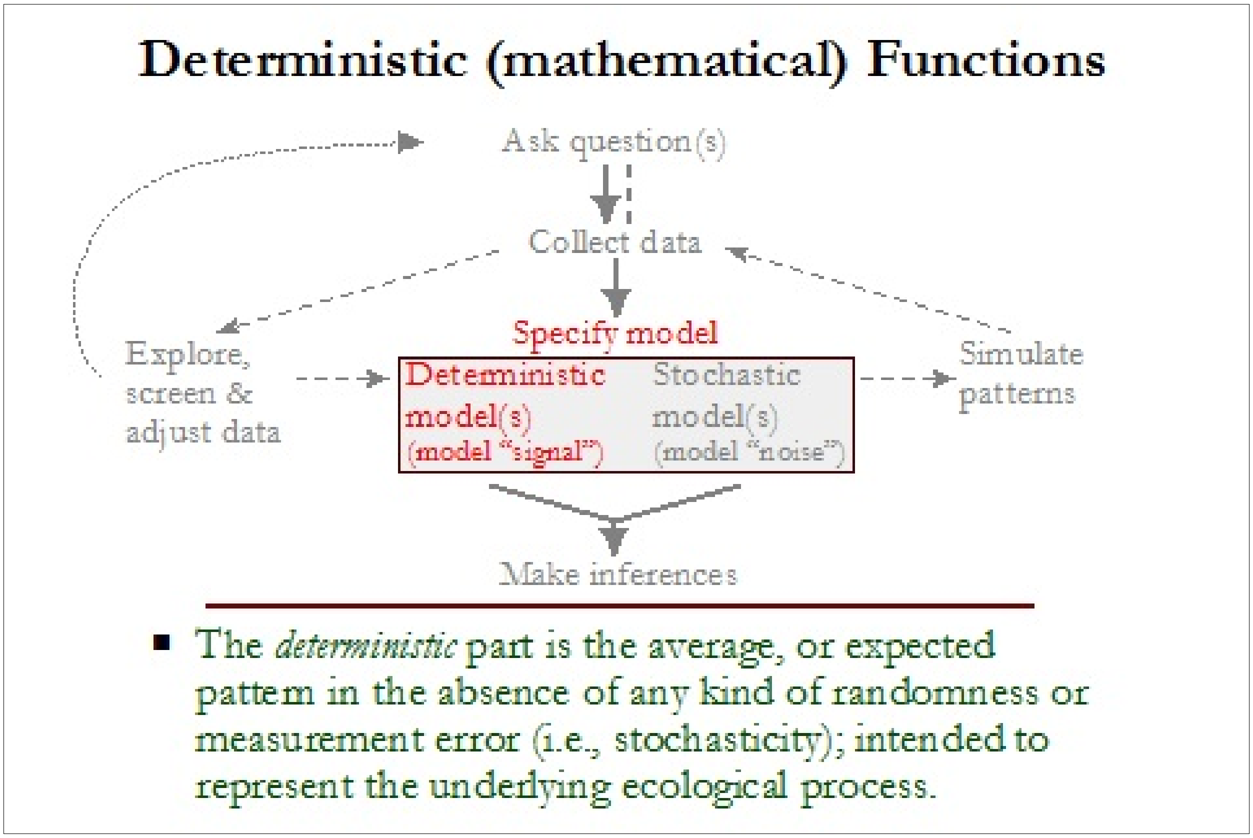
Analysis of Environmental Data

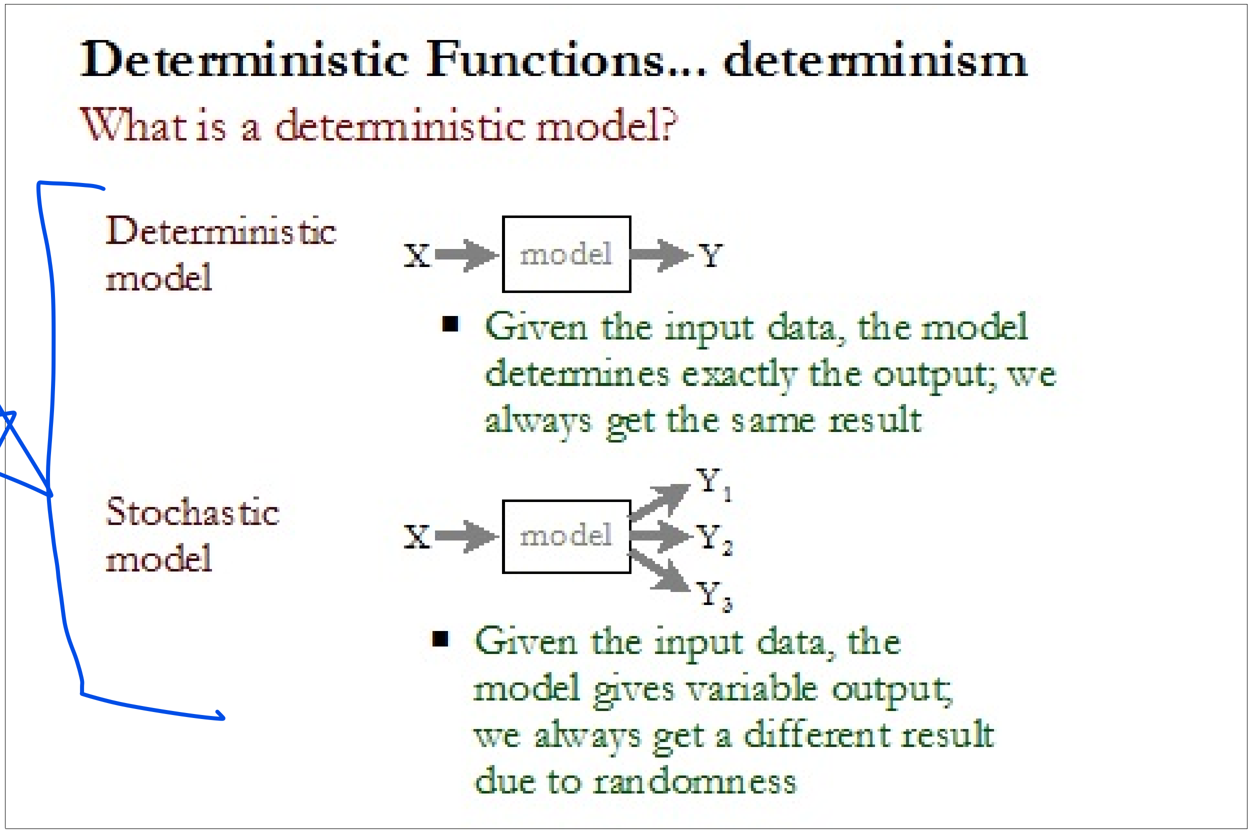
Chapter 4. Conceptual Foundations:

Deterministic Functions



1. What is a deterministic (mathematical) function

Most statistical models are comprised of a *deterministic* model(s) and a *stochastic* model(s). The deterministic part is the average, or expected pattern in the absence of any kind of randomness or measurement error (i.e., stochasticity). The deterministic model can be *phenomenological* (i.e., relationship based on the observed patterns in the data), *mechanistic* (i.e., relationship based on underlying ecological theory), or even a complex individual-based simulation model. More on these distinctions below. Importantly, the deterministic model is intended to represent the underlying ecological process, and estimating the parameters of this model is usually the focus of statistical modeling.



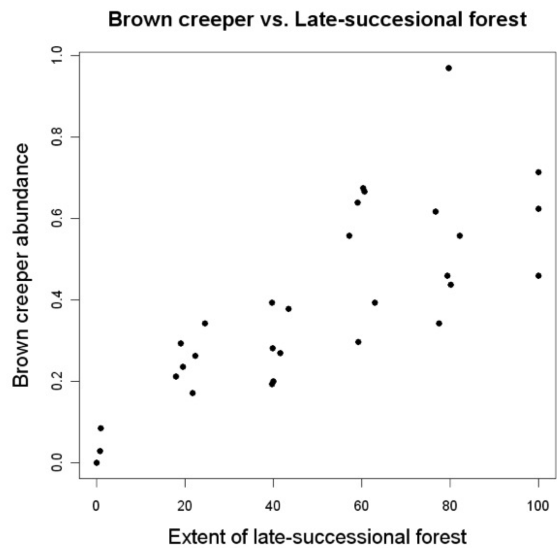
One way to distinguish the deterministic model from the stochastic model is as follows:

- *Deterministic model* – in a deterministic model, given the input data and parameter values, the model determines exactly the output, such that we always get the same result. If the deterministic model perfectly described the environmental system under consideration and there was no uncertainty or source of error, then given the value of the independent variable (x) and the model parameters, we would be able to predict the value of the dependent variable (y) exactly (i.e., with no uncertainty).

