldots + \varepsilon_i$$

Cyanobacteria biomass = Nitrogen + Phosphorus
Cyanobacteria biomass = Iron + Phosphorus
Cyanobacteria biomass = Iron + Phosphorus + Temperature
Cyanobacteria biomass = Iron + Phosphorus + Temperature
+ Molybdenum + Wind strength + Turbidity
Cyanobacteria biomass = Temperature + Fish

Etc…

1. Quick review of model selection

Why not compare all possible models using P-values?

1. Quick review of model selection

Why not compare all possible models using P-values?

So… compare subset…

1. Quick review of model selection

Why not compare all possible models using P-values?

So… compare subset…

Using *a priori* hypotheses to compare thoughtfully selected models

2b. Test smaller subset of models, if possible…
Focused comparisons to test specific *a priori* hypotheses

	R^2	P
Iron	0.32	0.16
Phosphorus	0.14	0.12
Nitrogen + Phosphorus + Iron	0.57	0.02
Iron + Phosphorus	0.49	0.02

….

1. Quick review of model selection

Why not compare all possible models using P-values?

So… compare subset…

Using *a priori* hypotheses to compare thoughtfully selected models

Or

Stepwise regression allowing exploration of combinations of variables

Stepwise Regression

$$Y_i = \beta_0 + \beta_1 X_{1,\mathrm{i}} + \varepsilon_i$$
$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,l} + \varepsilon_i$$
$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \beta_3 X_{3,l} + \varepsilon_i$$
$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_3 X_{3,l} + \varepsilon_i$$

and so on….

-At each step, enter or remove a predictor based on whether that predictor improves model fit

Stepwise Regression – very strong cautions

-**No** guarantees that final model is
- -Best fit of all models
- -Scientifically meaningful

- Still influenced by multicollinearity (X-X correlations)
  - -Stepwise is better in this respect than "forward" and "backward" model selection methods that preceded it

- Still subject to Type I error problems (i.e., P = 0.05 means 1 in 20 models is likely to be wrong)

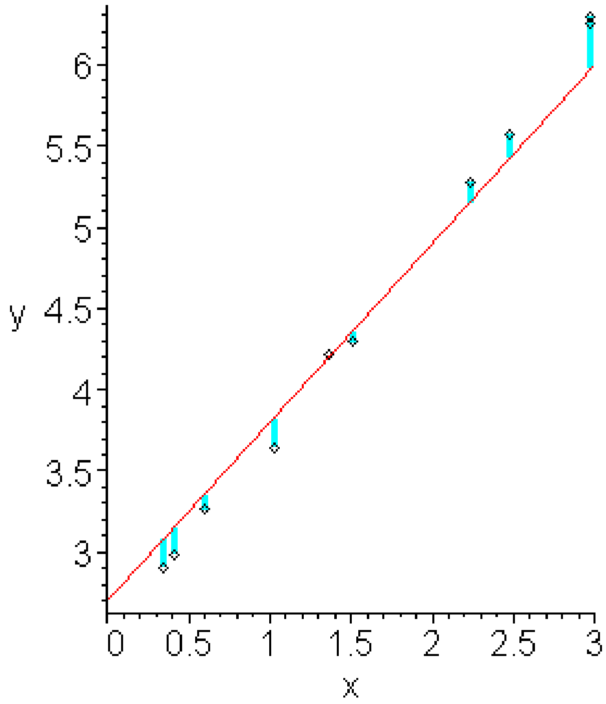
2. Using AIC to compare models

Getting away from the traditional probability tests

2. Using AIC to compare models

Getting away from the traditional probability tests

Just assess the fit of the model (e.g. the line or the plane) to the data,



$$L = -\frac{n}{2}\left[\log\left(2\pi\frac{SSE}{n}\right)+1\right]$$

Note that there are other formulations of *L*

L is log-likelihood

2. Using AIC to compare models

Getting away from the traditional probability tests

Just assess the fit of the model (e.g. the line or the plane) to the data,

$$\mathrm{AIC} = -2L + 2p$$

$$L = -\frac{n}{2}\left[\log\left(2\pi\frac{SSE}{n}\right) + 1\right]$$

and then penalize yourself for how complicated you make it

Note that there are other formulations of *L*

L is log-likelihood

p is number of parameters

2. Using AIC to compare models

Results in one number (AIC) that assesses model fit

$$\mathrm{AIC} = -2L + 2p$$

$$L = -\frac{n}{2}\left[\log\left(2\pi\frac{SSE}{n}\right) + 1\right]$$

L is log-likelihood

p is number of parameters

2. Using AIC to compare models

Results in one number (AIC) that assesses model fit

AIC is used to compare a suite of models

$$\mathrm{AIC} = -2L + 2p \qquad L = -\frac{n}{2}\left[\log\left(2\pi\frac{SSE}{n}\right)+1\right]$$

