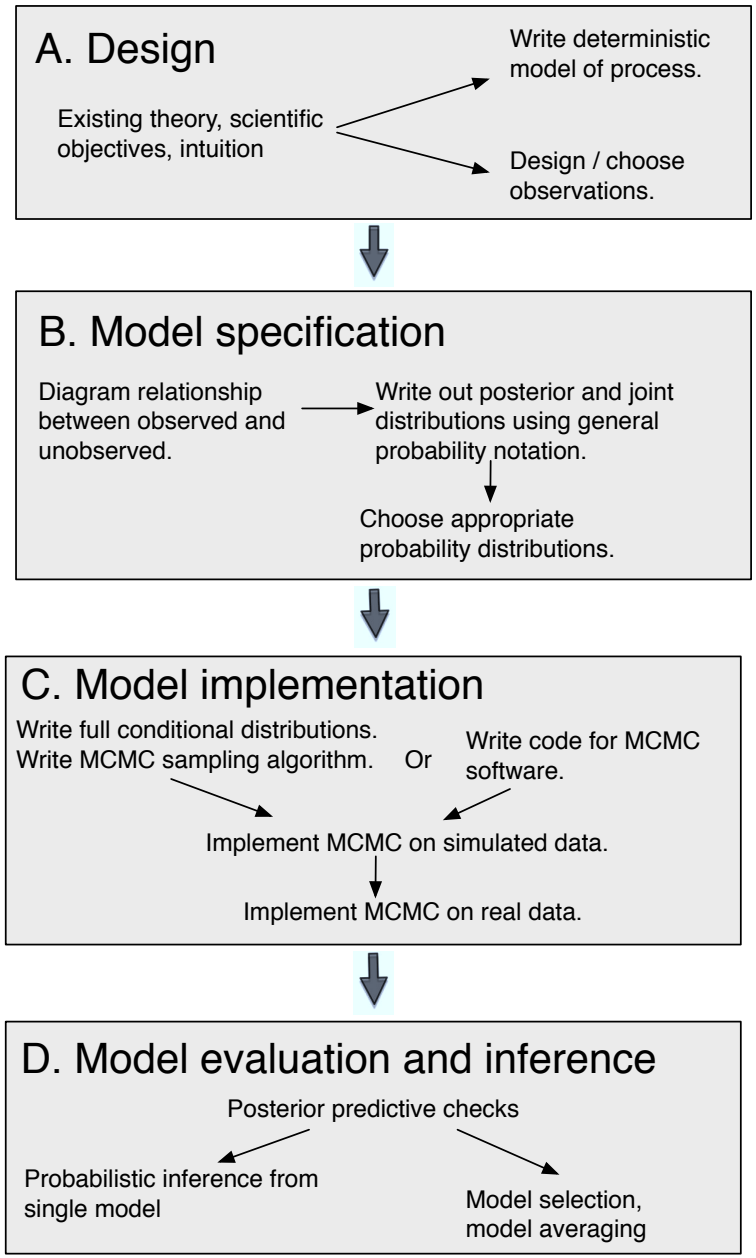
Bayesian Regression

ESS 575 Models for Ecological Data

N. Thompson Hobbs

March 21 2017

Where are we?

