# introductory concepts for Bayes

env/bio 665 Bayesian inference for environmental models

Jim Clark

2020-01-08

## Contents

---

*"Probability does not exist"*, probabilist Bruno de Finetti

---

# resources

## software

- Rstudio
- R vignette on Getting started with R

```
source('../clarkFunctions2020.r')
```

## readings

**Why Big Data Could Be a Big Fail**, Jordan on potential and limitations of Big Data (misleading title).

**Why environmental scientists are becoming Bayesians**, Clark on proliferation of Bayes in environmental science, *Ecol Letters*.

**Bayesian method for hierarchical models: Are ecologists making a Faustian bargain?**, Lele and Dennis offer contrarian view, *Ecol Appl*.

**Redefining statistical significance**, classical and Bayesian statisticians compromise, if $P$ values are to used, then $P = 0.05$ is way too high, *Nature*.

## objectives

1. recognize and generate proper **notation** for a simple model
2. identify the basic elements of a Bayesian model
3. identify the **deterministic vs stochastic** variables in a model
4. interpret a **hierarchical model**
    - construct a simple **graphical model**
    - assemble **parameters, process, and data** as a hierarchy
    - describe a **regression model** with notation and graph
5. using a graphical model, explain the benefits of a **generative model**
6. articulate advantages and disadvantages of **observational and experimental evidence**
7. define **Simpson's Paradox** and indentify when it could be operating.

## assignment due 13 January

An assignment appears at the end of this vignette. The last assignment question involves a group discussion of methods in ecology. Here are a few instructions.

- Each student is responsible for written responses for each problem.
- Designate a coordinator who makes sure that everyone is involved and prepares the group summary to be presented in class.
- The coordinator turns in the group assignment with the names of team members and their contributions to it.
- Meet and work in or out of class. Agree on the mode of communication within the group e.g., email).
- Consult with me if a conflict arises that can't be worked through by the team.

All of you will have taken an introductory statistics course, but many of you will not have had exposure to a Bayesian perspective. In this vignette I introduce some of basics and connect them to some of the more familiar topics you know from an intro statistics course.

## what makes it Bayes?

In Bayesian analysis, all quantities are uncertain and assigned probability distributions. Data analysis involves i) a process, represented by a **model** that summarizes how it works, ii) evidence, which comes in the form of observations, or **data**, and additional **prior** information that needs consideration, and iii) unknowns, or **parameters**, that must be estimated by combining all of the above.

Goals of analysis can be several:

- **Estimation** to learn about parameters that quantify relationships.
- **Model evaluation** to determine whether or not the model describes the data well enough to be useful.
- **Variable selection** to evaluate how adding/subtracting varibles may improve the model.
- **Prediction** to anticipate processes or data, for model evaluation, or for understanding. Predictions can be 'in-sample' or 'out-of-sample', the latter refering to times/locations other than where data were collected. They can be functions of outputs, such as sensitivity.

In **Bayesian analysis**, all of the above are based on and expressed in the form of probability distributions.

## expansion of Bayes in environmental science

Bayesian analysis was embraced, fostered, and promoted by statisticians who wanted methods that were **axiomatic**. Unlike other branches of mathematics, statistical methods of the mid 20th century did not follow from a set of foundational axioms, i.e., accepted principles. In the absence of a unifying framework, statisticians and consumers of their methods could not agree on concepts important for interpretation, like confidence intervals and predictive intervals. This interview with David Lindley (link) includes his first-hand recollections and contributions to the Bayesian revolution that gained steam in the mid 20th century. There are many review papers and excellent books on the history of Bayesian analysis.

As Lindley points out, the early books written about Bayesian statistics had titles like "Theory of Probability", reflecting the view that these developments were really about a coherent (axiomatic) framework that could be defended on basic principles. A fundamental departure from what is now sometimes called 'conventional statistics' is the **prior distribution**, which, in turn, emphasizes a conceptual challenge with the notion of "probability" itself. de Finetti's quote at the beginning of this vignette that probability 'does not exist', relates to the concept of **subjective probability**; you and I often assign different probabilities to the same events. Beyond conceptual differences on what probability means, the early 'Bayesians' could agree

that subjective belief can be coherent if it follows the rules of probability. Coherent inference requires a prior belief that is updated by data.

Broad dissemination of Bayesian methods required computational developments, especially Markov chain Monte Carlo (MCMC), which traces its roots at least to the Manhattan Project. Inroads into statistics during the 1980's presaged the benchmark 1990 publication of Gelfand and Smith's (link) application of Gibbs sampling, highlighting its widespread potential. Robert and Casella (link) use the term 'epiphany'. Essentially one framework could admit nearly all high-dimensional problems. It contributed to the modern potential of 'big data', particularly where estimates of uncertainty are needed. In modern Bayes, modeling and computation are intimately connected.

Practitioners from many disciplines find Bayes attractive, because probability theory provides a conceptual foundation that all of us can apply to notions like **uncertainty**. The interview with Michael Jordan on big data, cited at the beginning of this vignette, makes the case for estimates of uncertainty. Jordan has made many contributions in this area, using Bayesian methods. Engineers, computer scientists, social scientists, and natural scientists may not agree on or appreciate the myriad of "tests" that make up the practice of traditional statistics. There can be a lot of confusion about what a *P value* means and whether or not it is even a valid way to interpret evidence (link). The probability interpretation of a posterior distribution is not burdened by the same confusion–the laws of probability offer a common language.

Some ecologists express concerns about use of prior distributions. Lele and Dennis worry that it could be used to 'cook' the results of a study. Many others find unconvincing the complaint that Bayes is especially susceptable to ethical transgression. If someone wanted to falsify results, it would be done through the data. Unlike data that often cannot be traced back to original observations, the prior (like the rest of the model) is transparent.

[*Still less transparent than data and models is computer code. As computation plays an increasing role in science, finding ways to evaluate analyses of high dimensional problems has become a challenge. Simply making code available may not help at a time when it is increasingly difficult to find reviewers with time to even read manuscripts, not to mention work through code.*]

Finally, there can be confusion that an informative prior facilitates cheating to "improve the fit". The best fit to the data emphasizes likelihood over prior; the best possible fit has the weakest possible prior. The role of the prior is different–it is given weight to assure a balance of information, quite the opposite of improving the fit.

Bayesian analysis has proliferated in environmental science, because many view it as simple, transparent, and coherent. Computational tools provide a common framework that accommodates models from simple to high-dimensional. This course provides an introduction.

Having said all of this, the course will engage a range of methods and welcome comparisons between them.

## hierarchies in data and processes

Environmental processes and observations can often operate and be summarized hierarchically. At a time when many ecologists were debating the differences between classical methods and simple Bayesian methods, I emphasized that the real benefit of Bayes was the capacity to build models hierarchically (link). Where process are not strictly hierarchical, a scientist may see advantages to imposing structure. **Structure in time** can include hours within days within seasons within years. Time can be structured by 'active periods' for insects or census intervals for trees or humans. **Spatial structure** can be defined by watersheds, estuaries, soil types, and land cover types. Voting districts, zip codes, or states can be used. Even where space and time are treated as continuous variables, they may be structured within discrete units.