Models must all be based on the same data set, with the same dependent variable

L is log-likelihood

p is number of parameters

2. Using AIC to compare models

Results in one number (AIC) that assesses model fit

AIC is used to compare a suite of models

$$\mathrm{AIC} = -2L + 2p$$

$$L = -\frac{n}{2}\left[\log\left(2\pi\frac{SSE}{n}\right)+1\right]$$

Models must all be based on the <u>same data set, with the same dependent variable</u>

e.g., don't compare fish growth model between tuna populations at different places

L is log-likelihood

p is number of parameters

2. Using AIC to compare models

	AIC	R2	P
Iron	**242**	0.32	0.16
Phosphorus	**260**	0.14	0.12
Nitrogen + Phosphorus + Iron	**228**	0.57	0.02
Iron + Phosphorus	**226**	0.49	0.02

2. Using AIC to compare models

	AIC	R2	P
Iron	**242**	0.32	0.16
Phosphorus	**260**	0.14	0.12
Nitrogen + Phosphorus + Iron	**228**	0.57	0.02
Iron + Phosphorus	**226**	0.49	0.02

Rule of Thumb → Lowest AIC, with a difference of **2** being strong support for a better model

2. Using AIC to compare models

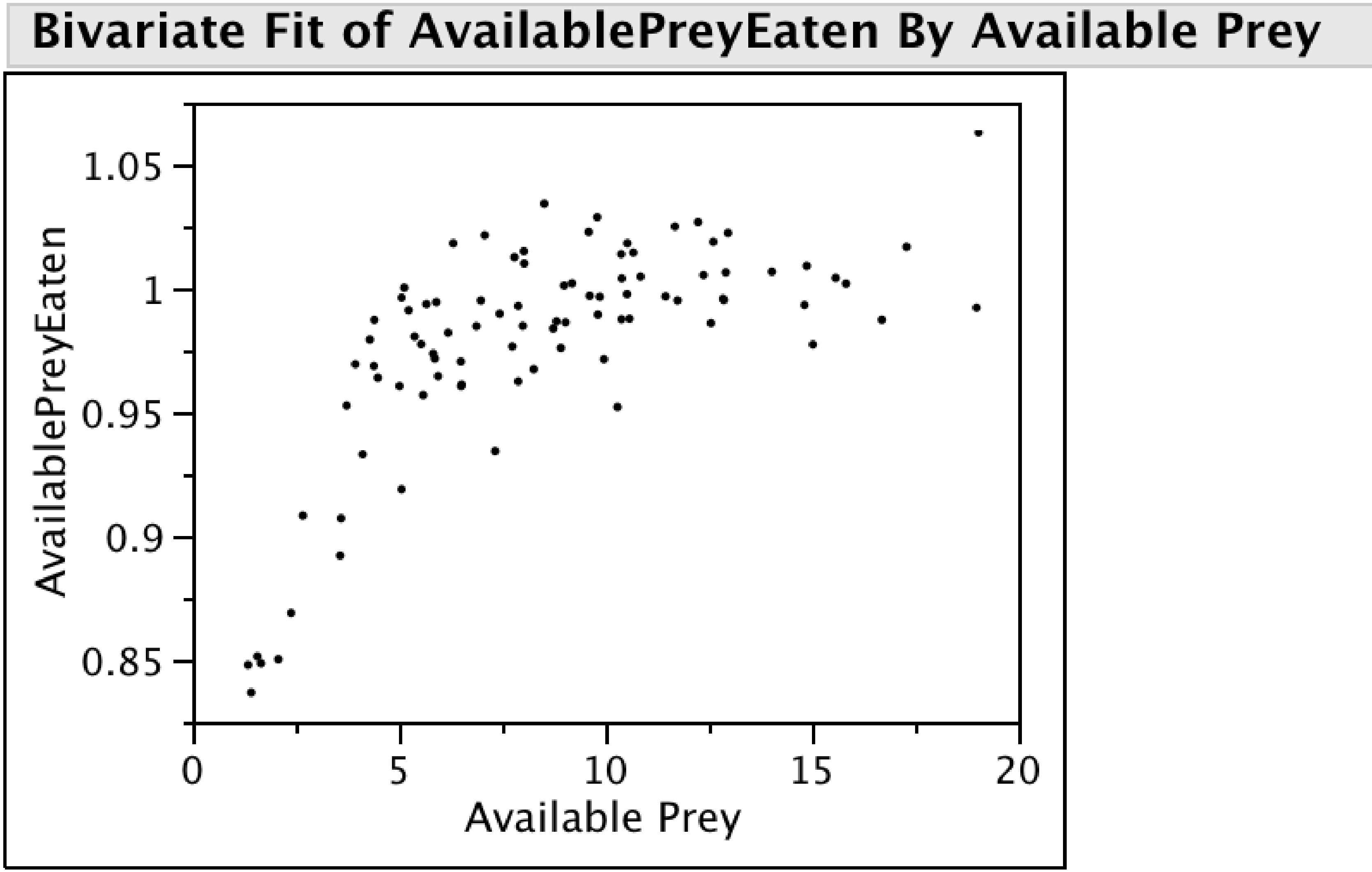
Sometimes software doesn't report AIC-calculate by hand