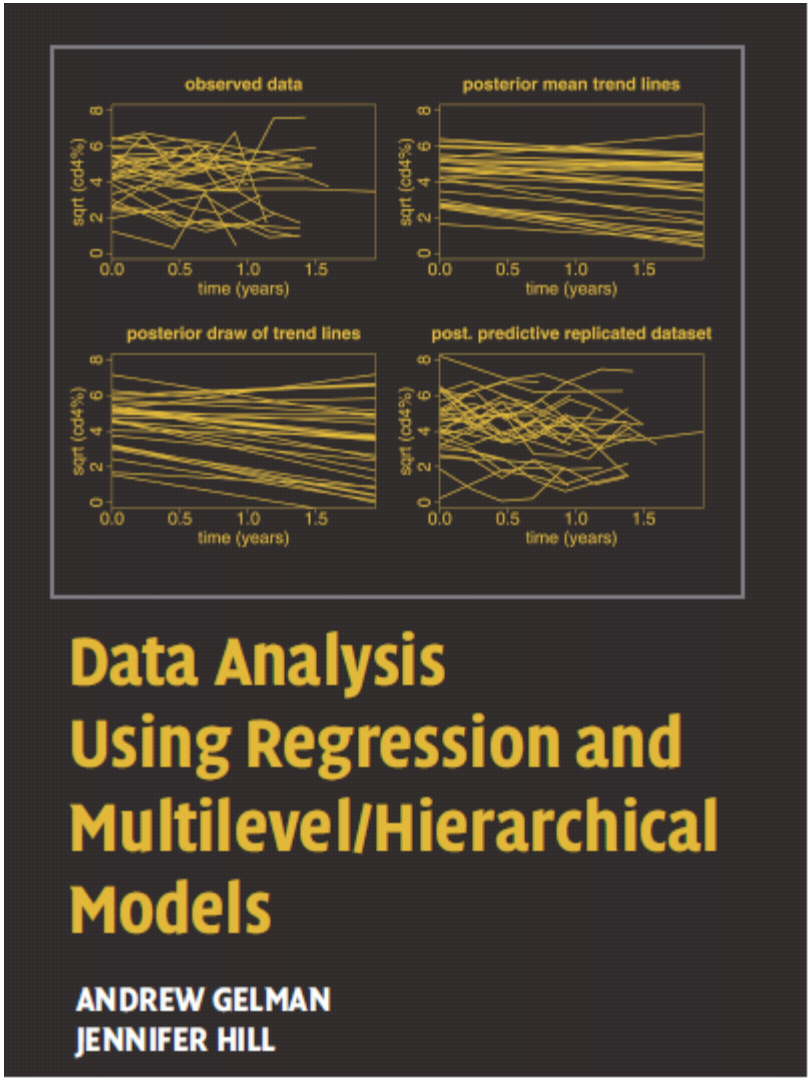

A. Design
Write deterministic model of process.
Existing theory, scientific objectives, intuition
Design / choose observations.
B. Model specification
Diagram relationship between observed and unobserved.
Write out posterior and joint distributions using general probability notation.
Choose appropriate probability distributions.
C. Model implementation
Write full conditional distributions.
Write MCMC sampling algorithm.
Or
Write code for MCMC software.
Implement MCMC on simulated data.
Implement MCMC on real data.
D. Model evaluation and inference
Posterior predictive checks
Probabilistic inference from single model
Model selection, model averaging

Learning outcomes

- Understand Bayesian inference using familiar examples.
- Appreciate one-to-one relationship between math and JAGS code.
- Be able to interpret coefficients of general linear models.
- Know how and why to center or standardize data.
- Be able to translate scalar linear equations into matrix equations.

A great follow-up

This book should be in your library:



The general Bayesian set-up

Recall that the posterior distribution of the unobserved quantities conditional on the observed ones is proportional to their joint distribution:

$$[\theta|y] \propto [\theta, y].$$

The joint distribution can be factored into a likelihood and priors for simple Bayesian models:

$$[\theta, \sigma^2] = [y \mid \theta, \sigma^2][\theta][\sigma^2]$$

A deterministic model of an ecological process is embedded in the likelihood like this...

$$[\boldsymbol{\theta}, \sigma^2] \propto [y \mid g(\boldsymbol{\theta}, x), \sigma^2][\boldsymbol{\theta}][\sigma^2]$$

Simple Bayesian regression models

As always, we start with a deterministic model,

$$\mu_i = \underbrace{g(\beta, x_i)}_{\text{deterministic model}}$$

where β is a vector of regression coefficients and $\mathbf{x}_i$ is a vector or predictor variables corresponding to observation y_i. We use likelihood to connect the predictions of our model to data:

$$\underbrace{[y_i \mid \mu_i, \sigma^2]}_{\text{stochastic model}}$$

$$[\beta, \sigma^2 \mid \mathbf{y}] \propto \prod_{i=1}^{n} [y_i \mid g(\beta, x_i), \sigma^2] [\boldsymbol{\theta}] [\sigma^2]$$

We choose appropriate deterministic functions (linear or non-linear) and appropriate probability distributions to compose a specific model. Simple and flexible.

Identical notation

$$y_i = g(\beta, x_i) + \epsilon_i$$
$$\epsilon_i \sim \text{normal}(0, \sigma^2)$$

is the same as

$$y_i \sim \text{normal}\big(g(\beta, x_i), \sigma^2\big),$$

but the second notation is much more flexible because it doesn't require additive errors.

You don't have to be normal!

Data (y-values)	Distribution	Mean function	Link
continuous, real valued	normal	$\mu = \beta_0 + \beta_1 x$	NA
discrete, strictly positive	Poisson	$\mu = e^{\beta_0 + \beta_1 x}$	$\log\left(\mu\right) = \beta_0 + \beta_1 x$
0 or 1	Bernoulli	$\mu = \frac{\exp\left(\beta_0 + \beta_1 x\right)}{\exp\left(\beta_0 + \beta_1 x\right) + 1}$	$\text{logit}\left(\mu\right) = \log\left(\frac{\mu}{1-\mu}\right) = \beta_0 + \beta_1 x$
0 – 1	beta	$\mu = \frac{\exp\left(\beta_0 + \beta_1 x\right)}{\exp\left(\beta_0 + \beta_1 x\right) + 1}$	$\text{logit}\left(\mu\right) = \log\left(\frac{\mu}{1-\mu}\right) = \beta_0 + \beta_1 x$
continuous, strictly positive	lognormal or gamma	$\mu = e^{\beta_0 + \beta_1 x}$	$\log\left(\mu\right) = \beta_0 + \beta_1 x$

