# Data and Models

M. Drew LaMar
September 19, 2016

> "Statisticians, like artists, have the bad habit of falling in love with their models."
>
> - Box

Introduction to Quantitative Biology, Fall 2016

# Class announcements

- Digital textbook from bookstore?
- Exam will be handed out THIS Friday
  - Lab on Thursday will be review
- Learning objectives for Chapters 4, 7, and 8 posted on Timeline
- Homework #1 and #2 solutions posted on Blackboard
- Lecture recording worked!!

# The Importance of Good Data

**Quote:** "Meaningful data of sufficient quantity are the grist of scientific bread."

1. Is the study sound so that an inductive inference can be justified?
   - Experimental design should be able to address predictions from…
   - …one or a number of scientific hypotheses that have been well thought out.
2. Are the data analysis methods sound? Relies on…
   - …adequate *modeling* and
   - objective approaches to model selection.

# Information and Statistics

**Quote:** "If data are collected in an appropriate manner, then there is **information** in the sample data about the process or system under study."

- Mathematical model is required (in most cases) to obtain information from the data.
- Inductive vs deductive reasoning
- Inductive: "inference of a generalized conclusion from particular instances" (sample -> population)
- Statistics adds **rigor** to the inductive process.
- The **inference** comes from a *model* that approximates the system or process of interest.

# Models in Science

Quantification is *essential* due to variation and complexity.

**Quote:** "Unless one is engaged in simple descriptive studies, they [the empirical sciences] must deal with mathematical models."

**Quote:** "We are not trying to model the data; instead, we are trying to model the information in the data."

**Quote:** "Data contain both **information** and **noise**; fitting the data perfectly would include modeling the noise and this is counter to our science objective."

# Models in Science

**Quote:** "Models must be derived to carefully represent each of the science hypotheses."

$$H_1 \Leftrightarrow g_1,\ H_2 \Leftrightarrow g_2, \ldots, H_k \Leftrightarrow g_k.$$

**Scientific Question:** What is the support or empirical evidence for the ith hypothesis (via its corresponding model), ***relative to others in the set.***

**Model Selection:** What is the the ***evidence*** for each of the hypotheses (and their associated models), ***given the data***.

# Models are Approximations

> "All models are wrong, but some are useful."
>
> - Box

**Example**: Population survival

$$n_{t+1} = s \cdot n_t$$

**Assumptions**:

- Population survival rate $s$ does not change over time.
- Each individual most likely has a different survival rate ($s$ represents the population average).
- Biotic and abiotic factors that influence survival rate are being ignored.

# Models are Approximations

**Discuss:** What about Hardy-Weinberg equilibrium? What are the assumptions and approximations that go into this model?

# Parameter estimation and model fit

Three common approaches have emerged for general parameter estimation:

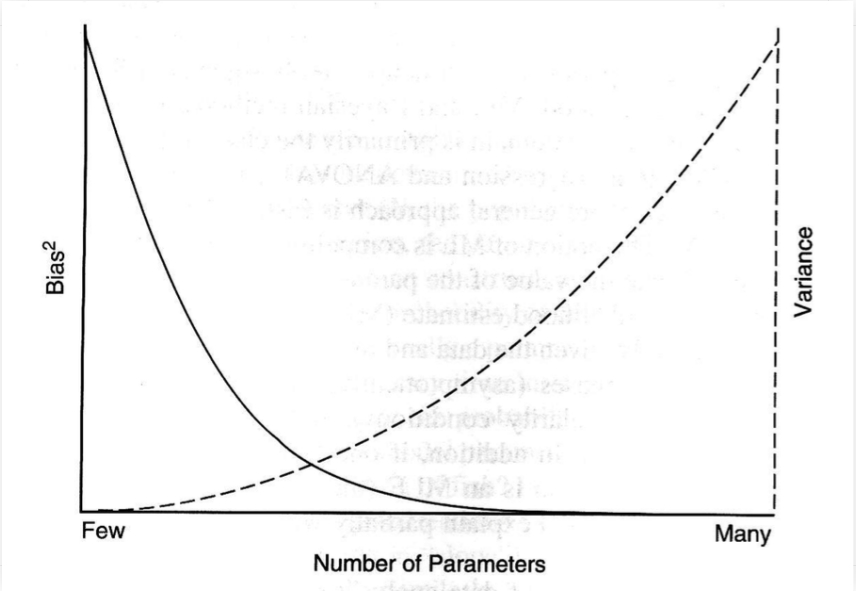
- least squares, LS (or "regression"),
- maximum likelihood, ML, and
- Bayesian methods.

**Definition:** The ***maximum likelihood estimate (MLE)*** is the value of the parameter that is most likely, given the data and model.

# The Principal of Parsimony

**Quote:** “A person new to statistical thinking often finds it difficult to relate data, model, and model parameters that must be estimated. These are hard concepts to understand and the concepts are wound into the issue of parsimony. Let the data be fixed and then realize the information in the data is also fixed, then some of this information is "expended” each time a parameter is estimated. Thus, the data will only “support” a certain number of estimates, as this limit is exceeded parameter estimates become either very uncertain (e.g., large standard errors) or reach the point where they are not estimable.“

# The Principal of Parsimony




"…too few parameters and the model will be so unrealistic as to make prediction unreliable, but too many parameters and the model will be so specific to the particular data set so to make prediction unreliable."

- Edwards

# The Principal of Parsimony

**Quote:** “Each time a parameter is estimated, some information is "taken out” of the data, leaving less information available for the estimation of still more parameters.“

**Quote:** "In model selection, we are really asking which is the best model ***for a given sample size***.”

In other words, what's the best model given the amount of information that we have?

**Quote:** “We are really asking - how much ***model structure*** will the data support?”