Burnham & Anderson. 1998. Model Selection and Multimodel Inference. Springer.

Mac Nally, R. 2000. Biodiversity and Conservation 9: 655–671
http://www.springerlink.com/content/w18t336xgm8q7773/

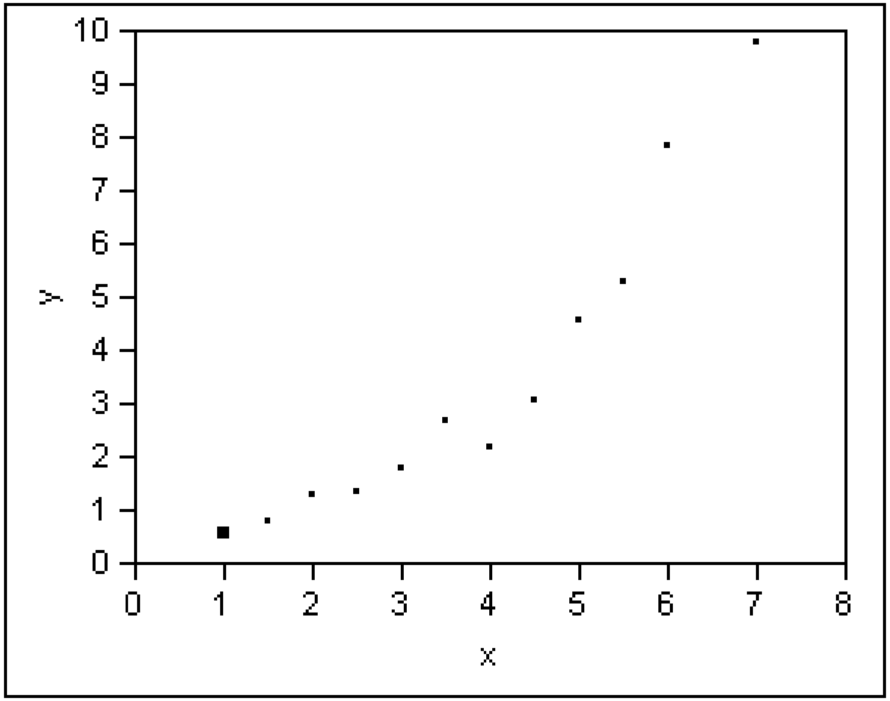
Hilborn, R., and Mangel, M. 1997. The Ecological Detective: Confronting Models with Data. Princeton University Press.