- *Stochastic model* – in a stochastic model, given the input data and parameter values, the model gives variable output, such that we always get a different result due to randomness. If there is some uncertainty in our model parameters, then we would expect for a given value of the independent variable (x) to generate a different value of the dependent variable (y) each time, since the model is imperfect. The stochastic model is simply the error in our ability to predict the outcome (dependent variable) for a particular input. All statistical models have a stochastic component.

Deterministic Functions... examples

Example: *Brown creeper abundance*



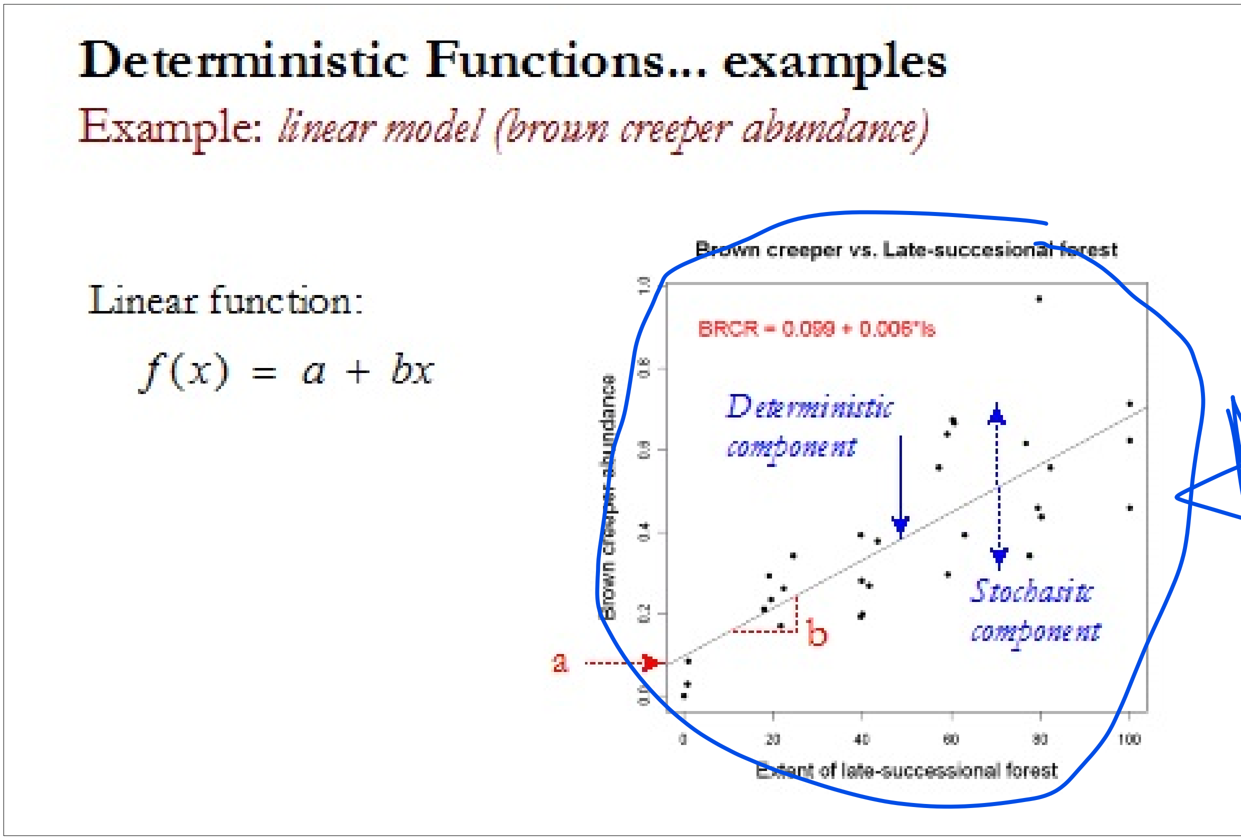
2. Examples of deterministic functions

2.1 Example 1 – Brown creeper abundance along forest succession gradient

In this example, the data represent the extent of late-successional forest and the relative abundance of brown creepers across 30 subbasins in the Oregon Coast Range.

- The first thing that you should notice about the plot is that brown creeper abundance appears to increase linearly with increasing percentage of the landscape comprised of late-successional forest. Thus, it would be logical to propose a simple *linear model* to represent the deterministic component of a statistical model for this data. Note that this is a phenomenological description of the relationship, as we derived it by observing the patterns in the data rather than hypothesizing it a priori (i.e., before looking at the data).

- The second thing that you should notice is that the relationship is not perfectly linear; i.e., there is considerable variability about the trend. This represents the *stochastic* component of the model. For now, we will ignore the stochastic component and focus solely on describing the deterministic component.

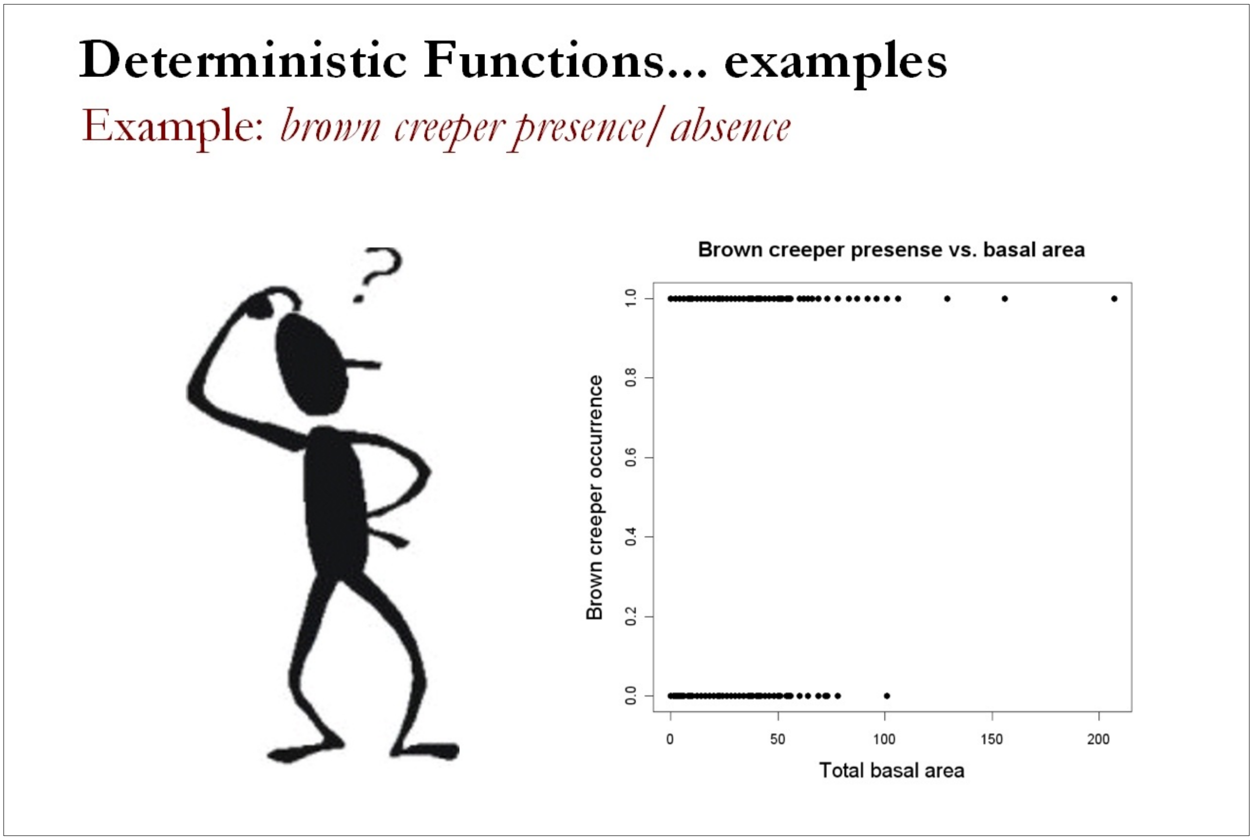


The linear function.– A simple linear function has the following form:

$$y = a + bx$$

where a is the y-intercept and b is the slope. Note, we can fit this model using a variety of methods (the usual method is Ordinary Least Squares, which finds the parameters that minimize the sum of the squared deviations from the fitted line), but for now, we will ignore the fitting procedure and focus solely on whether the linear model does a good job of describing the pattern.

Does the fitted linear model (solid line in the figure) do a good job in describing the relationship between brown creeper abundance and the extent of late-successional forest?