Lots of flexibility as a modeler

Continent-wide Adélie penguin population dynamics

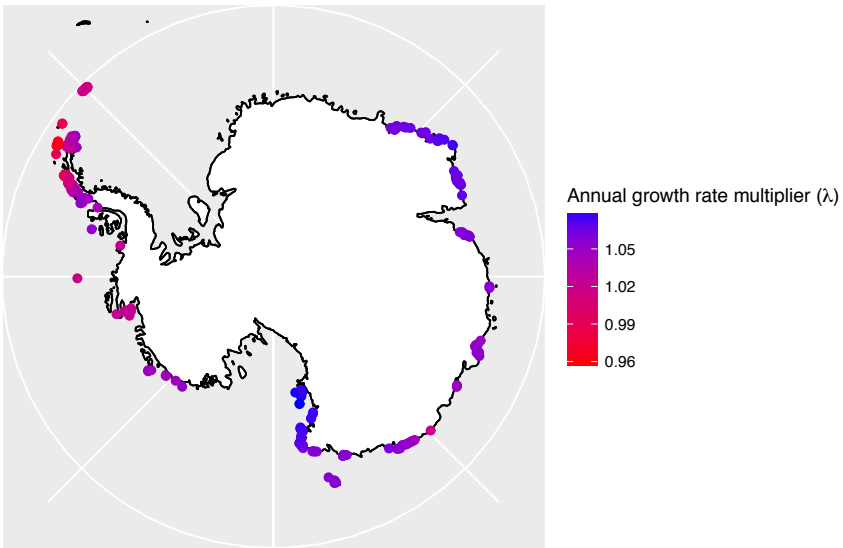
$$g(\beta, z_{i,t}) = \log(z_{i,t-1} e^{(\beta_{0,i} + \beta_1 \text{wsic}_{i,t} + \beta_2 \text{ssic}_{i,t} + \beta_3 \text{krill}_{i,t})\Delta t})$$

$$y_{i,t} \sim \text{Poisson}(z_{i,t})$$

$$z_{i,t} \sim \text{lognormal}(z_{i,t} \mid g(\beta_{0,i}, \beta_1, \beta_2, \beta_3, z_{i,t-1}), \sigma^2_{\text{process}})$$

$$\beta_{0,i} \sim \text{normal}(\mu_{\text{site}}, \varsigma^2_{\text{site}})$$

Comment about moment matching the mean here.



Normal data, continuous and real valued

$$
\begin{aligned}
[\beta_0, \beta_1, \sigma \mid \mathbf{y}] &\propto \prod_{i=1}^{n} \text{normal}\big(y_i \mid g(\beta_0, \beta_1, x_i), \sigma^2\big) \times \\
&\quad \text{normal}\big(\beta_0 \mid 0, 1000\big)\, \text{normal}\big(\beta_1 \mid 0, 1000\big) \times \\
&\quad \text{uniform}\big(\sigma \mid 0, 100\big) \\
g(\beta_0, \beta_1, x_i) &= \beta_0 + \beta_1 x_i
\end{aligned}
$$

```
b0 ~ dnorm(0, .001)
b1 ~ dnorm(0, .001)
sigma ~ dunif(0, 100)
tau <- 1/sigma^2
for (i in 1:length(y)){
  mu[i] <- b0 + b1 * x[i]
  y[i] ~ dnorm(mu[i], tau)
}
```

Exercise

What is the interpretation of β_0? Of β_1?

Poisson, discrete and positive

$$
\begin{aligned}
[\beta_0, \beta_1 \mid \mathbf{y}] &\propto \prod_{i=1}^{n} \text{Poisson}\big(y_i \mid g(\beta_0, \beta_1, x_i)\big) \times \\
&\quad \text{normal}\big(\beta_0 \mid 0, 1000\big)\, \text{normal}\big(\beta_1 \mid 0, 1000\big) \\
g(\beta_0, \beta_1, x_i) &= e^{\beta_0 + \beta_1 x_i}
\end{aligned}
$$

```
b0 ~ dnorm(0, .001)
b1 ~ dnorm(0, .001)
for(i in 1:length(y)){
  log(mu[i]) <- b0 + b1 * x[i]
  y[i] ~ dpois(mu[i])
}
```