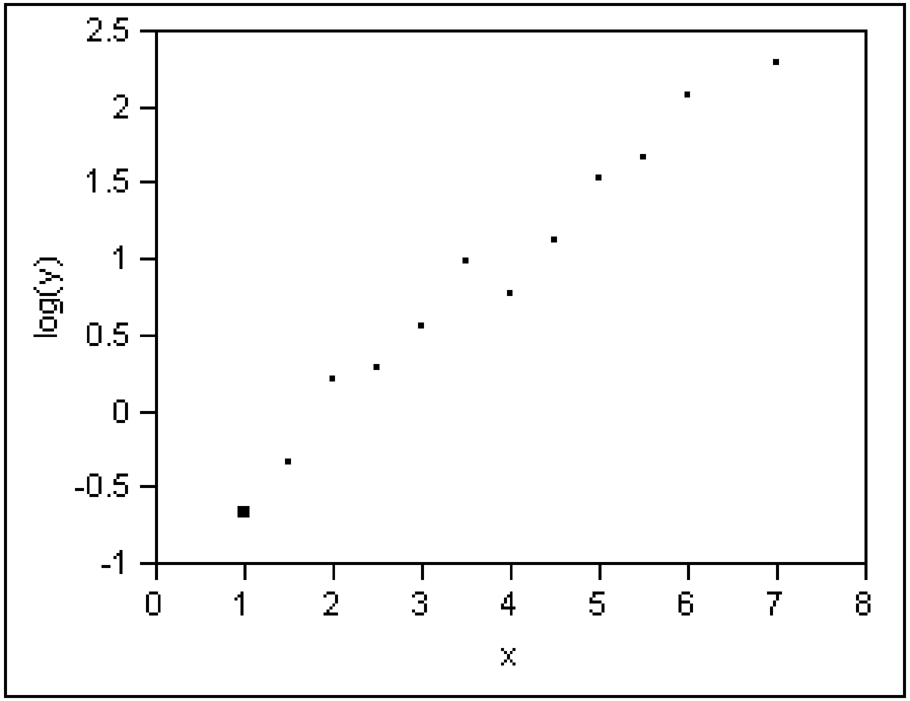
3. Building nonlinear models



3. Building nonlinear models

Some relationships considered "inherently linear" can be transformed easily to discern linear relationship (remember Log transforms)

