Lecture 1: Introduction to statistical modeling in ecology

WILD6900

updated 2018-11-26

REVIEW SYLLABUS

General philosophy - You are voluntarily in grad school. You are voluntarily taking this course. Therefore, I assume you want to learn this material and I do not need to use punative measures to force you to learn. My objective is not to give a specific grade distribution. My goal is for you to learn. If you put in the time, you will be able to master the material. If you do, you will be an A.

STATISTICAL MODELING IN ECOLOGY

A quick note on notation

In this course, I will try to follow the notation used in *Hobbs and Hooten 2015*. Deterministic functions will be denoted using a lowercase letter followed by parentheses:

$$g()$$

with arguments specific to that function in the parentheses. Remember that these are *deterministic*, meaning that when we put in specific values and data, we will always get the same answer. There is no uncertainty. Deterministic functions are used in virtually all ecological modeling, usually to represent our hypothesis about how our system works.

In many cases, the outcome of a certain process cannot be perfectly predicted. There is uncertainty (e.g., a coin flip). Although we cannot predict the outcome of these *stochastic* processes with certainty, we can describe them probabilistically. A *random variable* is a quantity that can take on values due to chance. The change of each value is governed by a probability distribution. Probability distributions will be denoted using square brackets:

$$[a|b, c]$$

which means a is a random variable conditional on the parameters b and c. We will discuss what | (read "conditional on") later so don't worry if that is a bit confusing right now.

What are we modeling?

All of us use models to make inferences about the state of some ecological process that we are interested in and why it changes across space, time, individuals, or populations. To understand these processes, we form hypotheses and then collect data. Our goal is to understand processes that we **cannot directly observe** based on quantities that we *can* observe. Data collection involves sampling from our population of interest and collecting observations of the system state(s).

Uncertainty enters every part of this process. Decreasing uncertainty is always our objective - that is what science seeks to do. However, to decrease uncertainty we must acknowlege where it enters our inference and

what type of uncertainty it is. Not all uncertainties behave the same way and the tools that we can use to reduce uncertainty depend on what type of uncertainty it is.

To understand where uncertainty enters the scientific process, it is useful to break down the modeling process into distinct components. This is not usually the way we are taught to approach ecological modeling but as you will hopefully see throughout the semester, thinking this way will improve not just the way your approach analyzing your data, but also how you plan your studies and communicate your results.

Process models (add example here)

Process models provide a mathematical description of the *state-variables* we are interested in and how they change over space and time. These represent the **true** value of our state variables at any given point in space or time.

State-variables are the ecological quantities of interest in our model (e.g., N, z). *Parameters* determine how state-variables change over space or time (e.g., r, K, ψ)