Process models and data can refer to discrete **groups**. There can be individuals within flocks, sample plots, or political parties. Flocks or herds can be assigned to species, sample plots to blocks (**stratification**), and political parties to countries. Hierarchies can be a natural way to organize data and models.

Structured variables can be **misaligned**, such as data sets reported by zip codes versus voting precincts. Misalignment introduces challenges for hierarchical modeling.

Structure, sometimes referred to as 'scale', can not only facilitate but also complicate, interpretation. **Simpson's Paradox** and the related **ecological fallacy** are important considerations for any analysis that involves multiple scales and/or hidden variables. This issues arise when there are unmeasured variables that affect observations, which is ubiquitous in environmental problems. They can occur when attempting to fit a model at one scale and predict at another scale. Examples will arise throughout the semester. In the next section I discuss why not only variables, but also model analyses that can be hierarchical–some parameters generate a model, other parameters and model together generate data.

## hierarchical models and Bayes theorem

The expansion of modern Bayes has gone hand-in-hand with hierarchical modeling. Hierarchical models simplify the synthesis of observations with processes and parameters. [Ironically, by simplifying analysis, hierarchical models have had the opposite effect of increasing the complexity of a typical model.] They commonly exploit the Bayesian paradigm. Hierarchical models don't have to be Bayesian, but they usually are. I introduce these concepts together.

The basic building blocks are **likelihood** and **prior distribution**. Hierarchical models organize these pieces into levels or stages. These are the knowns (data and priors) and the unknowns (latent processes and parameter values). It is natural to think of inference this way:

$$[\text{unknowns} \mid \text{knowns}] = [\text{process, parameters} \mid \text{data, priors}]$$

If notation is unfamiliar: $[A]$ is the probability or density of event $A$, $[A, B]$ is the **joint probability** of $A$ *and* $B$, and $[A|B]$ is the **conditional probability** of $A$ given $B$. The Appendix summarizes notation.

This structure from Mark Berliner (example here) comes naturally to a Bayesian. It can be unwieldy to the non-Bayesian and can involve improper distributions (e.g., when omitting prior distributions). A hierarchical model typically breaks this down as

$$[\text{process, parameters} \mid \text{data, priors}] \propto$$
$$[\text{data} \mid \text{process, parameters}][\text{process} \mid \text{parameters}][\text{parameters} \mid \text{priors}]$$

The left hand side is the joint distribution of unknowns to be estimated, called the **posterior distribution**. This structure is amenable to simulation, using **Markov Chain Monte Carlo (MCMC)**. **Gibbs sampling** is a common MCMC technique, because the posterior can be factored into smaller pieces that can be dealt with one at a time, as simpler conditional distributions.

The elements of the hierarchical model relate directly to **Bayes' theorem**. To simplify, consider a joint distribution of data $D$ and parameters $P$,

$$[D, P] = [D|P][P] = [P|D][D]$$

Rearranging I obtain Bayes' theorem,

$$[P|D] = \frac{[D|P][P]}{[D]}$$

which can be read as 'the probability of parameters given data".

# regression as parameters to process to data

In this example, I introduce just the basic elements of an analysis involving regression. I use this example to make some comparisons with classical regression and to introduce a hierarchical model.

## least squares, maximum likelihood, Bayes

This is a good time to read this section of the appendix. First, I introduce a subscript $i = 1, \ldots, n$ to indicate the $n$ observations of a response $y_i$, each with corresponding predictors in a length-$p$ vector $\mathbf{x}_i$. The index $i$ is *exchangeable*, meaning that the order of observations doesn't matter. The responses make up a length-$n$ vector $\mathbf{y}$. The predictors make up a $n \times p$ matrix $\mathbf{X}$.

**the traditional view**

A traditional regression model looks like this:

$$y_i = \mu_i + \epsilon_i$$
$$\mu_i = \beta_1 + \beta_2 x_{i,2} + \cdots + \beta_p x_{i,p} = \mathbf{x}_i'\boldsymbol{\beta}$$

(again, notation is summarized in an Appendix.) There is an observed response variable $y_i$ and predictors in a length-$p$ **design vector** $\mathbf{x}_i$. Here is an example of a design vector for an intercept and two slope variables,

$$\mathbf{x}_i = (1, x_{i2}, x_{i2})'$$

Note that the first element of the design vector is $x_{i,1} = 1$, corresponding to the intercept for the model. The error has an expection $E[\epsilon] = 0$ and variance $Var[\epsilon] = \sigma^2$. This is called a **linear model**, because it is a linear function of the parameters $\boldsymbol{\beta}$.

Typically, I want to estimate the parameters in the model. To do this, I have to decide on a criterion that constitutes a 'good fit'. The simplest assumption could be to find parameter values that minimize the leftover ('residual') variation. Without specifying a distribution for $\epsilon$, I could minimize the expected value of the squared error,

$$mse = \min_{\boldsymbol{\beta}} E[\epsilon^2] = \min_{\boldsymbol{\beta}} E\left[\sum_{i=1}^{n}(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2\right]$$

with respect to parameters $\boldsymbol{\beta}$. Within the brackets I am summing the squared difference between the observed $y_i$ and the value that would be predicted by the model, $\mathbf{x}_i'\boldsymbol{\beta}$. By taking derivatives with respect to $\boldsymbol{\beta}$, setting them equal to zero, and solving for $\boldsymbol{\beta}$, I obtain the **least-squares estimates** $\hat{\boldsymbol{\beta}}$. This is the **method of least-squares**. So far, there is no mention of probability or uncertainty.

If I specify a normal distribution for errors, $\epsilon \sim N(0, \sigma^2)$, I can write a **likelihood function** for the observations, where $N$ indicates a normal distribution, with mean zero and variance $\sigma^2$. Due to special properties of the normal distribution, if $\epsilon \sim N(0, \sigma^2)$ and $y = \mu + \epsilon$, then $y \sim N(\mu, \sigma^2)$, where parameters are $\mu$ (mean) and $\sigma^2$ (variance).

Because the regression model assumes independent observations, I can write the likelihood function like this:

$$L(\boldsymbol{\beta}, \sigma^2) = \prod_{i=1}^{n} N(y_i|\mathbf{x}_i'\boldsymbol{\beta}, \sigma^2)$$

The independence assumption means that the probability of all $n$ observations is equal to the product of the probilities of each individual observation,

$$\prod_{i=1}^{n} N(y_i|\mathbf{x}_i'\boldsymbol{\beta}, \sigma^2) = N(y_1|\mathbf{x}_1'\boldsymbol{\beta}, \sigma^2) \times \cdots \times N(y_n|\mathbf{x}_n'\boldsymbol{\beta}, \sigma^2)$$

The **maximum likelihood estimate** comes from maximizing the likelihood function with respect to parameters. I obtain the same estimates as I did with least squares. This is true *if $\epsilon$ is normally distributed*. For least squares and this maximum likelihood approach, the likelihood is treated as a function of the parameters.