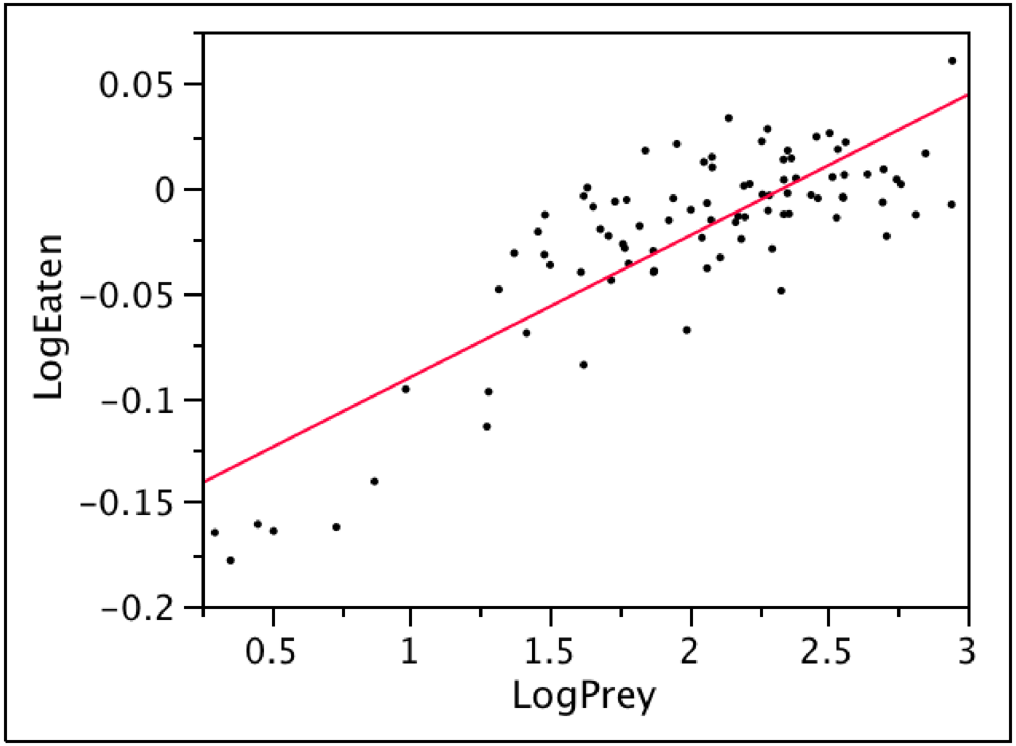




3. Building nonlinear models

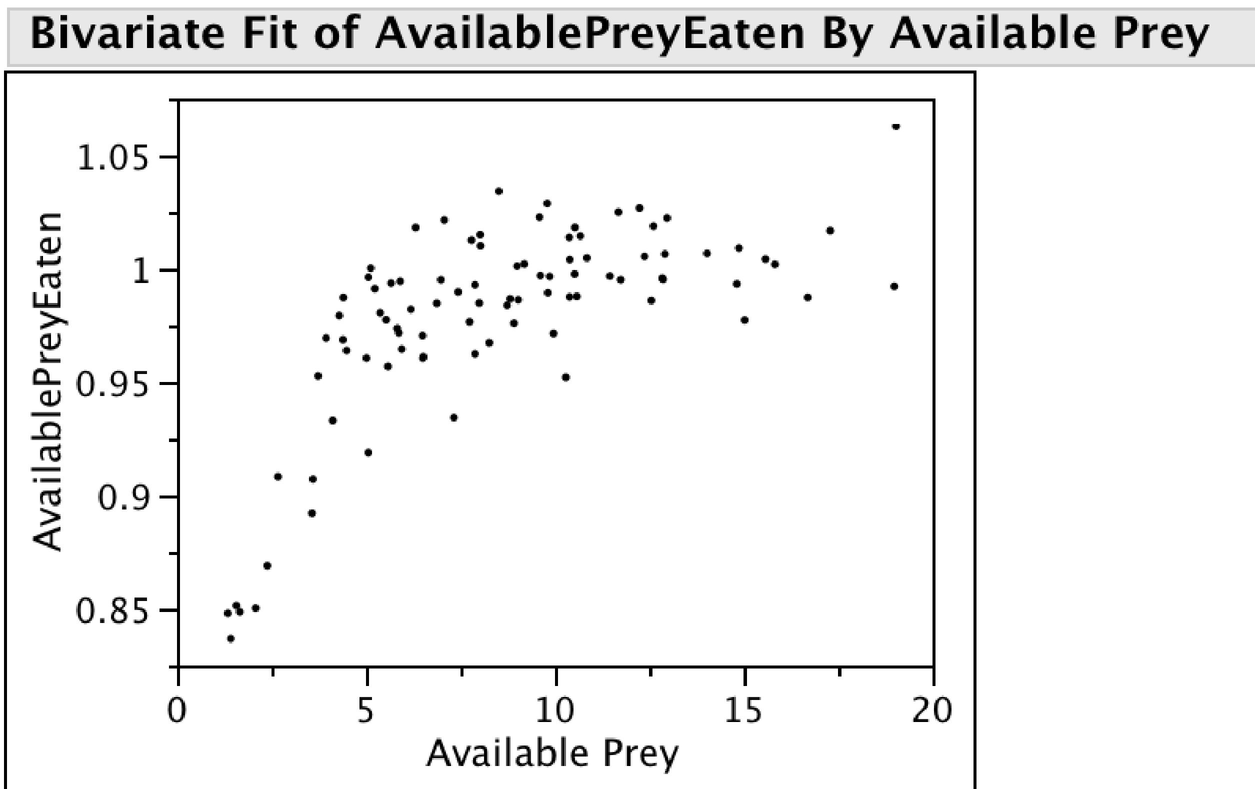
Some relationships considered “inherently linear” can be transformed easily to discern linear relationship (remember Log transforms)

But some relationships more complex functions



3. Building nonlinear models

Test hypothesis about mechanistic relationships by building and comparing nonlinear and linear models



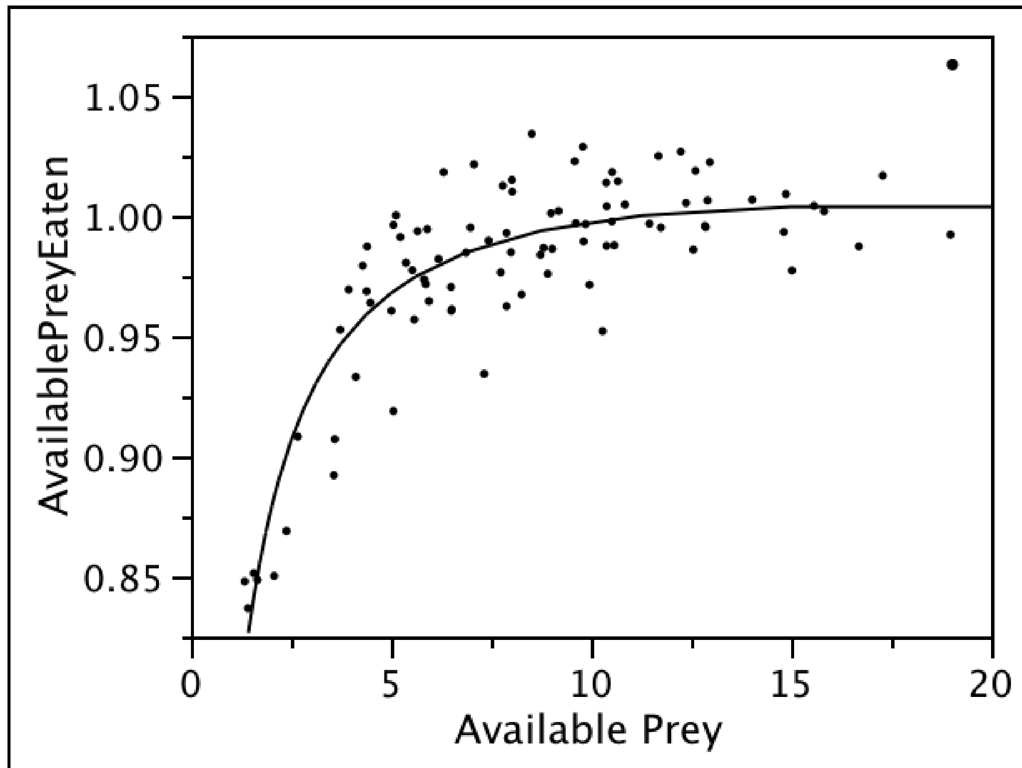
3. Building nonlinear models

Test hypothesis about mechanistic relationships by building and comparing nonlinear and linear models