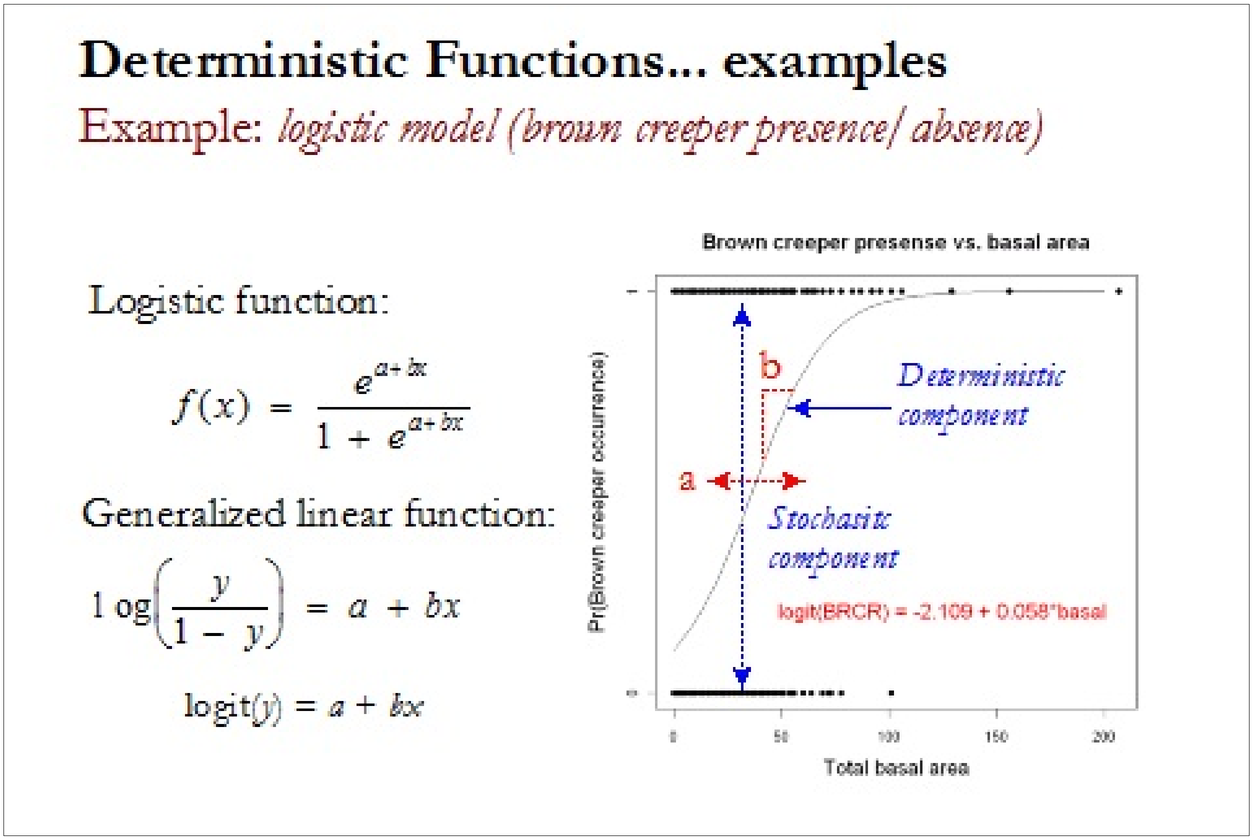


2.2 Example 2 – Brown creeper presence/absence along basal area gradient

In this example, the data represent the total basal area of trees and the presence/absence of brown creepers across 1,046 sample plots in the Oregon Coast Range.

- The first and most important thing that you should notice about the plot is that the response variable is binary; either 1 or 0, indicating the presence or absence, respectively, of brown creepers at the plot. Thus, the deterministic model must honor the 0-1 bounds of the response variable. In addition, you should notice that it appears that the proportion of presences increases as the total basal area increases.

- Given the binary response, it would be logical to use a simple *logistic model* to represent the deterministic component of the statistical model for this data. The logistic model has a characteristic sigmoid shape and is very commonly used for binary presence/absence data. Note, it is less clear in this particular case whether the logistic model is a phenomenological or mechanistic description of the relationship, since the parameters don't necessarily have a biological basis for their meaning.



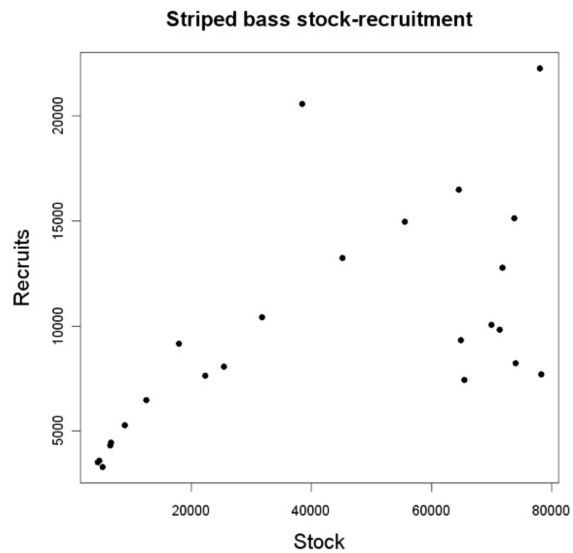
The logistic function.– A simple 2-parameter logistic function has the following form:

$$y = \frac{e^{a+b \cdot x}}{1 + e^{a+b \cdot x}}$$

where *a* is the "location" parameter that shifts the curve left or right, and *b* is the 'scale' parameter that controls the steepness of the curve. In this case, *y* represents the probability of 'presence'. Again, we can fit this model using a variety of methods (the usual method is known as Iteratively Reweighted Least Squares), but for now, we will ignore the fitting procedure and focus solely on whether the logistic model does a good job of describing the pattern. Does it?

Deterministic Functions... examples

Example: *striped bass stock-recruitment*

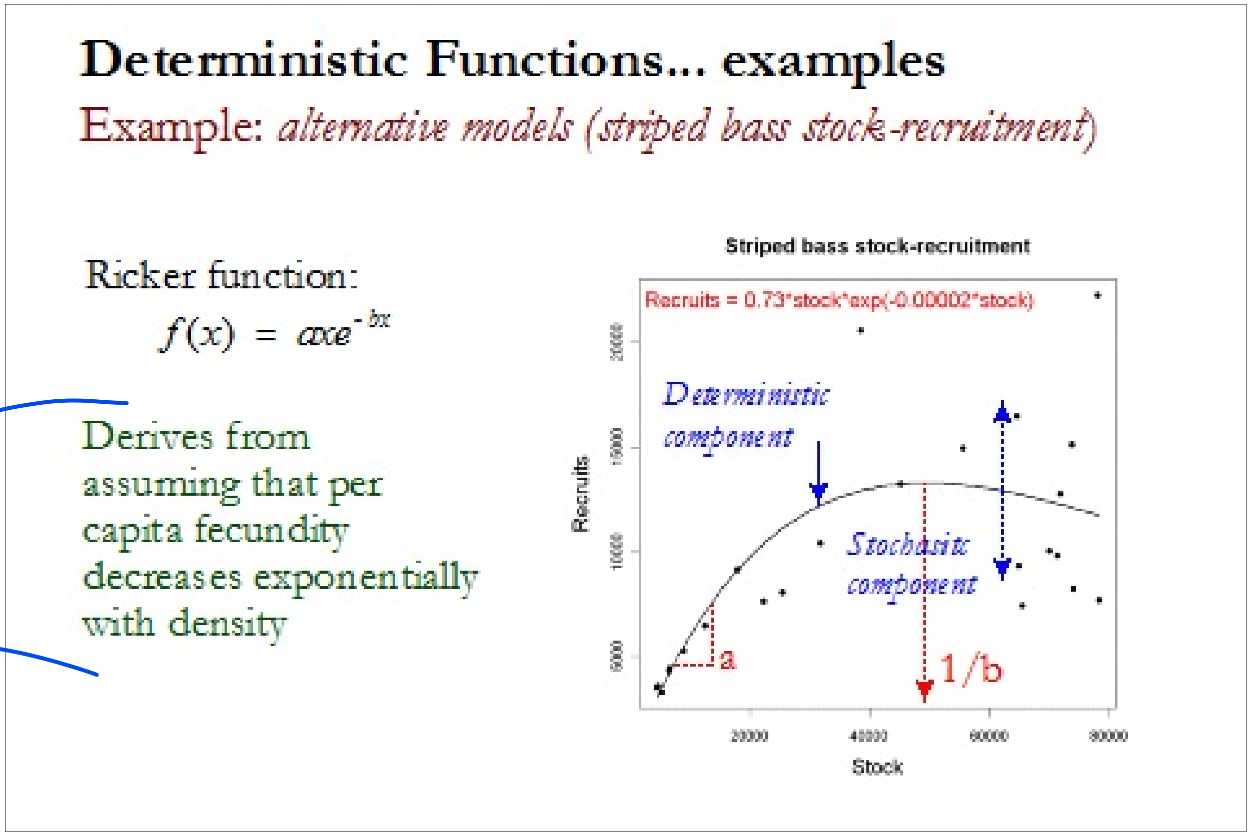


2.3 Example 3 – Striped bass stock-recruitment

In this last example, the data represent estimated stock levels (number of females) and recruits (age 1 numbers appropriately lagged to match the producing stock) of summer striped bass over a 24 year survey period.