---

Exercise 1. write out the multiplication for $\mathbf{x}'\boldsymbol{\beta}$.

---

### a Bayesian view

A Bayesian regression model is written a bit differently:

$$[\boldsymbol{\beta}, \sigma^2|\mathbf{y}, \mathbf{X}] \propto [\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}, \sigma^2][\boldsymbol{\beta}, \sigma^2]$$

Inserting the likelihood on the right-hand side gives

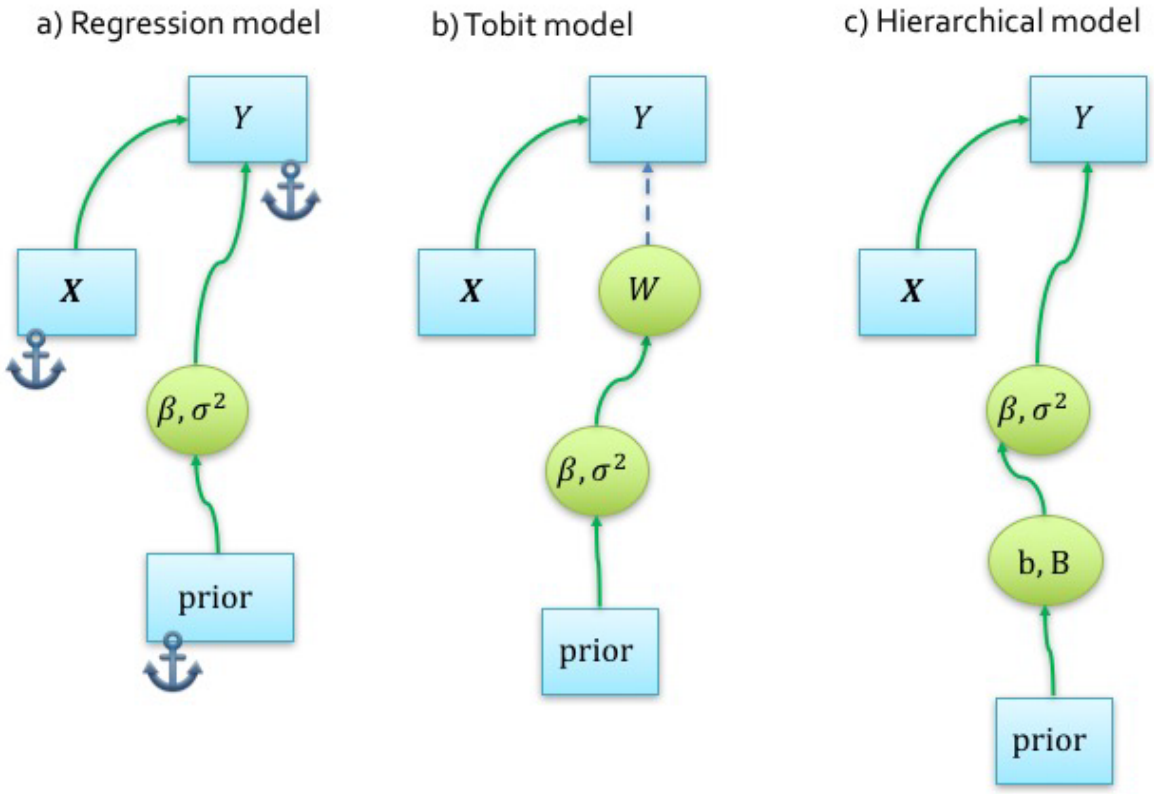
$$\prod_{i=1}^{n} N(y_i|\mathbf{x}_i'\boldsymbol{\beta}, \sigma^2)[\boldsymbol{\beta}, \sigma^2]$$

The left-hand side is "parameters given data", or the **posterior distribution**. The likelihood is unchanged. However, in Bayesian analysis the likelihood is treated not as a function, but rather as a distribution. Rather than optimization, I will engage distribution theory. To highlight this difference, in the first line of this equation, I write it as a generic distribution, $[\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}, \sigma^2]$, the observed response conditional on the observed predictors in $\mathbf{X}$ and the parameters. The second factor on the right indicates an (unspecified) prior distribution for parameters, using the bracket notation for a distribution. The **prior distribution** allows me to introduce information beyond the observations, and to evaluate a **posterior distribution**. Technical details follow later, for now I focus on general concepts.

There are similarities and differences in these approaches. If the prior distribution is flat, the posterior distribution for this regression model will have the same marginal means and nearly identical standard errors as I obtain by traditional methods. The impact of the prior distribution becames important when it is made informative and as models increase in size.

## model graph

Models can be represented as graphs. Below are three graphs, each describing a different model. The Bayesian regression model is shown in (a).




*Three regression variations in graphical form. Solid arrows are stochastic. Circles are unknown (estimated). Symbols in (a) emphasize variables that 'anchor' the fit–observed variables and prior parameter values are known, whereas others are estimated. a) Unknown parameters and known predictors $X$ generate response $Y$. b) The Tobit model could be viewed as hierachical or not. $Y$ is a deterministic function of $W$ (see text). c) The hierarchical model has an additional stochastic stage.*

In this diagram, there are boxes, circles, dashed lines, and solid lines. The graph provides a visual interpretation of the model. I discuss symbolism in the next sections.

**known and unknown variables**

Varibles are nodes in the graph. The known variables are boxes, including data and prior parameter values. They anchor the model fit, because they are constant. The unknown variables and parameters are circles. In the regression example, these are the regression coefficients, the length-$p$ vector $\boldsymbol{\beta}$, and the residual variance $\sigma^2$.

#### stochastic and deterministic variables

Both the response $y_i$ and the predictors $\mathbf{x}_i$ are known. However, the predictors are treated as deterministic, whereas the response is treated as stochastic before it is observed. Specifically, $y_i$ is assigned a distribution, the likelihood. The vector $\mathbf{x}_i$ is not assigned a distribution. In observational studies it is not always clear what should be a predictor and what should be a response.

#### stochastic and deterministic arrows

Solid arrows in the graph indicate stochastic relationships, meaning that knowledge of a parent node does not tell us the value at the child node (or vice versa). Deterministic relationships can often be omitted from the graph. One deterministic relationship is shown in part (b) as a dashed line. This example will be discussed below.