Specify form of curve describing hypothesized relationship

(d * 2 * Available Prey + b * 2 * Available Prey ^ 2) /

(1 + c * Available Prey + d * Available Prey + (Available Prey^2) * b))

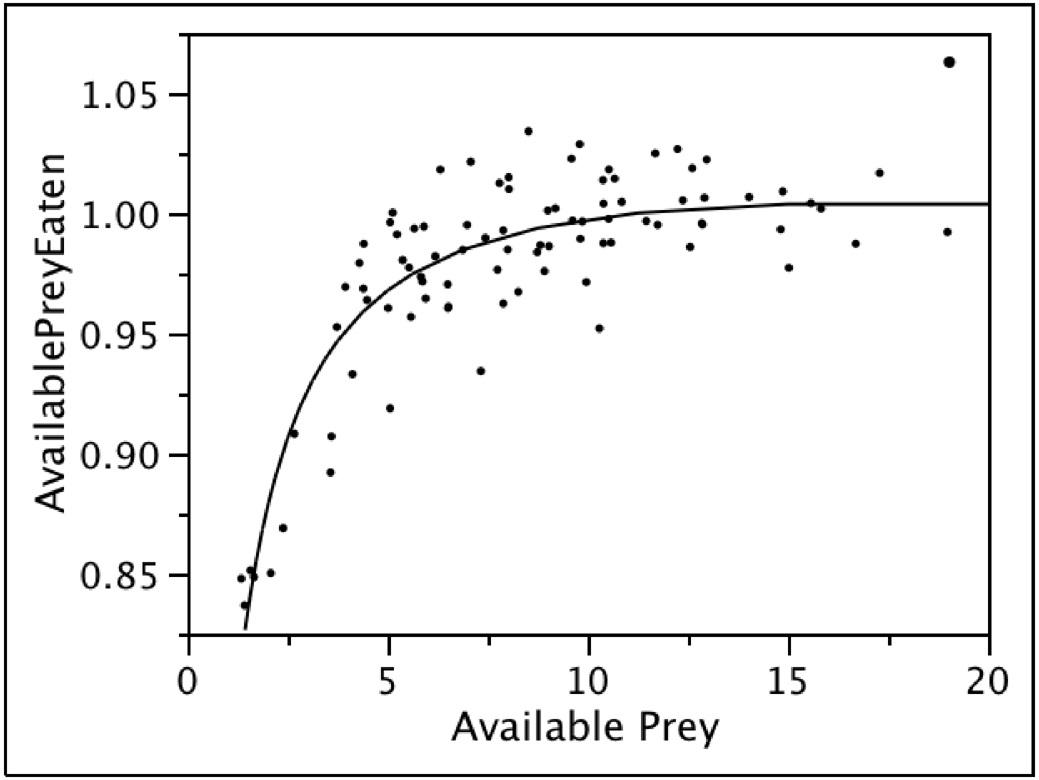


3. Building nonlinear models

Test hypothesis about mechanistic relationships by building and comparing nonlinear and linear models