- The first thing that you should notice about the plot is that recruitment initially increases with increasing stock levels, but then appears to either level off or peak and decline when the stock gets relatively large. This might be expected with density-dependent population growth if per capita fecundity decreases exponentially with density. Thus, there is potentially a mechanistic explanation to the pattern observed and we might propose a model that represents this mechanism.

- The second thing that you should notice is that there is considerable variability about the mean stock-recruitment relationship, and this variability increases considerably as the stock level increases (i.e., variance increases with the mean). This represents the *stochastic* component of the model. For now, we will ignore the stochastic component and focus solely on describing the deterministic component.



The Ricker function.– The Ricker function is one of several functions that has been used to define this sort of relationship, and has the following form:

$$y = a \cdot x \cdot e^{-b \cdot x}$$

where it starts off growing linearly with slope *a* and has its maximum at $x=1/b$. The Ricker function derives from assuming that per capita fecundity decreases exponentially with density. While it has a mechanistic interpretation in this case, the Ricker function is also widely used as a phenomenological model for environmental variables that start at zero, increase to a peak, and decrease gradually back to zero.

Deterministic Functions... examples

Example: *alternative models (striped bass stock-recruitment)*

Ricker function:

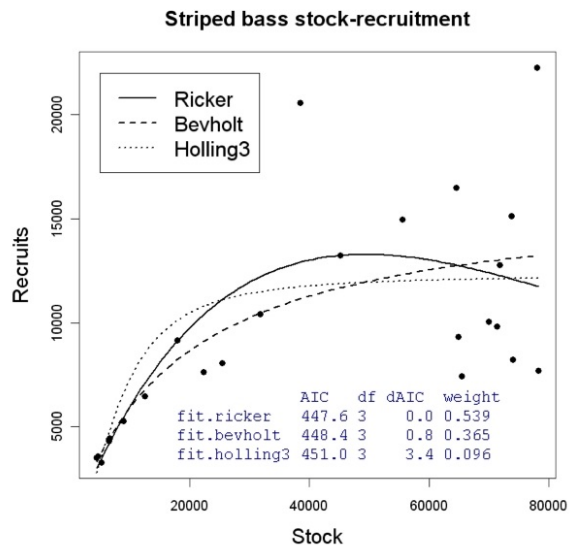
$$f(x) = axe^{-bx}$$

Beverton-Holt function:

$$f(x) = \frac{ax}{b + x}$$

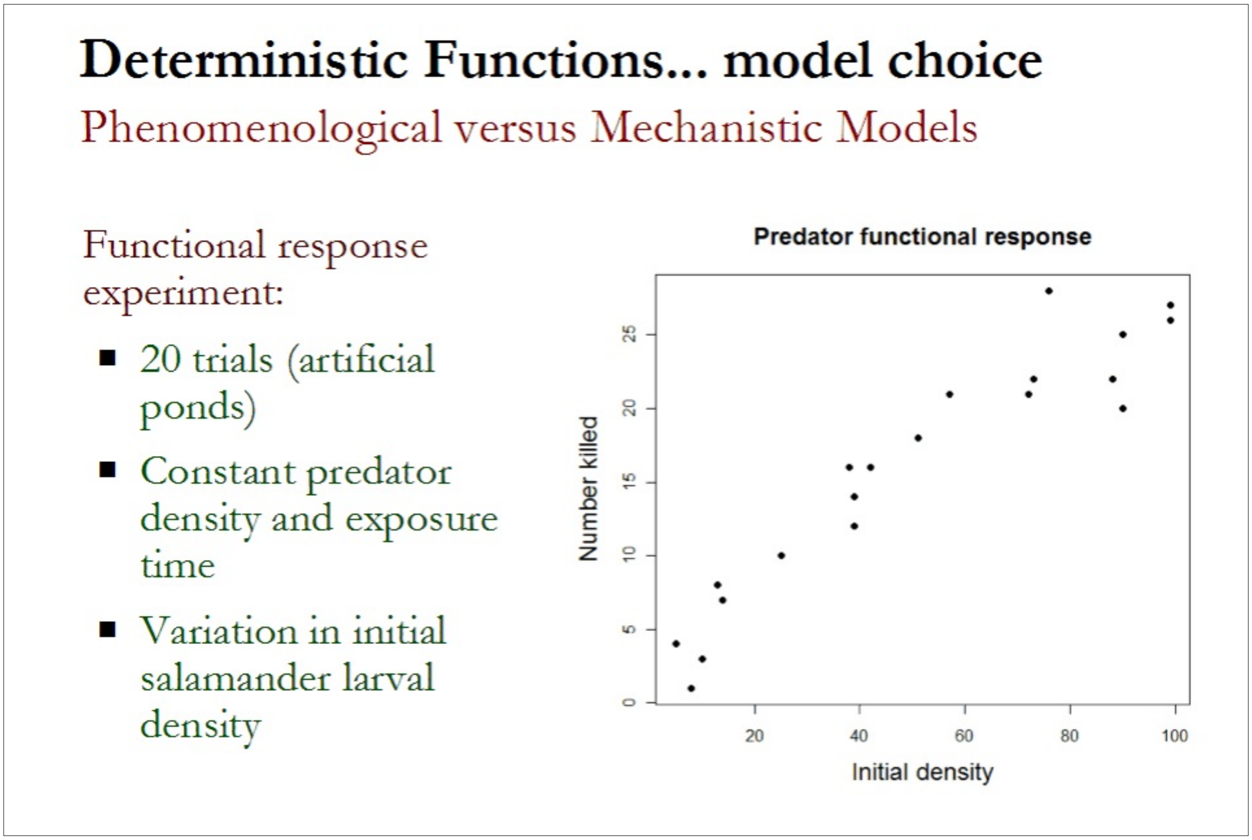
Holling type III function:

$$f(x) = \frac{ax^2}{b^2 + x^2}$$



As is often the case with environmental data, the Ricker function is not the only plausible mechanistic model for this data set. The Beverton-Holt function is also commonly used to describe stock-recruitment patterns and derives from assuming that over the course of the season the mortality rate of young-of-the-year is a linear function of their density (Quinn and Deriso 1999).

Importantly, while the Ricker and Beverton-Holt functions (and others) are logical mechanistic models for describing this stock-recruitment relationship, its is important to realize that there are several other functions that could be used to describe the pattern phenomenologically; i.e., without an explicit mechanistic explanation, and in some cases may fit the data better than the mechanistic function. We explore this issue further in the next section.



3. Phenomenological versus mechanistic functions

Phenomenological vs mechanistic

As noted at the beginning, it is usually a good idea to have an a priori idea of the expected form of the deterministic model, since this explicitly ties the model to environmental theory. A deterministic model constructed in this manner, such that the model parameters have a mechanistic relationship to an environmental process, is sometimes referred to as a *mechanistic* model. In this case, the model parameters have a direct environmental interpretation. However, there may be times when a purely *phenomenological* description of the data is sufficient; that is, when the deterministic model is derived based on the patterns observed in the data and not on underlying theory per se. In this case, the model parameters do not have an explicit environmental interpretation.

To illustrate the distinction, let's consider the results of a hypothetical functional response experiment. The data shown here are derived from a hypothetical study on predator-prey relationships involving larval salamanders and predacious aquatic invertebrates. The study involves 20 trials in artificial ponds containing a constant predator density and exposure time. The trials vary in the initial salamander larval density. The response variable is the number of salamander larvae killed. The figure shows the relationship between initial larval density and the number of larvae killed by aquatic predators.