or

```
mu[i] <- exp(b0 + b1 * x[i])
y[i] ~ dpois(mu[i])
```

Exercise

What is the interpretation of β_0? Of β_1

Hint– Expand $e^{\beta_0+\beta_1 x_i}$

Poisson with offset

$\log(u_i) = \textit{offset}$ for observation i

$$
\begin{aligned}
[\beta_0, \beta_1 \mid \mathbf{y}] &\propto \prod_{i=1}^{n} \text{Poisson}\left(y_i \mid g\left(\beta_0, \beta_1, x_i, u_i\right)\right) \times \\
&\quad \text{normal}\left(\beta_0 \mid 0, 1000\right) \text{normal}\left(\beta_1 \mid 0, 1000\right) \\
g\left(\beta_0, \beta_1, x_i, u_i\right) &= u_i e^{\beta_0 + \beta_1 x_i}
\end{aligned}
$$

```
b0 ~ dnorm(0, .001)
b1 ~ dnorm(0, .001)
for(i in 1:length(y)){
  log(mu[i]) <- log(u[i]) + b0 + b1 * x[i]
  y[i] ~ dpois(mu[i])
}
```

Bernoulli, data 0 or 1 (aka logistic)

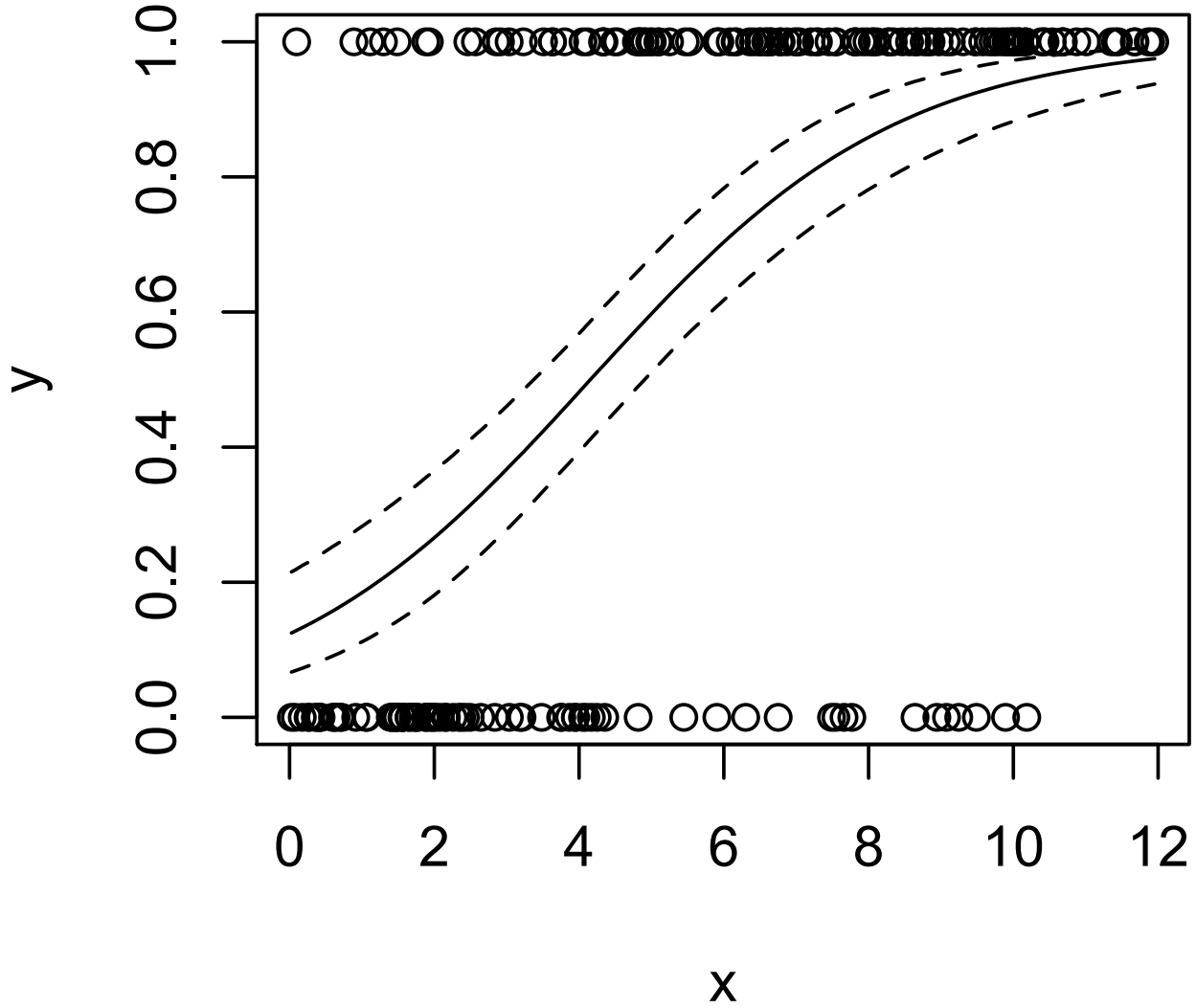
$$
\begin{aligned}
[\beta_0, \beta_1 \mid \mathbf{y}] &\propto \prod_{i=1}^{n} \text{Bernoulli}\left(y_i \mid g\left(\beta_0, \beta_1, x_i\right)\right) \times \\
&\quad \text{normal}\left(\beta_0 \mid 0, 2\right) \text{normal}\left(\beta_1 \mid 0, 2\right) \\
g\left(\beta_0, \beta_1, x_i\right) &= \frac{e^{\beta_0 + \beta_1 x_i}}{e^{\beta_0 + \beta_1 x_i} + 1}
\end{aligned}
$$

```
b0 ~ dnorm(0, .5)
b1 ~ dnorm(0, .5)
for(i in 1:length(y)){
  logit(p[i]) <- b0 + b1 * x[i]
    y[i] ~ dbern(p[i])
 }
```