Specify form of curve describing hypothesized relationship

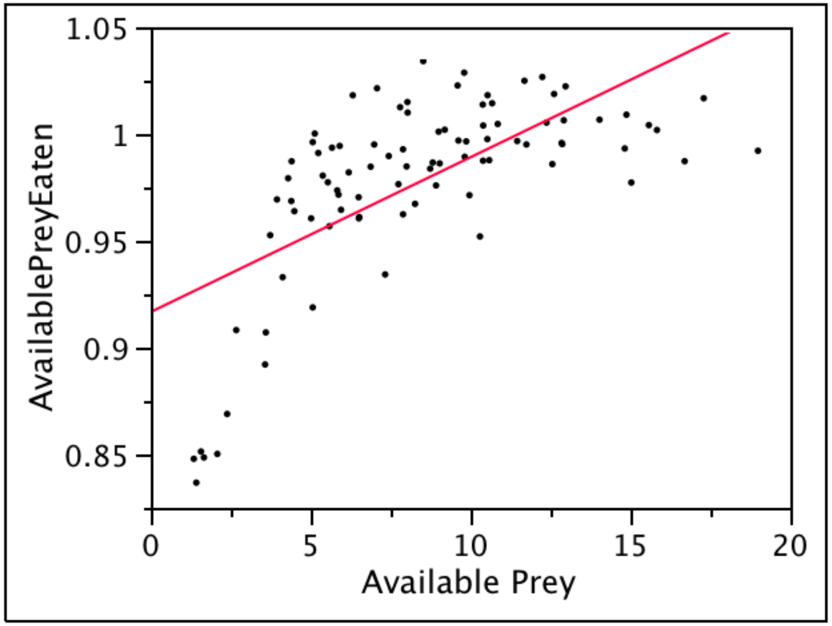
Algorithm optimizes parameter values

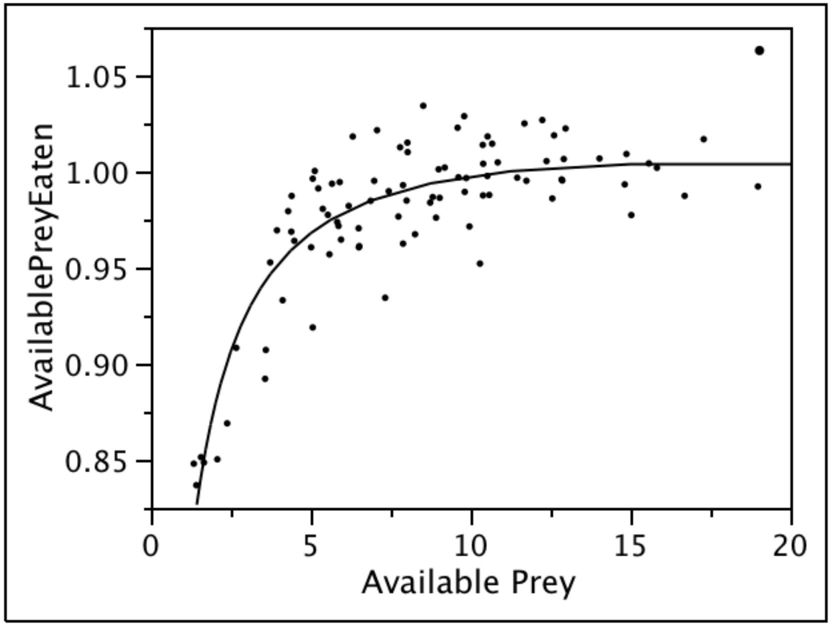


4. Comparing fit of competing models

Test hypothesis about mechanistic relationships by building and comparing nonlinear and linear models

Compare fit of competing models





4. Comparing fit of competing models

Test hypothesis about mechanistic relationships by building and comparing nonlinear and linear models

Compare fit of competing models

	SSE	L	p	AIC
Linear	90	-79.54	2	163.08
Nonlinear	40	-64.04	3	134.09

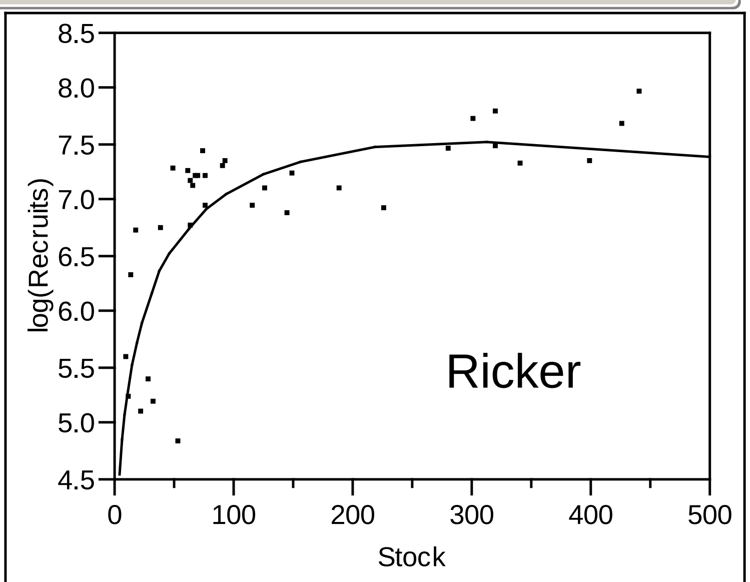
Note these numbers weren’t generated from data in previous graphs

4. Comparing fit of competing models: other examples

Test hypothesis about mechanistic relationships by building and comparing nonlinear and linear models