Deterministic Functions... model choice

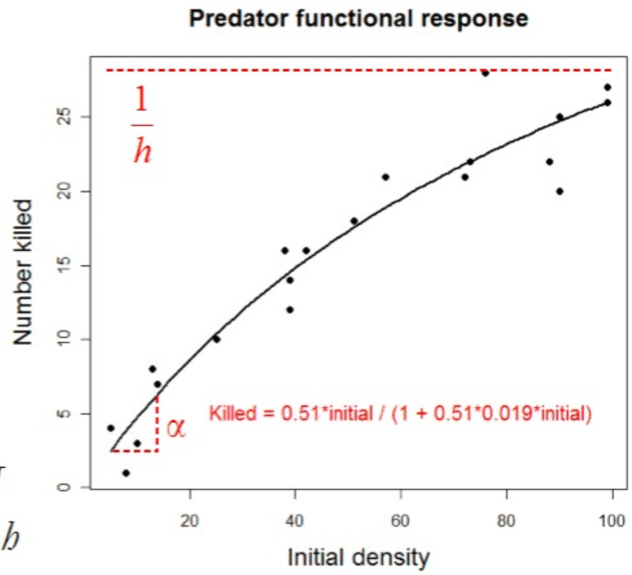
Phenomenological versus Mechanistic Models

Holling type II functional response function:

$$P(N) = \frac{\alpha N}{(1 + \alpha h N)}$$

P = predation rate
N = initial larval density
α = attack rate
h = handling time

@low density: $P(N) \approx \alpha N$
@high density: $P(N) \approx 1/h$



Given the experimental design, we would almost certainly have an a priori model of the functional response relationship expected. The standard model for saturating functional responses is the Holling type II (also known as Beverton-Holt and Michaelis-Menten) response,

$$P(N) = \frac{\alpha N}{(1 + \alpha h N)}$$

where P = predation rate (number eaten per predator per unit of time), N = initial starting larval density, α = baseline attack rate, and h = handling time. The Holling type II function assumes that the per capita predation rate of larvae decreases hyperbolically with density (= α / $(1+\alpha hN)$). In this case, the initial slope is α and the asymptote is $1/h$. Ecologically, this makes sense because at low densities the predators will consume prey at a rate proportional to the attack rate ($P(N)\approx\alpha N$) while at high densities the predation rate is entirely limited by handling time ($P(N)\approx 1/h$). It makes sense that the high-density predation *rate* is the inverse of the handling *time*: if a predator needs half an hour to handle (capture, swallow, digest, etc.) a prey, and needs essentially no time to locate a new one (since the prey density is very high), then the predation rate is 1/(0.5 hour)=2/hour.

Deterministic Functions... model choice

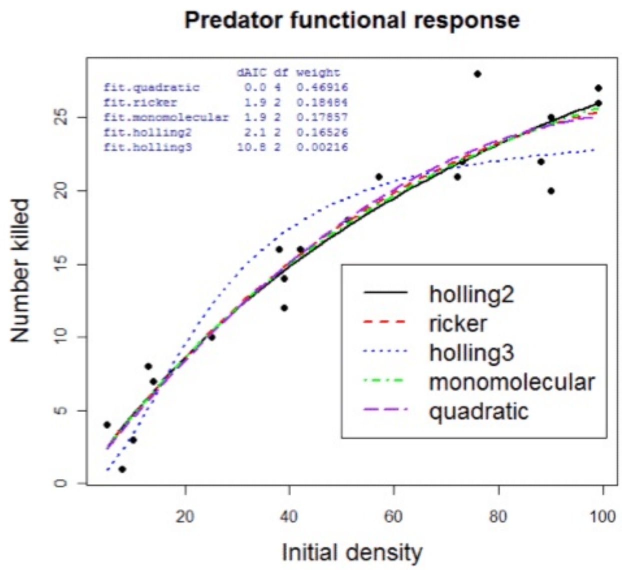
Phenomenological versus Mechanistic Models

Alternative mechanistic models:

- HollingII (Beverton-Holt; Michaelis-Menten)
- HollingIII
- Ricker
- Monomolecular (von Bertalanffy; skellem)

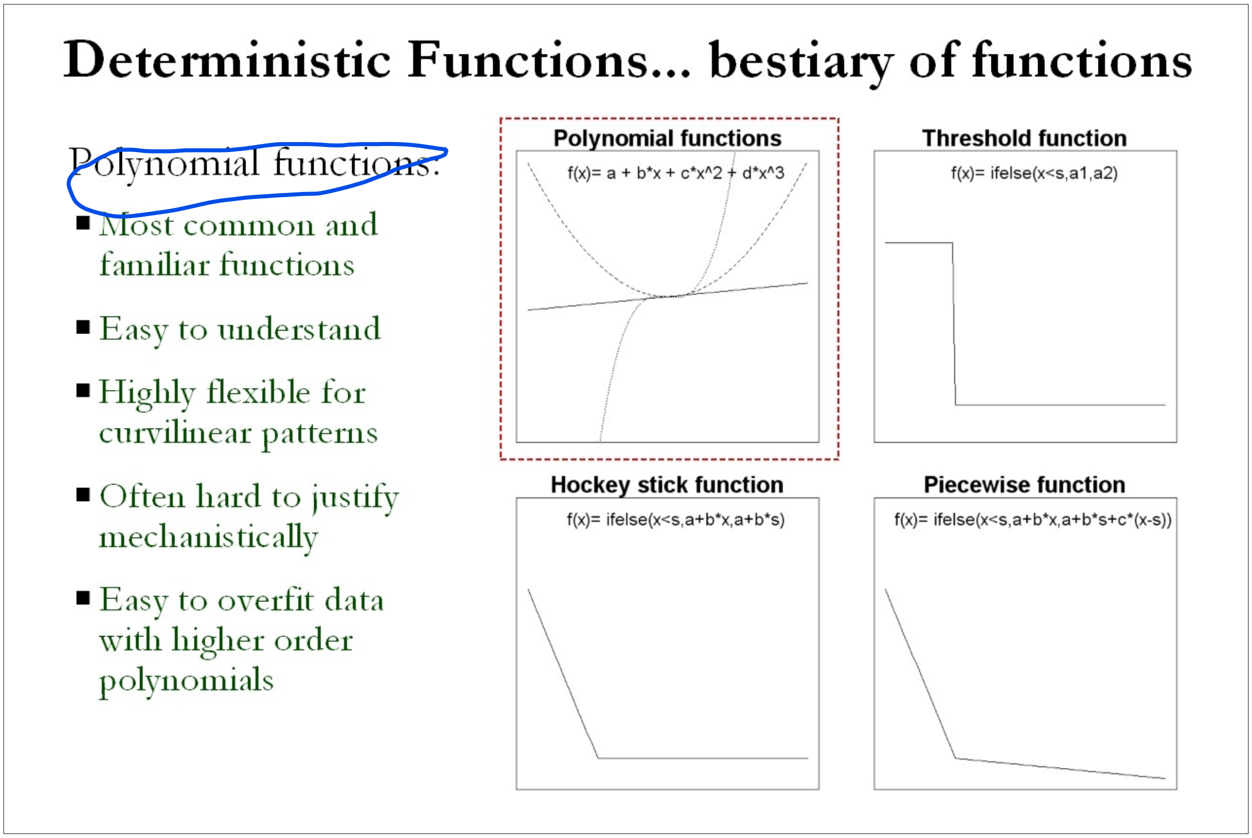
Phenomenological models:

- Quadratic polynomial
- Others?



While the Holling type II model certainly makes sense, it is not the only plausible mechanistic model to describe the functional response relationship. Holling type III, Ricker and Monomolecular (also known as von Bertalanffy and skellem) are also suitable candidates, each with its own mechanistic interpretation. As you can see from the figure, all of the models do a pretty good job of fitting the data. Indeed, the AIC criterion indicates that the Holling II, Ricker and Monomolecular models are all just about as good as each other. Based on the data, there would be little reason for choosing one over the other.

While these mechanistic models all make sense because they have an explicit environmental interpretation, it is important to recognize that there may be times (perhaps more often than not) when we don't have a plausible mechanistic model for the data a priori. In such cases, we can fit any suitable deterministic function that describes the relationship well. In the example here, the quadratic polynomial function fits the data extremely well, better than any of the mechanistic models in fact, and there are undoubtedly many other functions that could equally describe the pattern well. The difficulty lies in the fact that while we may be able to describe the pattern well, there is no underlying environmental theory tied to the model. Thus, we can describe the phenomenon well, but not the underlying mechanism. Despite these limitations, you will likely have many occasions where you will need to fit data on a phenomenological basis.