or

```
p[i] <- inv.logit(b0 + b1 * x[i])
y[i] ~ dbin(p[i])
```

Bernoulli, data 0 or 1 (aka logistic)



Exercise

What is the interpretation of the line, i.e. the model fit? What is the interpretation of β_0? Of β_1?

Interpretation of the line, odds ratios, and odds

The interpretation of the line is the probability that $y = 1$ at a given x, i.e., $[y_i = 1 \mid x_i]$. We use these predictions to construct odds ratios:

$$\overbrace{\log\left(\underbrace{\frac{[y_i = 1 \mid x_i]}{[y_i = 0 \mid x_i]}}_{\text{odds}}\right)}^{\text{odds ratio}} = \beta_0 + \beta_1 x_i$$

Interpreting the line and the intecept

The interpretation of the intercept is difficult. Nonsense at x=0. Need to evaluate at some other point, for example the mean of x,
inverse logit$\big(\beta_0 + \beta_1 \text{mean}(x)\big)$, which can be more easily accomplished by rescaling the data, discussed shortly. If data are rescaled such that $x = 0$ at the mean of x, then then e^{β_0} is the odds that $y = 1$ at the mean of x. For example, if $e^{\beta_0} = 2$, it is twice as likely that $y = 1$ than $y = 0$.

Interpreting slopes

Odds ratios and odds:

$$\overbrace{\log\left(\underbrace{\frac{[y_i = 1 \mid x_i]}{[y_i = 0 \mid x_i]}}_{\text{odds}}\right)}^{\text{odds ratio}} = \beta_0 + \beta_1 x_i$$

β_1 is the additive change in the odds ratio per unit change in x. Exponentiating both sides we see that e_1^{β} is the multiplicative change in the odds that $y = 1$ given x per unit change in x.

Guidance

I advise using the inverse logit function

$$g(\beta_0, \beta_1, x_i) = \frac{e^{\beta_0 + \beta_1 x_i}}{e^{\beta_0 + \beta_1 x_i} + 1}$$

with specified x values as a basis for interpreting the model. Odds and, worse, odds ratios, can be difficult to understand and communicate.