## an application with R: tree abundance and climate

I use an application to illustrate some differences between a traditional and Bayesian analysis. Here are some R objects I use in this example:

**R packages**: `gjam`, `repmis`

**R objects**: `list`, `data.frame`,`numeric`, `numeric vector`, `numeric matrix`, `source_data`

**R functions**: `as.formula`, `attr`, `cbind`, `lm`, `names`, `par`, `plot`, `points`, `predict`, `summary`

To learn about a `function`, I use the help page:

```
help(lm)
```

In addition to these objects, I introduce the connection between a variables subscript and the location in an `array`, which can be a `vector` or a `matrix`.

A traditional regression looks like the graph in part (a), but it does not include the prior distribution for parameters. Here is a linear regression fitted to the abundance of white oak (*Quercus alba*) in the USDA Forest Inventory and Analysis (FIA) program data. I use a file sourced from the internet, the github website where you can deposit and share data and code.

#### load a package

R software includes built-in functions and packages. A package is a bundle of functions that together offer tools to process a related set of problems. To use a function that is contained in a package, I need to install the package. Here is the installation of two packages:

```
install.packages('repmis')  # read from github
install.packages('gjam')    # extract file
```

You can use a function without loading the entire package using the syntax `package::function`.

### load a remote data set

In the following code I first set a variable `biomass` to the the column for *Quercus alba*. I get the data from a remote repository using the function `source_data(fileName)` in the package `repmis`:

```
d <- "https://github.com/jimclarkatduke/gjam/blob/master/forestTraits.RData?raw=True"
repmis::source_data(d)
```

```
## [1] "forestTraits"
```

```
data    <- forestTraits$xdata                              # compressed
tmp     <- gjam::gjamReZero(forestTraits$treesDeZero) # decompress
biomass <- tmp[,"querAlba"]                                # response
data    <- cbind(biomass, data)                            # all variables
```

The function `gjamReZero` is a function in the packages `gjam` that decompresses a file containing mostly zeros.

### plotting data

The `data.frame data` includes both continuous variables and factors. It is formatted as observations (rows) by variables (columns). I discuss this OxV format in unit 2.

First I plot some variables in `data` using the R function `plot`:

```
par(mfrow=c(1,2),bty='n')
plot(data$temp, data$deficit, xlab='Winter temperature',
     ylab='Moisture deficit', cex=.2)
plot(data$temp, data$moisture, xlab='Winter temperature',
     ylab='Site moisture', cex=.2)
```

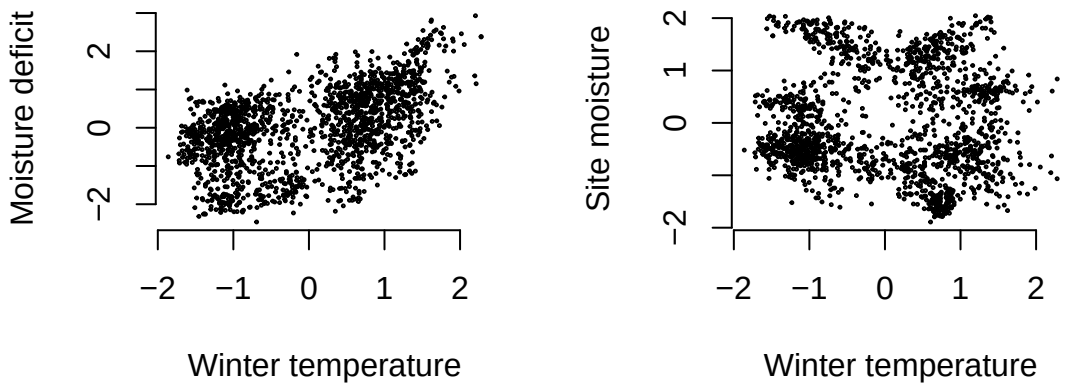



Figure 1: *Some variables in the data.frame data*

The syntax `data$temp` indicates that `data` is a type of R object termed a `list`. In this instance, `data` is a specific type of `list` called a `data.frame`. The `$temp` part of this name means that this `list data` has an object with the `name temp`. Use the `help(plot)` page to interpret the arguments to the function `plot`.

### variables: from `data.frame` to model `lm`

I want a model for the response `biomass` with predictors. Here are the names of variables in `data`:

```
names(data)
```

```
## [1] "biomass"  "temp"     "deficit"  "moisture" "u1"       "u2"       "u3"
## [8] "stdage"   "soil"
```

These are the column headings in `data`. To see this, I can look at the first two rows:

```
 data[1:2,]
```

```
##   biomass      temp    deficit  moisture         u1          u2         u3
## 1       4 1.2165433 0.03637914 0.6870299 0.01182693 0.003898011 0.04438465
## 2       1 0.1825447 0.20708706 1.6655992 0.02679904 0.000000000 0.87181296
##        stdage      soil
## 1 -0.16697961 reference
## 2 -0.02907271 reference
```

With the exception of `"soil"`, these are continuous variables. The variable `soil` is a `factor` with these `levels`:

```
attr(data$soil,'levels')
```

```
## [1] "EntVert"   "Mol"       "reference" "SpodHist"  "UltKan"
```

These are soil 'orders', based on parent material.

I start with the standard function `lm` in R for linear regression. In the `formula` below I have specified a model containing `moisture` or `deficit` and with quadratic effects, as `I(moisture^2)`. The `I()` function in a `formula` indicates that I want to evaluate the argument as it appears in the `function I`.

High correlation in predictor variables will mean that I cannot descriminate their contributions. These variables are ok. A call to `pairs(data, cex=.1)` will show all combinations of variables in `data`.

I can fit a linear regression with reponse `biomass` to several combinations of variables using `lm` and plot responses.

```
par(mfrow=c(1,2),bty='n', mar=c(4,4,1,1))
plot( data$moisture, biomass, cex=.2)                # wet sites
```

```
fit1 <- lm(biomass ~ moisture, data)
p1   <- predict(fit1)
fit2 <- lm(biomass ~ moisture + I(moisture^2), data) # quadratic
p2   <- predict(fit2)                                # predictive mean
points(data$moisture,p1, col='green', cex=.2)        # add to plot
points(data$moisture,p2, col='blue', cex=.2)

#repeat with additional variables
plot( data$deficit, biomass, cex=.2)
form <- as.formula(biomass ~ soil + moisture*stdage + temp + deficit)
fit3 <- lm(form, data)
p3   <- predict(fit3)
points(data$deficit, p3, col='green', cex=.2)
```