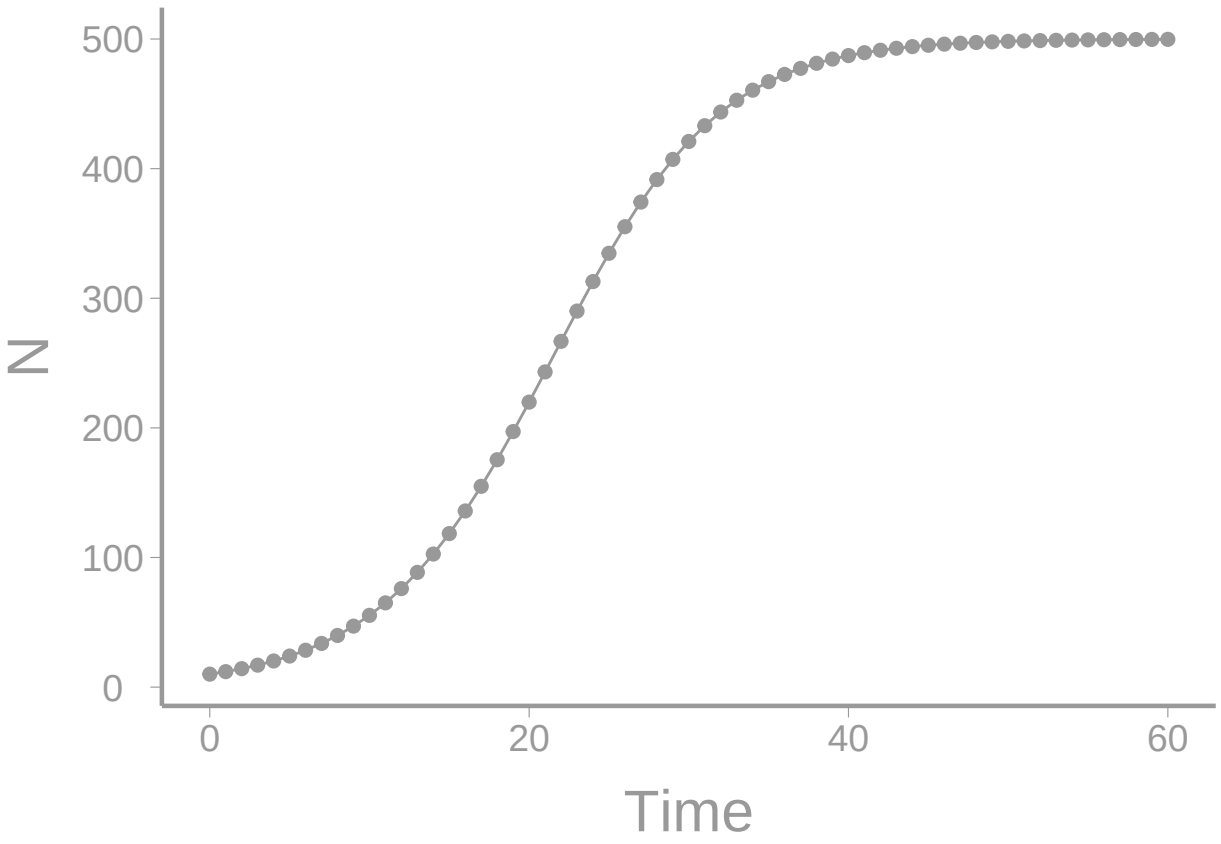
Process models are deterministic - they describe our hypothesis about how the system works; i.e., they describe the state of the system and the parameters that influence it.

Process models are abstractions - they inherently leave out a lot of details about the system in order to focus on the processes that we are most interested in or think are most important. We treat all the other sources of variation as a source of *uncertainty* - that is, unexplained variation in the state of the system. We can represent these uncertainty stochastically by defining a parameter σ_p^2 (p is for process) that subsumes all of the unmodeled sources of variation in the system. This allows us to model the *probability distribution* of the state-parameters:

$$[z|g(\theta_p, x), \sigma_p^2]$$

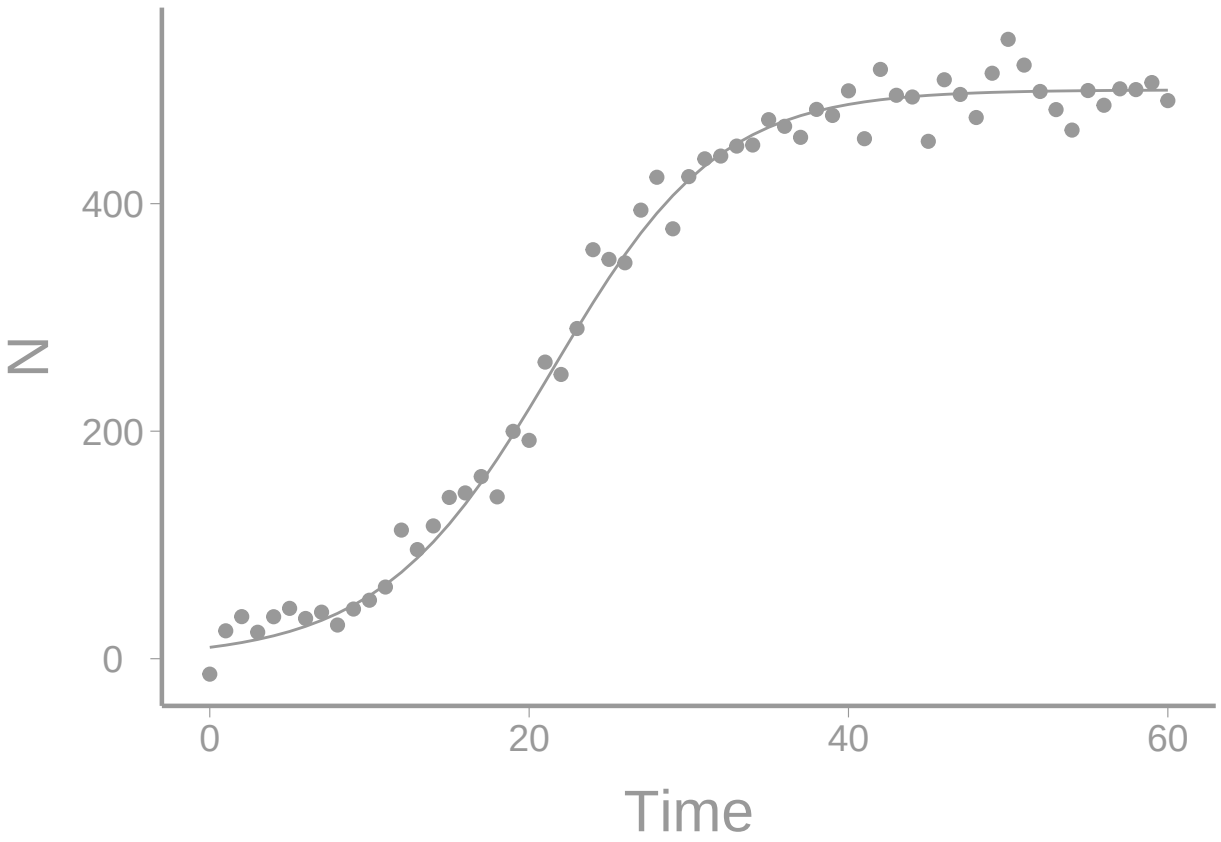
where $g(\theta_p, x)$ is a deterministic function that describes our hypothesis about how some predictor(s) x influence the state variable z based on parameter(s) θ_p. For example, perhaps we are interested in predicting the population growth on species a as a function of abundance in the previous year. We hypothesize that population growth rate will be highest at low densities and lowest (maybe even negative) at high density. This leads us to believe that the *discrete logistic equation* is a good descripter of our system, so:

$$N_{t+1} = g(\theta_p, x) = N_t e^{r_0[1-(N_t/K)]}$$



In this case N is both our state variable (z) and a predictor in the model (x). $\theta_p = (r_0, K)$ are the parameters that control the relationship beteween N_t and N_{t+1}.

In the above example, all of the populations sizes fall exactly on the predicted line, which represents the implausible scenario that process uncertainty (σ_p^2) is zero. More likely we would expect something like:



In this case, there are clearly processes other than density-dependence that are influencing population size in each year. So $sigma_p^2 > 0$.

Remember that in this example, we are modeling the **true** value of N, *not* our observation of it. One way to interpret $sigma_p^2$ is as a measure of how well our process model fits reality - the smaller σ_p^2, the better our model represents the system. So to minimize process uncetainty, we need *a better model*. This is critical. No amount of additional data collection will lower σ_p^2.

In our example, maybe we have enviromental covariates (rainfall, temperature, etc.) that we also think are important. To reduce process uncertainty, we would need to modify our process model to include these effects (we'll learn more about hwo to do this later in the semster).

Sampling models

To obtain probability distributions about our state-variables and parameters, we need data. Data are samples of the true population. This introduces another source of uncertainty because our sample will not perfectly represent the true state of the system. As for the process model, we can represent sampling uncertainty σ_s stochastically using a probability model:

$$[u_i|z, \sigma_s^2]$$

In our population size example, suppose we conduct transect or point counts to estimate abundance. The area of our counts (we'll call it a) is not the entire area of our population ($a < A$). If we want to estimate N_t, we need a model linking our counts (call it n_t) to the true population. If we assume individuals are uniformly distributed across our study area, then perhaps we could use:

$$\left[n_t \middle| \frac{N_t}{a}, \sigma_s^2\right]$$

in this case, our count n_t will be different if we had chosen different transect routes or points simply due to chance. This is waht $sigma_s^2$ represents. Separating σ_s^2 from σ_p^2 is important because we *can* lower σ_s^2 by collecting larger sample sizes, increasing replication (i.e., doing a better job of sampling the population). For example, if we could sample the entire area A, σ_s^2 would be 0.