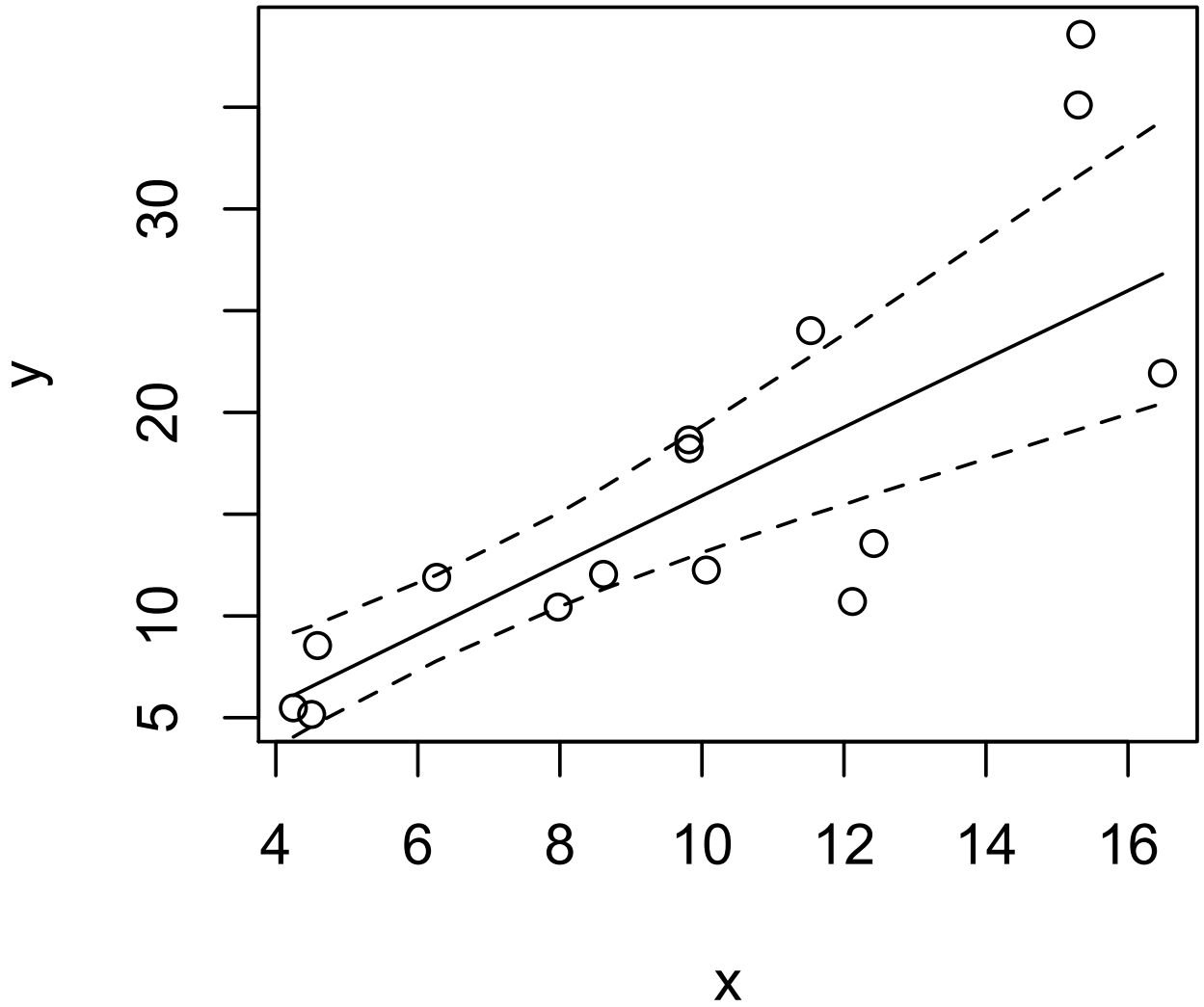
lognormal, data continuous and > 0

$$
\begin{aligned}
[\beta_0, \beta_1, \sigma \mid \mathbf{y}] &\propto \prod_{i=1}^{n} \text{lognormal}\left(y_i \mid \log\left(g\left(\beta_0, \beta_1, x_i\right)\right), \sigma^2\right) \times \\
&\quad \text{normal}\left(\beta_0 \mid 0, 1000\right) \text{normal}\left(\beta_1 \mid 0, 1000\right) \times \\
&\quad \text{uniform}\left(\sigma \mid 0, 5\right) \\
g\left(\beta_0, \beta_1, x_i\right) &= e^{\beta_0 + \beta_1 x_i}
\end{aligned}
$$

Talk about the interpretation of σ.

```
b0 ~ dnorm(0, .001)
b1 ~ dnorm(0, .001)
sigma ~ dunif(0, 5)
tau <- 1/sigma^2
for(i in 1:length(y)){
  mu[i] <- exp(b0 + b1 * x[i])
  y[i] ~ dlnorm(log(mu[i]), tau)
}
```

lognormal, data continuous and > 0



lognormal, data continuous and > 0

$$
\begin{aligned}
\left[\beta_0, \beta_1, \sigma \mid \mathbf{y}\right] &\propto \prod_{i=2}^{n} \text{lognormal}\left(y_i \mid \log\left(g\left(\beta_0, \beta_1, y_{i-1}, H_i\right)\right), \sigma^2\right) \times \\
&\quad \text{normal}\left(\beta_0 \mid 0, 1000\right) \text{normal}\left(\beta_1 \mid 0, 1000\right) \times \\
&\quad \text{uniform}\left(\sigma \mid 0, 5\right) \\
g\left(\beta_0, \beta_1, y_{i-1}, H_i\right) &= y_{i-1} e^{\beta_0 + \beta_1 y_{i-1}} - H_i
\end{aligned}
$$