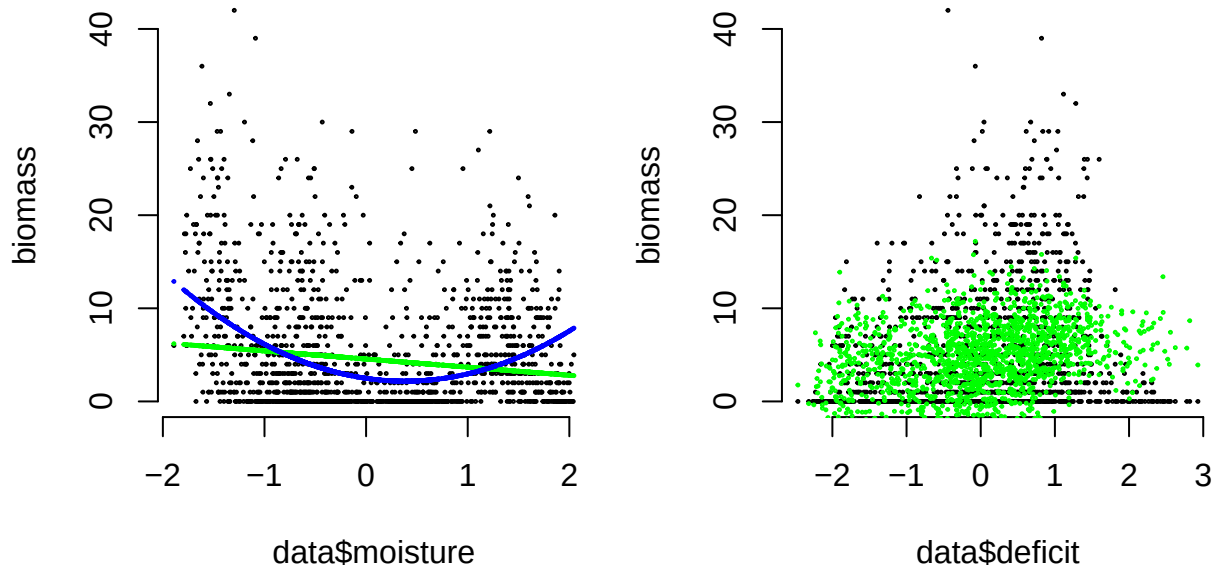



Figure 2: *Biomass fitted to moisture and climate deficit.*

For more on use of variables names in the above block of code, see the note about R environments.

Plots show the predictions for a linear and quadratic model, `fit1` and `fit2` and for a large model with main effects, an interaction `moisture*stdage` and the `factor soil`.

### a model summary

I can use the function `summary` to look at the estimates in a fitted objects, e.g.,

```
summary(fit3)
```

```
##
## Call:
## lm(formula = form, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.3953  -3.2229  -0.6428   2.3601  30.2392
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       2.4527     0.4952   4.953 8.07e-07 ***
## soilMol           1.6367     0.9588   1.707   0.0880 .
## soilreference     3.3204     0.5185   6.404 1.98e-10 ***
## soilSpodHist     -1.2109     0.5888  -2.057   0.0399 *
## soilUltKan        4.6979     0.8526   5.510 4.17e-08 ***
## moisture         -0.9775     0.1275  -7.669 2.98e-14 ***
## stdage            2.5918     0.1309  19.800  < 2e-16 ***
## temp              1.1360     0.1623   7.000 3.75e-12 ***
## deficit           0.7610     0.1503   5.062 4.62e-07 ***
## moisture:stdage  -0.6469     0.1374  -4.709 2.70e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.07 on 1607 degrees of freedom
## Multiple R-squared:  0.3613, Adjusted R-squared:  0.3578
## F-statistic:   101 on 9 and 1607 DF,  p-value: < 2.2e-16
```

Standard products of a traditional regression include a table of **point estimates** for the coefficients in $\boldsymbol{\beta}$, with their **standard errors**, which are a measure of uncertainty. There is a **Student's t statistic** for each estimate, a measure of how far the estimate is from a hypothesized value of zero. For each estimate there is a **P value**, the probability for obtaining a value of t at least this large under the hypothesis that the coefficient is equal to zero. There is a **F statistic** and P value for the entire model. The **degrees of freedom** is the sample size $n$ minus the number of fitted parameters (2). The **R-squared** is interpreted as the 'variance explained' by the model. The **residual standard error** is the estimate of parameter $\sigma = \sqrt{\sigma^2}$. The **95% confidence intervals** for these estimates would be approximately the point estimates $\pm 1.96$ times the standard errors, or:

```
##                  estimate   0.025   0.975
## (Intercept)       2.4530  1.4850  3.4210
## soilMol           1.6370 -0.2378  3.5110
## soilreference     3.3200  2.3070  4.3340
## soilSpodHist     -1.2110 -2.3620 -0.0598
## soilUltKan        4.6980  3.0310  6.3650
## moisture         -0.9775 -1.2270 -0.7283
## stdage            2.5920  2.3360  2.8480
```

```
## temp              1.1360  0.8187  1.4530
## deficit           0.7610  0.4671  1.0550
## moisture:stdage  -0.6469 -0.9154 -0.3783
```

Together, this combination of outputs for the linear regression would contribute to an interpretation of how abundance of this species responds to climate and habitat.

---

Exercise 2. repeat this analysis using function `lm` with a different species as a response. Interpret the analysis**

---

## what's different about Bayes?

Leaving technical detail for subsequent units, this section compares the previous result with a Bayesian analysis, incorporating a non-informative prior distribution. As mentioned above, a Bayesian analysis combines the likelihood with a prior distribution. In this analysis, the prior distribution is taken to be non-informative for coefficients and residual variance $\boldsymbol{\beta}, \sigma$. Here is a Bayesian analysis to compare with `fit3`:

```
fitb <- bayesReg(form, data)
```

```
##
## Coefficients:
##                  median std error   0.025    0.975 not zero
## intercept         2.473    0.4856   1.515    3.443        *
## soilMol           1.602    0.9632 -0.2426     3.46
## soilreference      3.31    0.5073   2.301    4.302        *
## soilSpodHist     -1.222    0.5851  -2.361 -0.06184        *
## soilUltKan        4.691    0.8492   3.066    6.369        *
## moisture        -0.9819    0.1288  -1.227  -0.7333        *
## stdage            2.591    0.1307   2.336     2.85        *
## temp              1.139    0.1604  0.8276    1.456        *
## deficit          0.7601    0.1506  0.4745    1.063        *
## moisture:stdage -0.6462    0.1361 -0.9165  -0.3774        *
##
##  * indicates that 95% predictive interval does not include zero
##
## Residual standard error 5.075, with 1607 degrees of freedom,
##  root mean sq prediction error 4.968.
```

The `function BayesReg` organizes output a bit differently from `lm`, reporting 95% credible intervals [values at (0.025 and 0.975)]. The output is simpler than that generated by `lm`–there are not a lot of statistics. However, I see that point estimates and standard errors for coefficients are nearly the same as I obtained with `lm`. I also see that the coefficients having

credible intervals that do not span zero in the Bayesian analysis are the same coefficients that `lm` flagged as 'signficant'.

The header for this section asks 'what's different about Bayes?'. The shift from a classical to Bayesian analysis did not change how I interpret effects of climate and habitat on biomass of white oak. However, the Bayesian model can be extended in a number of ways. In the next section I illustrate a hierarchical version having an additional stage.