4. Bestiary of deterministic functions

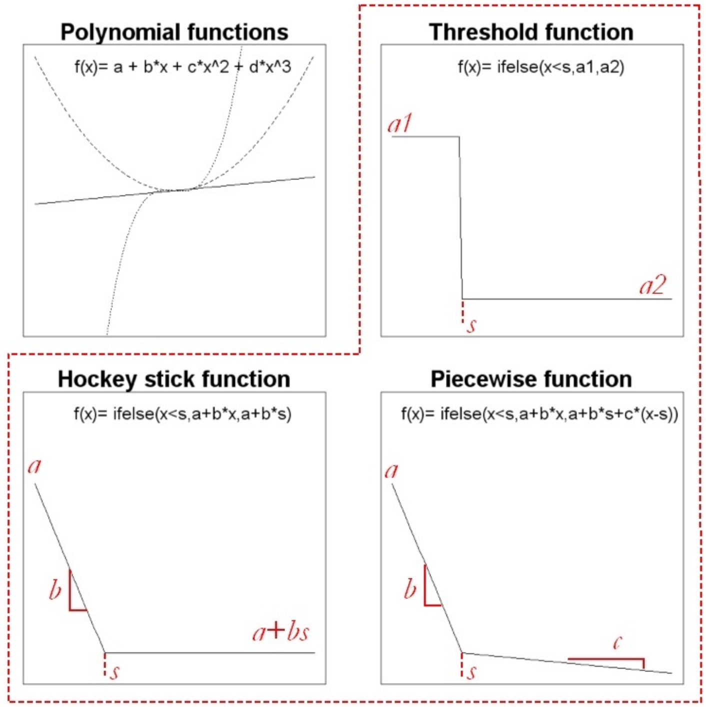
The realm of possible deterministic functions in environmental models is nearly infinite. A detailed understanding of the environmental system under study and a firm grasp of mathematics is all that is required to create your own deterministic function. Easier said than done, right? Fortunately, for most of us most of the time, we do not need to invent new mathematical functions because a very wide range of functions already exist. Thus, it behooves us to be aware of the range of extant possibilities. Here we will briefly review a bestiary of functions that are useful in environmental modeling (summarized from from Bolker 2008, but do see Bolker for a much more complete description of these functions), recognizing that this is but a sample of the many possibilities.

1. Polynomial functions.–Polynomial functions, including linear, quadratic and cubic functions shown here, are the most common and familiar functions. They are easy to understand and highly flexible for describing linear and curvilinear patterns; consequently, they have received widespread use. Unfortunately, they are often hard to justify mechanistically because the parameters are rarely derived from environmental theory. In addition, because of their flexibility it is very easy to overfit data with higher-order polynomials. In most cases, it is unwise to consider using polynomials higher than third-order (i.e., cubic).

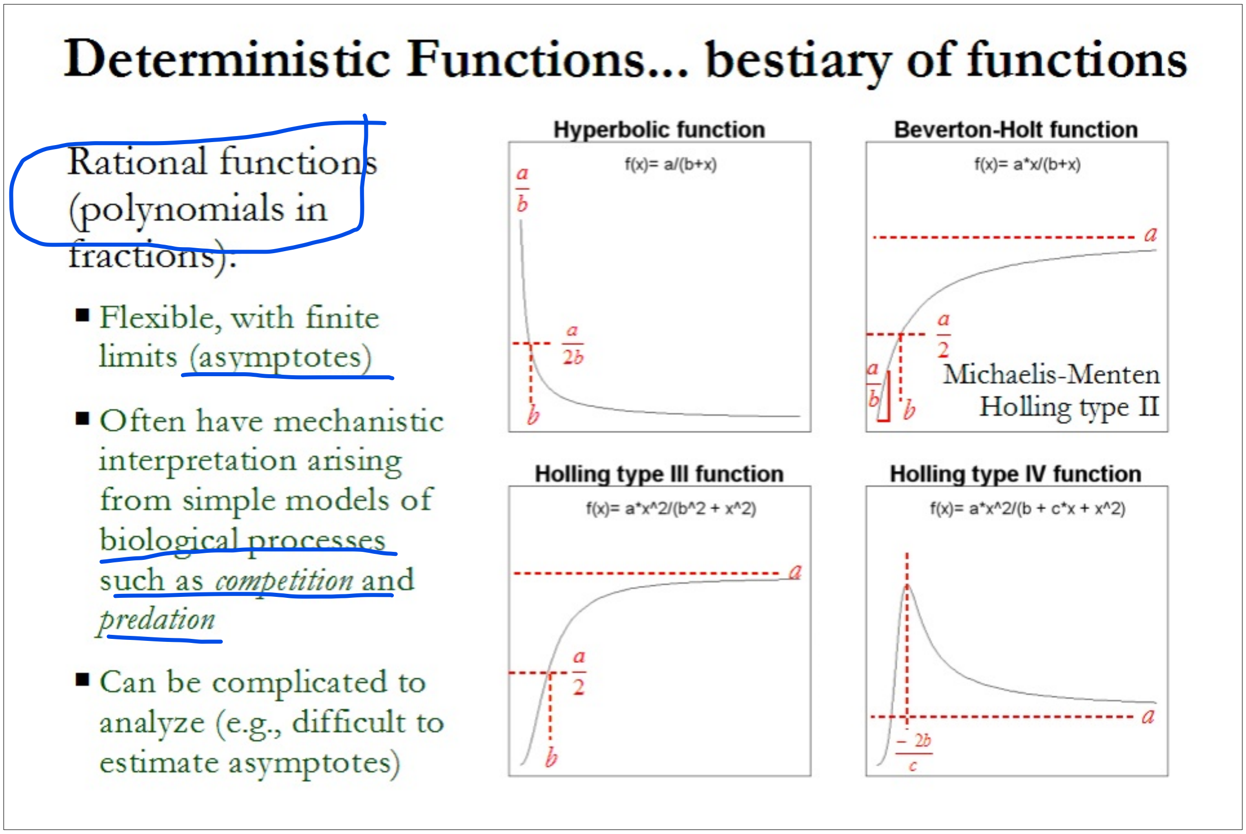
Deterministic Functions… bestiary of functions

Piecewise polnomial functions:

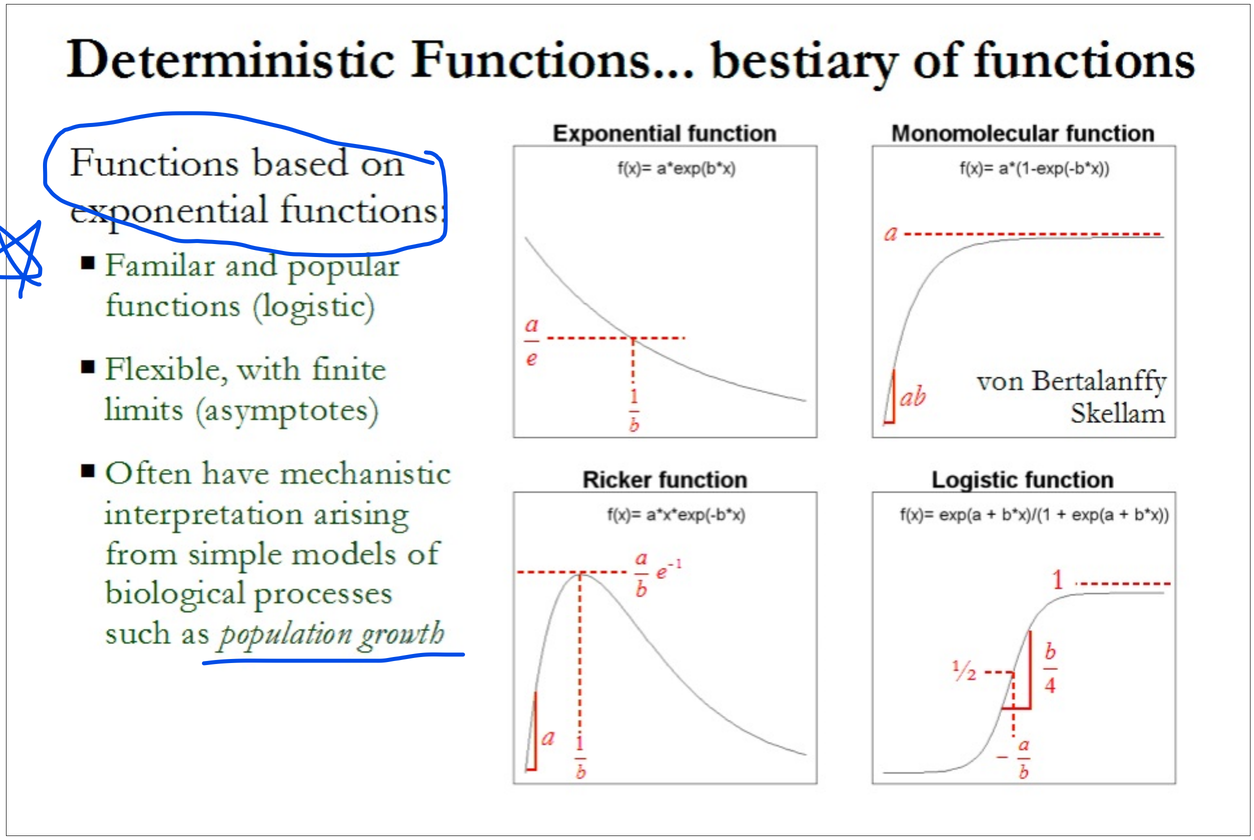
- Flexible for threshold-like patterns and for setting function limits
- Easy to understand
- Probably unrealistic in most cases, since abrupt thresholds are unlikely in ecological systems



2. Piecewise polynomial functions.–You can make polynomials (and other functions) more flexible by using them as components of piecewise functions. In this case, different functions apply over different ranges of the predictor (*x*) variable. Examples include the simple threshold function, hockey stick function, and more generalized piecewise function shown here. Like polynomials, piecewise polynomials are also easy to understand since they involve piecing together two or more polynomials. Piecewise polynomials are extremely flexible for fitting threshold-like patterns and/or phenomenologically as a simple way to stop functions from dropping below zero or increasing indefinitely when such behavior would be unrealistic. Using a piecewise function means that the rate of change (derivative) changes suddenly at some point. Such a discontinuous change may make sense, for example, if the last prey refuge in a reef is filled, but transitions in environmental systems usually happen more smoothly.