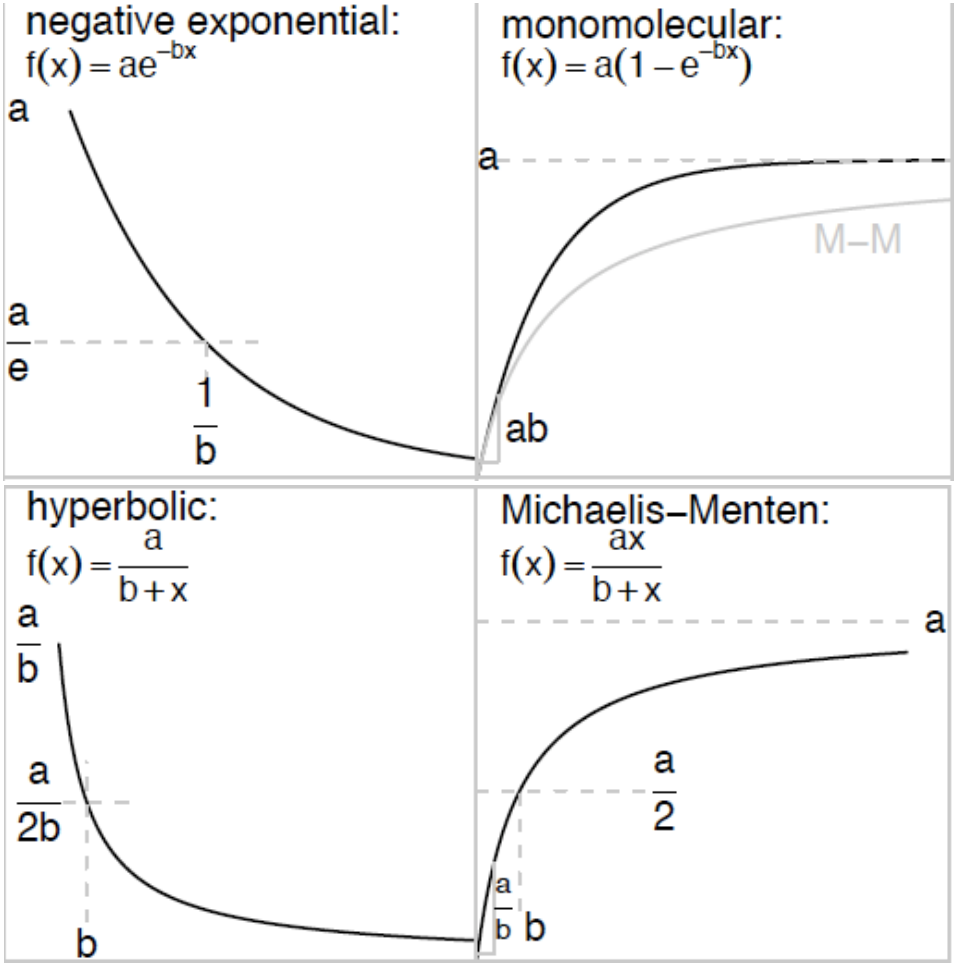
Talk about the bounding trick.

```
b0 ~ dnorm(0, .001)
b1 ~ dnorm(0, .001)
sigma ~ dunif(0, 5)
tau <- 1/sigma^2
for(i in 2:length(y)){
  mu[i] <- y[i-1] * exp(b0 + b1 * y[i-1]) - H[i]
  y[i] ~ dlnorm(log(max(.000001, mu[i])), tau)
}
```

Exercise

What is the interpreation of β_0? β_1?

Nonlinear regression



Figures c/o Bolker, B. 2008. *Ecological Models and Data in R*. Princeton University Press, Princeton, NJ. USA.

Centering and standardizing

The remainder of the slides apply to all of the general linear models, but I will use a simple linear for normally distributed data as an example.

Centering predictor data

$$y_i = \beta_0 + \beta_1 (x_i - \bar{x})$$

Why complicate things?