## what about zero?

At least one problem with this analysis is that regression does not allow zeros. The likelihood for the regression model defines a normal density on the real line, $(-\infty, \infty)$. Recall that the probability of any value on the real line is zero. On the other hand the probability for any interval on the real line must be greater than zero. Observed values of $Y$ are recorded to a degree of precision defined by data collection. Although they are discrete, we model them as continuous variables. Continous variables could exist to an arbitrarly degree of precision. So discrete values, like zero, are not permitted. Here is a histogram of the data and the fraction of zeros:

```
hist(biomass, nclass=50)                                    # distribution
```

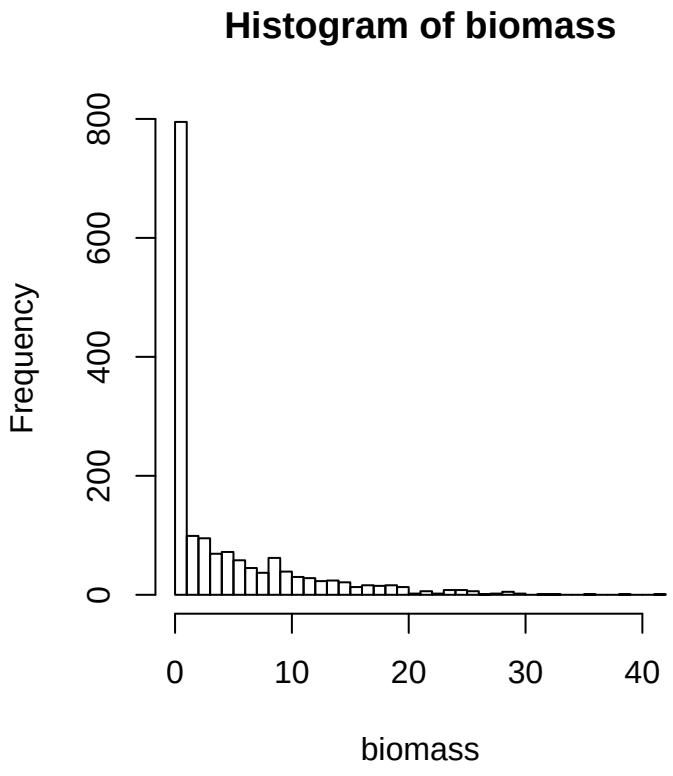



Figure 3: *Response has many zeros.*

```
length(which(biomass == 0))/length(biomass)
```

```
## [1] 0.3896104
```

Biomass data diverge from the assumptions of the model in that observations $Y \in [0, \infty)$– there is **point mass** at zero. Positive biomass values are continous, but zero is discrete. In fact, many of the observed values are zero–see the histgram of the data above. If there were a few zeros, with most values bounded away from zero I might argue that it's close enough. That's not the case here. Standard diagnostic plots, such as `plot(fit1)` make this clear.

If these were discrete data, I might turn to a zero-inflated Poisson or negative binomial model. These are examples of generalized linear models (GLMs) to be discussed in later units. These models sometimes work ok, provided there are not too many zeros, but here zeros dominate. Regardless, these GLMs describe discrete data, so I cannot use either.

I cannot add a small amount to each observation and then transform the data to a log scale. If I do that, every coefficient I estimate depends the arbitrary value I used.

## Tobit regression

Part (b) of the graph includes an additional stage for a variable $W$. This variable has a normal distribution; it is defined on $(-\infty, \infty)$. The observed $Y$ is a censored version of $W$. It is equal to the response $Y$ whenever $Y > 0$. When $Y = 0$, then the latent variable $W$ is negative:

$$y_i = \begin{cases} w_i & w_i > 0 \\ 0 & w_i \leq 0 \end{cases}$$

Treating $Y$ as a censored version of $W$ allows me to combine continuous and censored data without changing the scale. In a Tobit model the censored value is zero, but it also works with other values.

With this model the regression moves to the latent $W$,

$$w_i \sim N(\mathbf{x}_i'\boldsymbol{\beta}, \sigma^2)$$

The model has become hierarchical. I fit this model in a Bayesian setting, again, requiring a prior distribution.

Now allowing for the zeros in $Y$, the fitted coefficients differ substantially from both previous models:

```
fitTobit <- bayesReg(form, data, TOBIT=T)
```

```
## fitted as Tobit model
```

```
##
## Coefficients:
##                  median std error  0.025   0.975 not zero
## intercept       -0.5852    0.7215 -2.028  0.7788
## soilMol          2.755     1.346 0.1406   5.417        *
## soilreference    4.817     0.756  3.353   6.296        *
## soilSpodHist    -6.071    0.9476 -7.896  -4.185        *
## soilUltKan       6.897     1.189  4.521    9.22        *
## moisture        -0.9123      0.18 -1.258 -0.5535        *
## stdage           4.261     0.211  3.857   4.665        *
## temp             2.327    0.2518  1.829    2.81        *
## deficit          1.019    0.2203 0.6066   1.446        *
## moisture:stdage -0.9487    0.2057 -1.359 -0.5515        *
##
##  * indicates that 95% predictive interval does not include zero
##
## Residual standard error 6.705, with 1607 degrees of freedom,
##  root mean sq prediction error 4.838.
```

```
par(mfrow=c(1,2),bty='n')
plot(biomass, p3, cex=.2, ylab='Predicted values')
points(biomass,fitTobit$predictY[,2], col=2, cex=.2)
abline(0,1,lty=2)
abline(h=0)
plot(summary(fit3)$coefficients[,1],fitTobit$coeff[,1],
     xlab='Linear regression',ylab='Tobit')
abline(0,1,lty=2)
```

---

Exercise 3. Using a different species and predictors, interpret differences between estimates for the Tobit and standard regression model.

---

I called the Tobit model 'hierarchical', but some could see it differently. The $W$ stage in the model is 'partially known'. Note the dashed line in part (c) of the graphical model. If I know the value of $W$, then I also know the value of $Y$. So in the 'generative direction' from $W$ to $Y$ I can view their relationship as deterministic. However, given $Y$ does not necessarily mean that I know $W$. If $Y = 0$, then $W$ is stochastic.

In summary, not only is the Tobit defensible as a valid model for continuous data with zeros–it also finds more effects and better predicts data than the traditional regression. Many value the hierarchical framework for its flexibility to admit additional stages.