3. Rational functions (polynomials in fractions).–Rational functions are ratios of polynomials. Examples include the hyperbolic function, Beverton-Holt function (also known as Michaelis-Menten and Holling type II functional response depending on the discipline), and Holling type III and type IV functional response functions. Rational functions are extremely flexible, simple to compute, and are typically used when there are finite limits (asymptotes) at the ends of their range. They often have a mechanistic interpretation arising from simple models of biological processes such as competition and predation. However, they can be complicated to analyze because the quotient rule makes their derivatives complicated. In addition, because they approach their asymptotes very slowly, estimating the asymptote can be difficult.

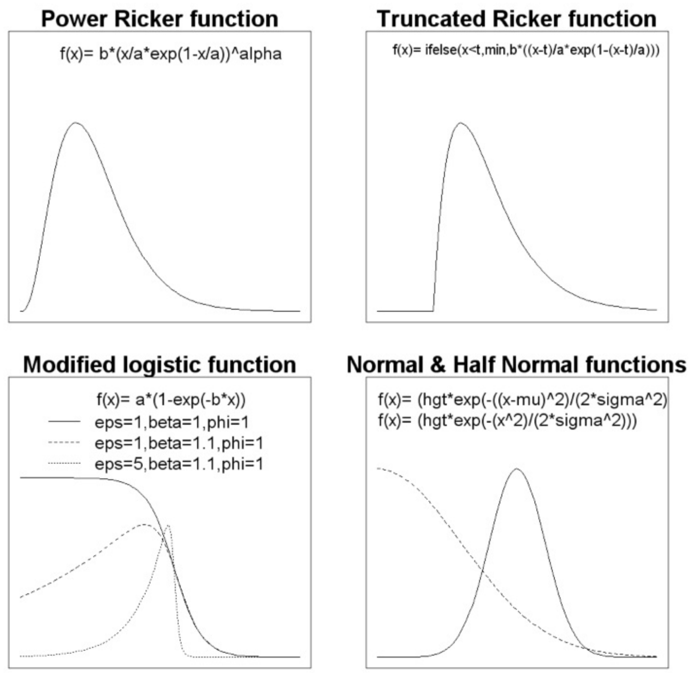


4. Simple exponentials and combinations of exponentials with other functions.–Simple exponentials include the exponential growth or decay function and saturating exponential growth functions such as the monomolecular function (equivalent to the simplest form of the von Bertalanffy growth curve in organismal biology and fisheries, and the Skellam model in population ecology). More complex functions involving exponentials in combination with other functions include the popular Ricker and logistic functions. All of these exponential functions are highly flexible and most have finite limits at the ends of their range like the rational functions. Like rational functions, these exponential functions also typically have a mechanistic interpretation arising from simple models of biological processes such as population growth.

Deterministic Functions... bestiary of functions

Functions based on exponential functions:

- Modifications of simpler functions to provide greater flexibility
- Mostly phenomenological models

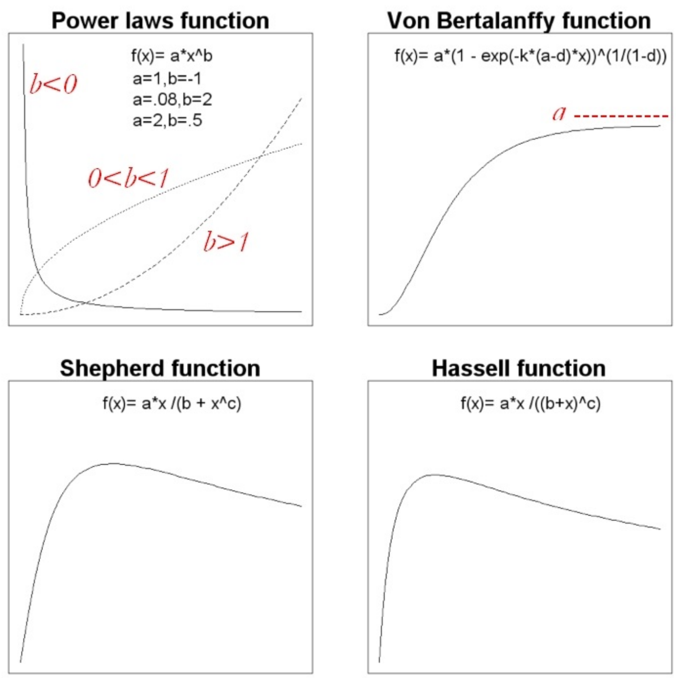


There are modifications to these functions, including the power Ricker, truncated Ricker and modified logistic, to render them more flexible in accommodating a wider range of patterns. However, these modified functions are typically phenomenological models designed to fit the data better than the original model, but lacking in an explicit mechanistic interpretation. The familiar normal (or Guassian) function (and associated half-normal function) is perhaps the most common phenomenological model in environmental applications.

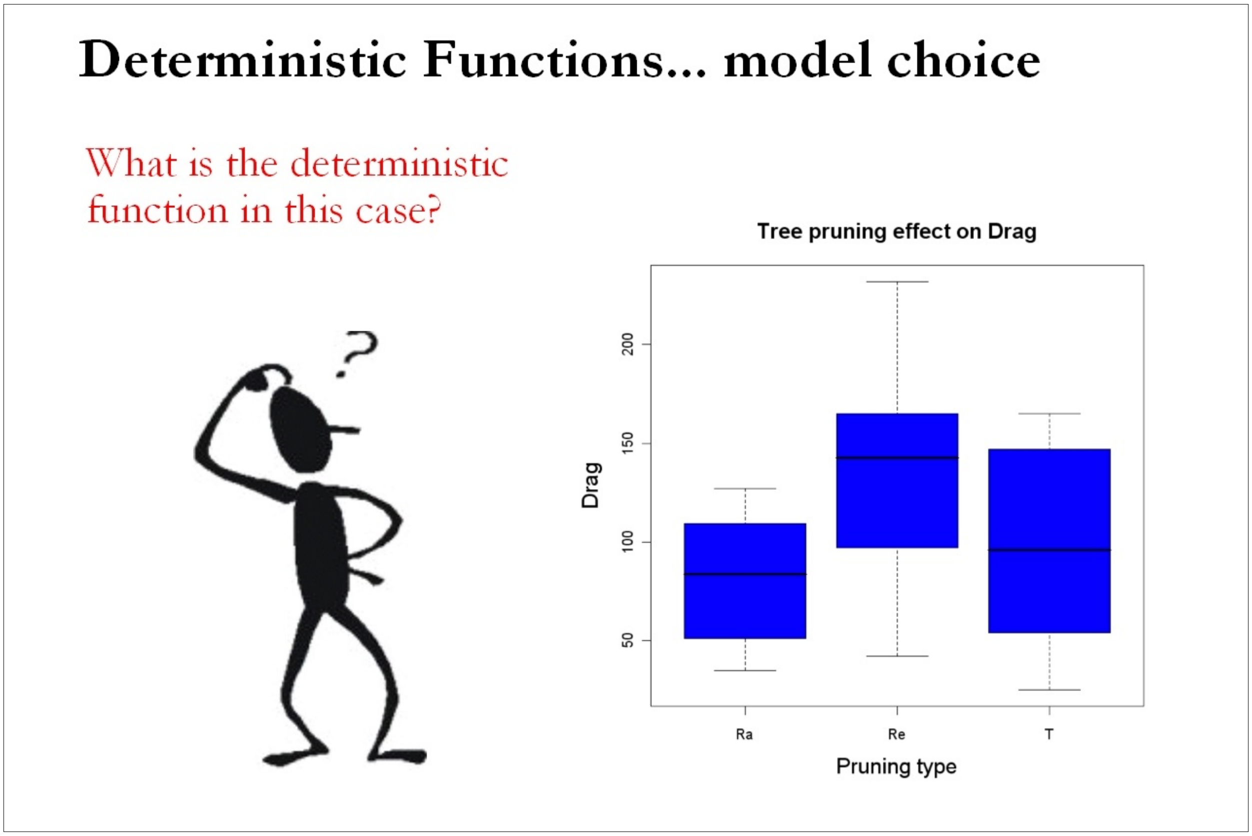
Deterministic Functions… bestiary of functions

Functions involving power laws:

- Flexible, especially since the power parameter is usually added to existing model that already allows for changes in location, scale and curvature
- Exponent sometimes derived from intrinsic geometric or allometric properties of the system and does not have to be estimated



5. Power laws.– The polynomials involved in the rational functions above were simple linear or quadratic functions. Environmental modelers sometimes introduce an arbitrary (fractional) power as a parameter instead of using only integer values; using power laws in this way is often a phenomenological way to vary the shape of a curve, although these functions may also have mechanistic derivations. Examples include the generalized version of the von Bertalanffy growth curve, a generalized version of the Beverton-Holt (Michaelis-Menten; Holling type II) function known in fisheries as the Shepherd function, and the closely related Hassell function. Power functions are extremely flexible, especially since the power parameter is usually added to an existing model that already allows for changes in location, scale and curvature. The exponent is sometimes derived from intrinsic geometric or allometric properties of the system and thus does not always have to be estimated.