Observation models

In the sampling model above, you may have noticed that $sigma_s^2$ only captures uncertainty that is due to the randomness of our sampling process. If we used the n_t as our observation, we would have to make the assumption that we counted every individual in our sampling area. This is almost never a good assumption in studies of animals (and even plants!). We almost always observe the system imperfectly. Animals are elusive and may hide from obsevers. Individual birds may not be singing during our point counts. Evem plants may be cryptic and hard to find even when sampling plots are small.

Another way to say this is that n_t is the *true* number of individuals in our sampling area but cannot count every individual. Observation uncertainty can lead to biased estimates of model parameters and so generally must be accounted for using an exlicit observation model:

$$[y_i|d(\Theta_o, u_i), \sigma_o^2]$$

In our count model, we might define a parameter ψ that is the probability that individual that is present in our sample is counted by the observer (we could further use a generalized linear model to account for the effects of, e.g., weather or observer skill, on ψ).

$$[y_t|\psi n_t, \sigma_o^2]$$

where σ_o^2 is uncertainty about the value of ψ.

Including an observation model is needed to correct for bias. If we assume that the observed value equals the true value when in fact it doesn't, than we will get biased estimates of z. Observation error is pervasive in ecological studies. In occupancy models we don't know if a site is truly unoccupied or if we failed to detect our study species. When we take morphological measureements, our instruments have some error that prevents us from knowing the true size of an individual or feature. Thinking hard about observation processes and how we can isolate them from sampling and process uncertainty will be an a key theme throughout the semester.

Parameter models

In a Bayesian world, there is one additional level of modeling (this is probably true in frequentist analyses too we just don't explicitly acknowledge it). Parameter models express what we know about our parameters *prior to* collecting data. For this reason, parameter models are more commonly referred to as *priors*. Every parameter in our model requires a probability distribution describing the prior probability we place of different values the parameter could take.

$$[\theta_p][\theta_o][\sigma_p^2][\sigma_s^2][\sigma_o^2]$$

These distributions can provide a lot or a little information about the potential value of each parameter. We will talk more about priors in the coming weeks.

The full model

With each of our models created, we are prepared to right out the full model:

$$\Big[\underbrace{z, \theta_p, \theta_o, \sigma_p^2, \sigma_s^2, \sigma_o^2, u_i}_{unobserved} \Big| \underbrace{y_i}_{observed}\Big] \propto \underbrace{[y_i|d(\Theta_o, u_i), \sigma_o^2]}_{Observation\ model} \underbrace{[u_i|z, \sigma_s^2]}_{Sampling\ model} \underbrace{[z|g(\theta_p, x), \sigma_p^2]}_{Process\ model} \underbrace{[\theta_p][\theta_o][\sigma_p^2][\sigma_s^2][\sigma_o^2]}_{Parameter\ models} \quad (1)$$

At this point, equation 1 probably looks pretty confusing. Don't worry - over the next few weeks we will take apart each piece of this process to understand how to build a model like this from the ground up using your data. This will involve choosing sensible process and observation models as well as choosing appropriate probability distributions for the stochastic quantities in the mode. We will also learn about how to estimate the model parameters.

For now, notice that all of the quantities that we do not directly observe (parameters and state-variables) are on the left side of the conditioning symbol "|". This means that they are treated as *random variables* that are governed by probability distributions. This distinction - treating all unobserved quantities as random variables and specifying probability distributions for each quantity, is what makes this model *Bayesian*.

Thoughts on the modeling process

The approach just outlined provides a rigorous strategy for modeling ecological data in way that explicitly acknowledges the hierachical nature of most data and separates different sources on uncertainty based on when and how it enters our inference. This process has huge advantages and is one we should all strive for.

However, in many cases, full separating process from sampling from observation uncertainty may not be possible (and in some cases not necessary). In other words, you will not always follow this process for every analysis you do. That's ok. You can still do science without following this approach to a T. But the process of *thinking hard* about what process your trying to model and how your observations do (or do not) represent the state variable of interest in enormously useful. It will help you design your sampling methods before you collect data (this is one of the most important advantages!), help you think through and make sense of complex analyses, and help you communicate your methods and results more effectively. If you take nothing else away from this course, I hope you realize the advantages of separating process, sampling, and observation processes.

WHY BAYESIAN?