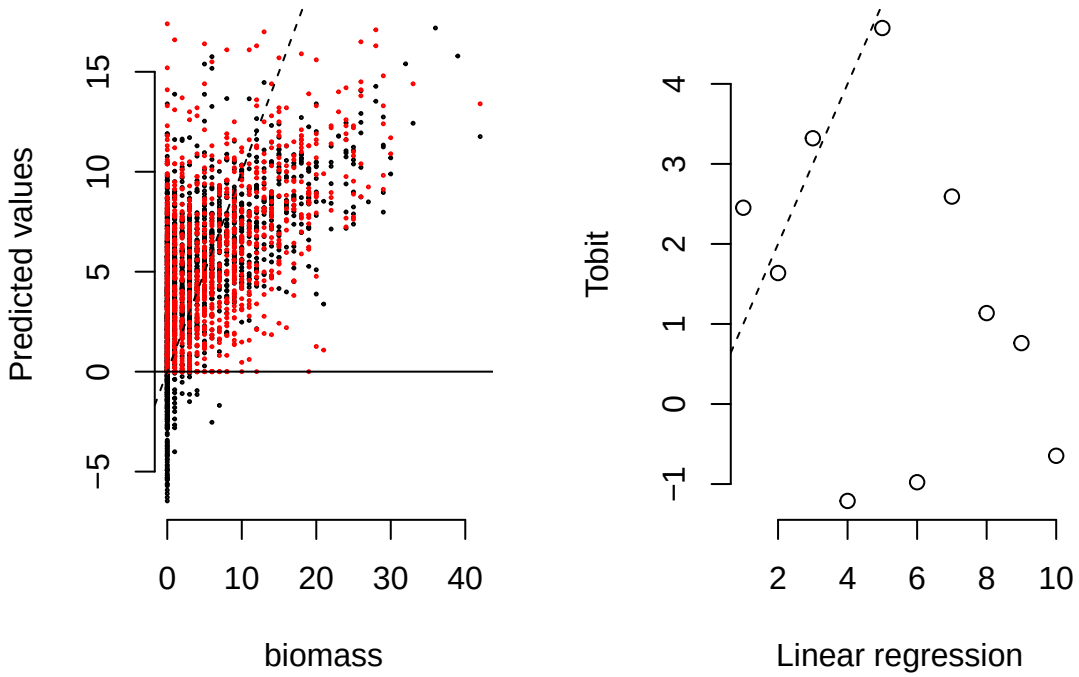



Figure 4: *Prediction from linear regression (black) and the Tobit model (red). Mean parameter estimates at right show large differences.*

# background concepts

## some ambiguous terms

The terms "process model" and "statistical model" can represent a false dichotomy (Berliner 2003). Hierarchical models typically include both elements. They are used for inference and to predict processes and data, with uncertainty. To take a simple recent example, development of spring phenological state could be written as a rate equation,

$$\frac{dh}{dt} = r(t; \mathbf{x}(t))\,(1 - h(t))$$

for development rate $r(t)$ that changes over time, in part due to environmental changes contained in $\mathbf{x}(t)$. Most ecologists would call this a 'process model'. However, when discretized, written in terms of predictors, and combined with a data model for ordinal observations, it's used for inference (Clark et a. 2014). Hierarchical modeling has made the distinction between statistical and process a false dichotomy.

## design

The common elements of a study design include **replication, stratification, and randomization** (RSR). The design of a study contributes to the information that can be gleaned

from it. Information can be evaluated in many ways, such as the precision a model provides on parameter estimates or predictions. There is a large literature on 'optimal design' and how it benefits from RSR. Design principles are implicit or explicit in many analyses discussed this semester.

## observational and experimental evidence

The two common sources of data include **experimental**, where the design is subject to some control, and **observational**, where variables are uncontrolled only to the extent possible from sample deployment. Experiments are deployed according to design considerations. The model for experimental data may be a model for the design (e.g., ANOVA), but it doesn't have to be. An experiment may be established to evaluate many processes, each analyzed with a different model.

Observational data have an advantage in being available at scales that could be not be studied with controlled experiments. Large observational networks, some global, are examples. Weather station data, now combined with satellite imagery, are the basis for modern understanding of ocean and atmospheric circulation. Epidemiological evidence for public health trends are widely used in the social sciences, as are polling data in public policy. Disease surveillance can be active or passive, the latter being easier to get, but harder to use.

The big concern with observational data is how to attribute cause and effect. As many variables change, interacting with one another as well as unmeasured variables, there may be no way to identify what's affecting what. Are people healthy because they exercise, or does the state of being healthly motivate exercise?

It can be hard to even decide what goes on the right and the left side of the model. Does high nitrogen mineralization rate promote N-demanding plants and microbes, or vice versa? What do I do when I know that both are true?

Experiments benefit from manipulation and control. Treatment effects might be isolated when other variables are controlled. Control is critical to inference in many settings.

Experimental results can be difficult to extrapolate outside the experimental frame, which could be the laboratory, growth chamber, greenhouse, or small field plot. Only a fraction of important questions are amenable to experimentation. Many ecologists and social scientists look for 'natural experiments'.

## data/prior/process balance

Every analysis strikes its own balance between data and model. In many cases data may dominate the interpretation. Many arguments about climate change focus on global mean temperatures, obtained by averaging station data that go back over a century. This seems to be a data-driven debate. ("what do the data say?")–until we confront the fact that the mean

of the stations is not the global mean. Models are used to translate unevenly distributed data to a less biased estimate of regional climate. The PRISM data are an example.

## ecological fallacy and Simpson's paradox

It's not hard to find examples where individual and group behavior appear to show different relationships. In recent elections, we now know that rich states voted democratic, while rich people voted republican (Gelman 2009). In the last few decades, the mean household income has increased, while the income of the mean household has declined. Ecological communities (species distributions and abundances) are poorly predicted by aggregating predictions for individual species (Clark et al. 2011). These are examples of 'Simpson's Paradox'.

Conversely, models of total numbers of species do not predict extinction threat, which operates at the scale of individual species, an example of the 'ecological fallacy'. The ecological fallacy refers to the fact that group behavior does not predict individuals within groups (rich people do not vote like rich states). Gerrymandered voting precincts exploit this fact.

## generative models go forward and backward: inference - prediction - simulation

A **generative model** is one that predicts the data that were used to fit the model. This means that a model works forward and backward. Take as an example a model $[Y|X,\theta]$. Prediction and simulation is the **forward** direction: start with parameters $\theta$ and predictors $X$ and generate data $Y$. Model fitting is the **inverse** direction, from data to parameter estimates, $[\theta|X,Y]$. In recent years I have been advocating inverse prediction of predictors, $[X|Y,\theta]$, as a basis for evaluating the importance of predictor variables. Generative models are needed if the model is to be evaluated with prediction–can model fitting recover the parameters used to simulate the data.

Many models are not generative. When they are not, evaluation is difficult. In this course prediction plays a large role in model checking.

## taxonomy of classical/Bayesian/algorithmic

There are many ways to learn from evidence. I have already mentioned that Bayes has attracted many scientists who appreciate simplicity, transparency, and an axiomatic foundation. The simplicity I speak of here refers to the fact that the posterior distribution stands in for a host of different 'tests' used in classical statistics. I also pointed out that adding a non-informative prior to a simple model may lead to an interpretation that is not much different from a classical approach. In such cases, a classical statistician can justifiably claim to hold the advantage of simplicity.