5. Choosing the right deterministic function?

Determining the appropriate deterministic function is fraught with difficulties in some studies, while in other cases, the function requires little to no thought. Take the example shown here from a study on the effect of three different tree pruning methods on the "drag" caused by wind. Note, drag is a measure of stress on the structure of the tree caused by the wind. You can imagine that an arboriculturalist would be very interested in knowing how various pruning methods effect drag and, ultimately, failure of the tree. The data in this figure show the measured drag under a range of wind speeds across three pruning methods for one tree species, Freeman maple. What is the deterministic model for analyzing the effect of pruning method on drag?

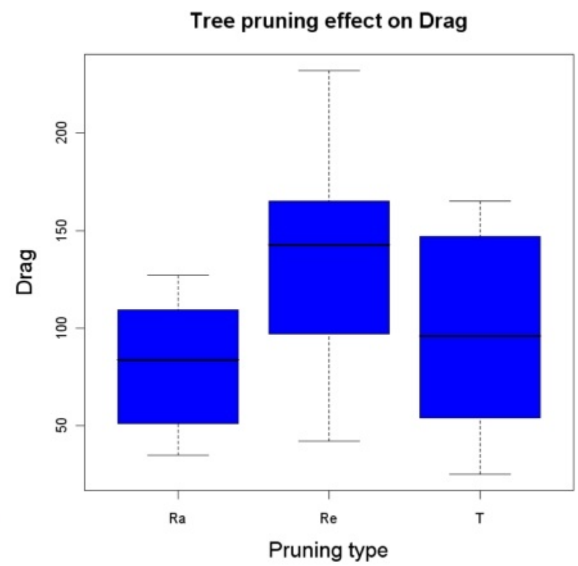
Deterministic Functions... model choice

Tree	Prune	Drag	X1	X2
29	Ra	35	0	0
40	Ra	60	0	0
43	Ra	35	0	0
...	...	...	...	...
26	Re	101	1	0
27	Re	63	1	0
28	Re	42	1	0
...	...	...	...	...
25	T	49	0	1
30	T	25	0	1
34	T	58	0	1
...	...	...	...	...

$$Drag = a + bX1 + cX2$$

(a=81, b=50, c=18)

E(Prune=Ra)=81
E(Prune=Re)=81+50=131
E(Prune=T)=81+18=99

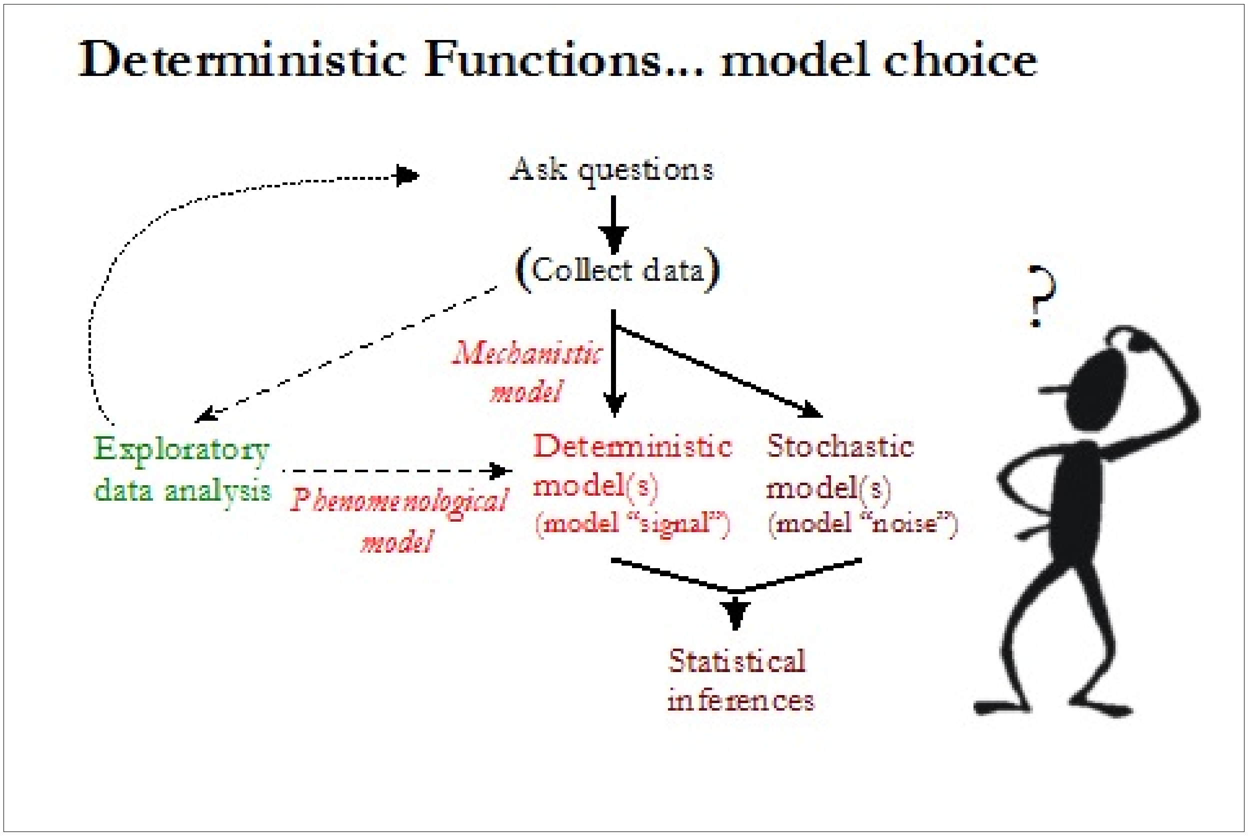


Since Pruning type (the independent variable) is a factor (i.e., categorical treatment variable), it gets transformed into "dummy variables" coded 0 or 1. With a three-level factor (Pruning types Ra, Re, and T), two dummy variables are required to uniquely identify an observation as belonging to one of the three levels. The first level (Ra) is distinguished by having a 0,0 coding for the two dummy variables; the second level (Re) is distinguished by having a 0,1 coding for the two dummy variables; and the third level (T) is distinguished by having a 1,0 coding for the two dummy variables. The deterministic model is simply a linear model of the form y = a + bx, only here we have two dummy variables to substitute for the original factor (Prune), so the final linear model becomes:

$$y\ (Drag) = a + bX1 + cX2$$

where *y* is the dependent variable, Drag, *a*, *b* and *c* are parameters to be estimated from the data, and *X1* and *X2* are the independent (dummy) variables. The derived estimates for the parameters (*a*, *b*, and *c*) allow us to predict the expected value (the mean) for an observation in any of the treatment levels. Specifically, the expected values for observations in each treatment level are computed as:

Expected Drag for observation in treatment Ra = 81 + 50*0 + 18*0 = 81
Expected Drag for observation in treatment Re = 81 + 50*1 + 18*0 = 131
Expected Drag for observation in treatment T = 81 + 50*0 + 18*1 = 99



In the modeling process, the choice of a deterministic model is critical because it is usually the way we represent the environmental process of interest. Estimating the parameters of the deterministic model is more often than not the focus of the statistical inference. This is particularly so if the model is mechanistic and the parameters have an explicit environmental interpretation. Thus, choosing a model carefully is of paramount concern. Ideally the model choice is made a priori, before you have looked at the data, but there will be many times where an initial examination of the data will provide important insights on the adequacy of a particular model and suggest a different model or perhaps competing models. Time spent carefully considering the right model for the question, given the data, is time well worth spending, as any inferences made are going to be contingent on the model, as are any insights gained from the study.