The use of Bayesian methods is growing rapidly in ecology. This largely due to powerful computers and the development of accessible software for fitting Bayesian models using *Markov chain Mote Carlo* methods (we'll talk more about what that means later). There is also a large amount of code online and in books that be used as the starting point for analyses. These developments have led to many people adopting Bayesian methods even if they don't fully understand what they're doing. It also led to some ecologists becoming full-on Bayesians and then a preditable backlash from those that view these methods as needlessly complex and overly trendy (i.e., statistical machismo).

My view is that the methods you'll learn about in this class, just like all statistical methods are *tools* to help you answer questions. Your job as a researcher is to choose the tools that best suite your question and your data. If that means using a t-test, great. Use that. But often in ecology we have to deal with many layers of complexity, which means using more complex methods. My goal for this course is to expand your toolbox.

Why are you interested in learning Bayesian methods?

Philosophical advantages

1) Probabilistic treatment of all unknown quantities
2) Coherent framework for incorporating prior knowledge into analysis
3) Proper accouting of uncertainty
4) Ease of estimating latent variables (and uncertainty)

Practical advantages

1) Ability to develop custom/complex models to suite your needs
2) Many statistical concepts (e.g., random effects) make more sense (no blackbox)
3) Expanded "toolkit" for quantitative analysis
4) Ability to keep up with the literature
5) Ability to review manuscripts/proposals that use Bayesian methods

Disadvantages

1) Computationally intensive
2) Few (no?) "canned" software

REPRODUCIBLE RESEARCH

One of the secondary objectives for this course is to help students develop a skillset for making their research more reproducible.

What makes research "reproducible"?

Many definitions (Goodman et al. 2016) but for the purposes of this class, we will define it as:

> The ability of independent researchers to reproduce scientific results using the original data and methods (adapted from Markwick et al. 2018)

This means that, given the raw data, someone other than you could recreate your analysis. At a minimum, this means someone could re-run your analyses and come up with exactly the same answer. Stricly speaking, reproducing results could also include the ability to recreate figures, tables, and even text as it appears in the original report or paper.

Obviously, reproducibility is a gradient from unreproducible (no data, no methods, etc.) to completely reproducible (the ability to start with raw data and reproduce all results, figures, tables, etc). Most studies fall in between. Fully reproducible research is *very* hard but that does not mean we shouldn't strive for moving our work closer to to that end of the spectrum.

Why make your research reproducible

Often times, making your work reproducible involves extra work. Why bother?

1) To help advance science - if your work can't be reproduced, it's not science

2) To meet requirements of journals/granting orgs

3) To make it easier to share your work with collaborators

4) To make it easier to revise your analysis later

   You always have at least one collaborator on every project - you future self. And your past self doesn't respond to email

Course philosophy

In my experience, courses like this one are the primary source of experience for students to work with data outside of their own research. But methods-based courses generally focus only on the modeling side of things. To do this, we give students clean, rectagular data sets that are ready to go into the model. This approach ignores the fact that **most** of your time will be spent wrangling your raw data to get it ready for analysis.

Those data wrangling tasks are extemely important but we don't teach them (even though we spend a lot of time teaching you how to collect and analyze data). So students have to figure out how to do this on their own. That is, at best, inefficient. Worse, it usually leads to the development of bad habits that can slow your progress, make it difficult to share your data and analysis with your advisor/collaborators, and worst of all lead to errors in your analysis.

In the "lab" activities in this course, we will emphasize the tasks of cleaning and processing raw data to prepare it for analysis. We will also use tools that allow us to document that process and make it easy to reproduce. In addition to the statistical knowledge you gain in the course, I hope you leave with a better workflow for collecting, entering, processing, documenting, analyzing, and reporting.

The past few years have seen a big growth in tools and practices for reproducible research. This guide from BES has a ton of good information and links to additional resources:

- Cooper, N. & Hsing, P. (2017) *A guide to reproducible code* British Ecological Society

References

Goodman, S. N., Fanelli, D., and Ioannidis, J. P. A. (2016) *What Does Research Reproducibility Mean?* Science Translational Medicine, 8, 341ps12–341ps12.

Marwick, B., Boettiger, C. & Mullen, L. (2018) *Packaging Data Analytical Work Reproducibly Using R (and Friends)*, The American Statistician, 72:1, 80-88, DOI: 10.1080/00031305.2017.1375986