The simplicity advantage of Bayes increases with the size of the model. Due to the large number of variables and, potentially, types of data, hierarchical Bayes can provide substantial advantages in environmental science.

The big data era has emphasized the value of more algorithmic approaches. I don't want to run a Gibbs sampler to search the internet. The list of recommended sites is already more than I can use; I don't want to pile on credible intervals or predictive intervals. I'm happy to have my cell phone tell me that traffic is stalled ahead, without the clutter of uncertainty estimates. Approximate answers that are 'close enough'. They mean more to me than generative modeling. Machine learning is extremely important, but often does not offer a model, parameter estimates, or uncertainty. For question-driven problems, scientists typically want a model, not an algorithm. There is a place for both approaches.

Each of these methods can coexist, but there can be clear preferences, depending on the application.

## recap

Am I a Bayesian? Many who make use of Bayesian models and software would not necessarily see themselves as particularly Bayesian. Models and software may be selected for convenience or other considerations that have little to do with philosophy. Avowed non-Bayesians may exploit MCMC methods, while making a point of omitting priors. Staunch Bayesians may sometimes omit prior where they do not affect results or interpretation, but include them elsewhere. Users of packages like INLA may not know that it uses prior distributions (it's Bayesian). Many big data applications bridge machine-learning and statistical worlds, exploiting approximations of many types. Bayes has expanded rapidly in large part for pragmatic reasons of simplicity and capacity to address large problems.

Bayesian analysis combines a prior distribution with a data model, or likelihood, to arrive at a posterior distribution of unknown parameters and latent variables. Deterministic variables are constant, whereas stochastic variables are defined by distributions. Known variables are constant, but they can be stochastic and, thus, can be defined by distributions, such as the response $Y$ in the regression model. Graphs are used to communicate model structure and organize algorithms. They provide a visual representation of hierarchical models. A common hierarhical structure is [data|process, parameters][process|parameters][parameters].

Basic regression models provide an opportunity to introduce similarities and differences. Least-squares, maximum likelihood, and Bayes share some common features at this simplistic level, but diverge as models increase in size.

Both experimental and observational data are analyzed in environmental sciences. They differ in design, control, and, often, scale. Fundamental design considerations include randomization, stratification, and replication.

Model fitting and prediction are 'backward' and 'forward' views. Generative models do both and, thus, can be fully evaluated (in both directions).

# assignment

1 - 3. Complete exercises 1-3.

4. In a Bayesian model, identify the marginal, joint, and conditional distributions.
    - What is random in the posterior distribution?
    - What is random in the likelihood?
5. What are the three elements of a design? What does each contribute?
6. *In groups:* A recent paper by (Carrara et al) compares four methods for estimating the strength of interactions between species. For each of these methods summarize the following:

- How is uncertainty in data handled? For example, is there a likelihood? If so, what is it? If not, how do you think uncertainty affects estimates and their interpretation?
- Are observations independent? If not, how does this affect the estimates and their interpretation?
- Are there fixed and random parameters? How might the differences affect the model.
- What is the role of computation in each model? Do Jordan's concerns apply to any of these methods?

# appendix

## note about R environments

To fully understand the block of code for the `lm` fit, I need to know that the variable `biomass` is defined in my global or "working" environment (see previous code block). I can enter `length(biomass)` and get an answer, because `biomass` exists in the global environment. The variables `deficit` and `moisture` have not been assigned in my global environment. They only exist within the `data.frame data`. When I call the function `lm` it knows to look for variables in my global environment or in `data`, because I have passed `data` as an argument. The functions `plot` and `points` do not look for variables in this way. When I call them, I must specify the `data.frame` with the variable name, `data$deficit`. Using R is the subject of Unit 2.

## notation

Here are some notation conventions used in these vignettes.

| notation | example | interpretation |
|---|---|---|
| italic | $x$, $X$ | scalar quantity, known |
| greek | $\alpha, \beta, \ldots$ | stochastic (fitted) variable, unknown |
| parentheses | $\phi(\mu, \sigma^2)$, $N(\mu, \sigma^2)$ | parametric function/density |
| curly brackets | $\{0, 1, \ldots\}$ | a set on objects |
| closed interval | $(-\infty, 0]$, $[0, \infty)$ | intervals include zero |
| open interval | $(-\infty, 0)$, $(0, \infty)$ | exclude zero |
| distributed as | $x \sim N(\mu, \sigma^2)$ | distribution or density |
| expectation | $E[\epsilon] = 0$ | expected value of a variable |
| variance | $Var[\epsilon] = \sigma^2$ | variance of a variable |
| bracket, distribution | $[A, B, C]$ | unspecified density or distribution |
| proportional | $f(x) \propto g(x)$ | differ by a constant, $f(x) = cg(x)$ |
| approximate | $\pi \approx 3.14159$ | approximately equal |
| real numbers | $\mathbb{R} = (-\infty, \infty)$ | note: also positive real $\mathbb{R}_+$ |
| is an element of | $\pi \in \mathbb{R}$ | $\pi$ is a real number |
| subset | $\{a\}$ and $\{a, b\} \subseteq \{a, b\}$ | in the set |
| proper subset | $\{a\}$, but not $\{a, b\} \subset \{a, b\}$ | cannot include all elements |
| union | $a \cup b$ | either $a$ or $b$ |
| intersection | $a \cap b$ | both $a$ and $b$ |
| sum | $\sum_{i=1}^{n} x_i$ | $x_1 + \cdots + x_n$ |
| product | $\prod_{i=1}^{n} x_i$ | $x_1 \times \cdots \times x_n$ |
| exponentiate | $e^x$, $exp(x)$ | $e \approx 2.71828$, inverse of $ln(x)$ |
| natural logarithm | $ln(x)$ | inverse of $e^x$ |

## matrices

| notation | example | interpretation |
|---|---|---|
| bold, l.c. | $\mathbf{x}$, $\mathbf{x}'$ | column and row vectors, respectively |
| bold, u.c. | $\mathbf{X}$ | matrix |
| dimensions | $\underset{n \times q}{\mathbf{X}}$ | $n$ rows, $q$ columns |
| subscript | $\mathbf{x}_i$, $\mathbf{X}_{ij}$ | element of an array |
| row vector | $\mathbf{X}_{i.}$ | row $i$ |
| column vector | $\mathbf{X}_{.j}$ | column $j$ |
| transpose | $\mathbf{X}'$ | rows become columns |
| matrix inverse | $\mathbf{A}^{-1}$ | solve systems of linear equations |
| identity matrix | $\mathbf{I}_p = \mathbf{A}\mathbf{A}^{-1}$ | $p \times p$ with 1's on the diagonal, zeros elsewhere |
| matrix determinant | $det\,(\mathbf{A})$ | obtain for a square matrix, e.g., covariance |
| kronecker product | $\underset{n \times m}{\mathbf{A}} \otimes \underset{p \times q}{\mathbf{B}}$ | $np \times mq$ matrix, defined in text |