Nonlinear Fit

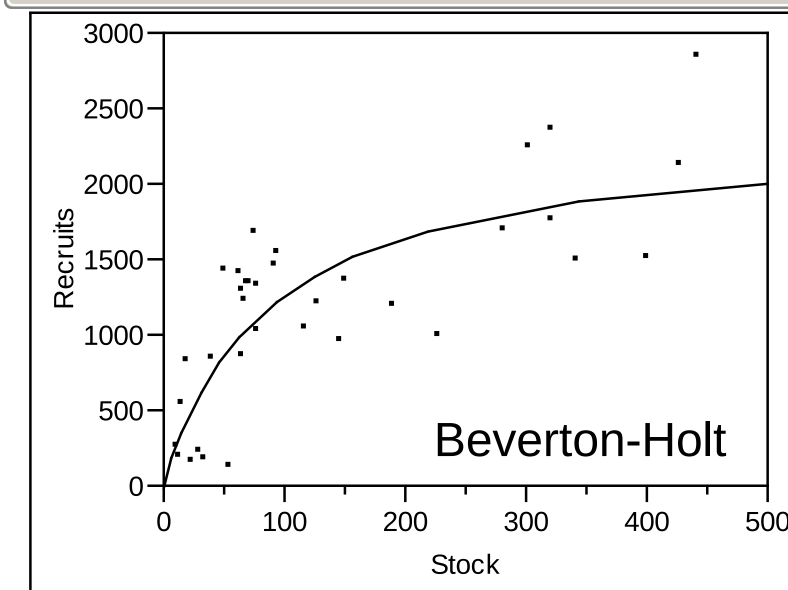
Plot



Parameter Estimate Low High

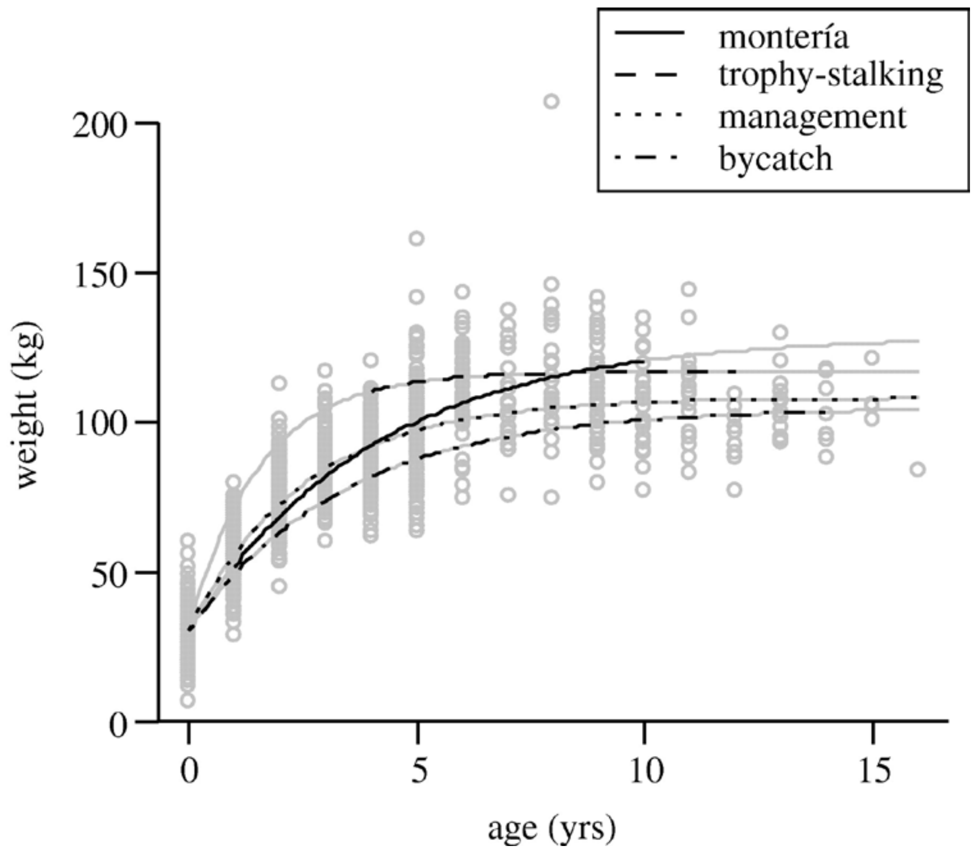
Nonlinear Fit

Plot



4. Comparing fit of competing models: other examples

Test hypothesis about mechanistic relationships by building and comparing nonlinear and linear models



Martinez et al. 2005. Different hunting strategies select for different weights in red deer. Biology Letters 3: 353-356.

Nonlinear models: Cautions

- Computationally intensive
- Need to provide pretty good guesses for initial parameter values – e.g., search literature before setting your guesses
- Algorithms do not always find best model

Nonlinear models: Cautions

- You can get similarly good fits with very different looking functions

- Different models will produce very different predictions when *extrapolating*