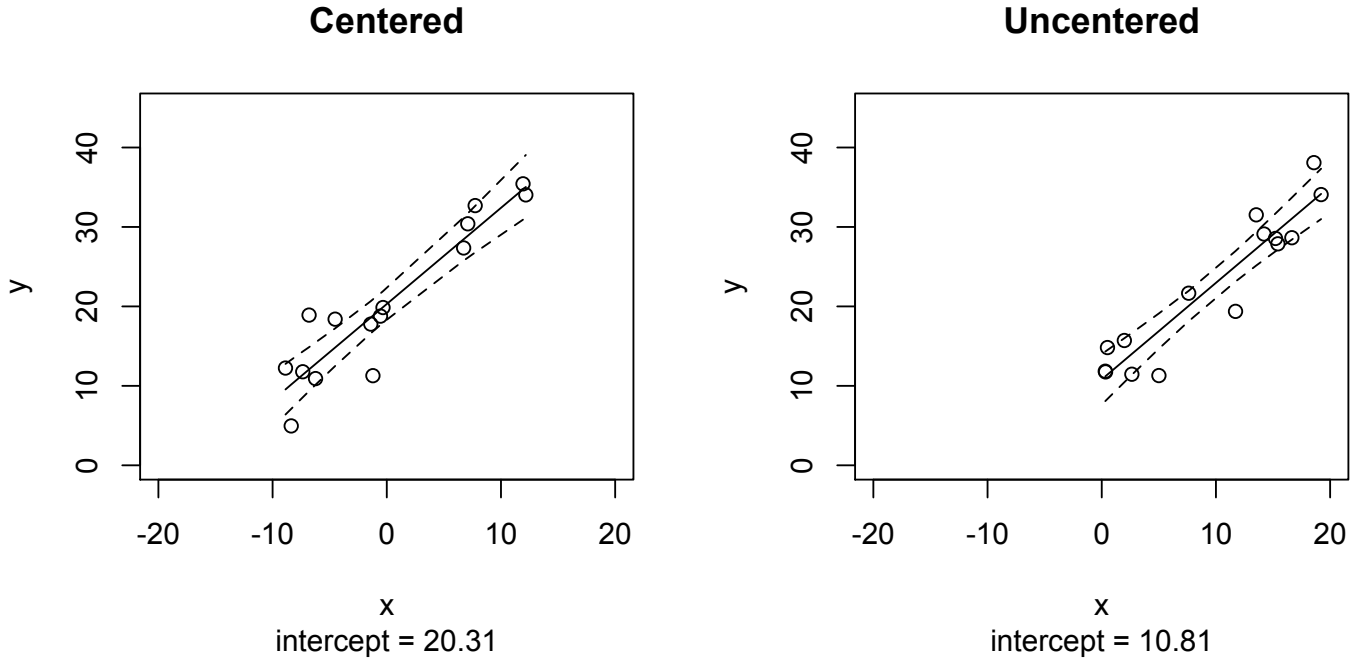
- To reduce autocorrelation in MCMC chain and speed convergence.
- To make the intercept more easily interpretable.

Centering predictor data

$$
\begin{aligned}
[\beta_0, \beta_1, \sigma \mid \mathbf{y}] &\propto \prod_{i=1}^{n} \text{normal}\left(y_i \mid g\left(\beta_0, \beta_1, x_i, \bar{x}\right), \sigma^2\right) \times \\
&\quad \text{normal}\left(\beta_0 \mid 0, 1000\right) \text{normal}\left(\beta_1 \mid 0, 1000\right) \times \\
&\quad \text{uniform}\left(\sigma \mid 0, 100\right) \\
g\left(\beta_0, \beta_1, x_i\right) &= \beta_0 + \beta_1\left(x_i - \bar{x}\right)
\end{aligned}
$$

```
b0 ~ dnorm(0, .001)
b1 ~ dnorm(0, .001)
sigma ~ dunif(0, 100)
tau <- 1/sigma^2
xBar <- mean(x)
for (i in 1:length(y)){
  mu[i] <- b0 + b1 * (x[i] - xBar)
  y[i] ~ dnorm(mu[i], tau)
}
b0_UC <- b0 - b1 * xBar
```

Recovering uncentered parameters



$$
\begin{aligned}
B_0 &= \beta_0 - \beta_1 * \bar{x} \\
B_1 &= \beta_1
\end{aligned}
$$

- For this to work properly, all of the coefficients in the model must be *added*.
- Slopes will not be the same if there is an interaction term or quadratic. In these cases, back transforming is not simple.

Standardizing predictor data

$$y_i = \beta_0 + \beta_1 \left(\frac{x_i - \bar{x}}{\sigma_x} \right)$$

Why complicate things?

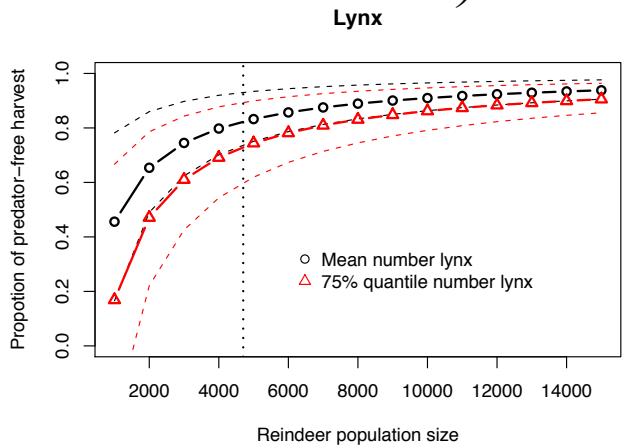
- To reduce autocorrelation in MCMC chain and speed convergence.
- To make the intercept more easily interpretable.
- To make parameters more easily comparable.

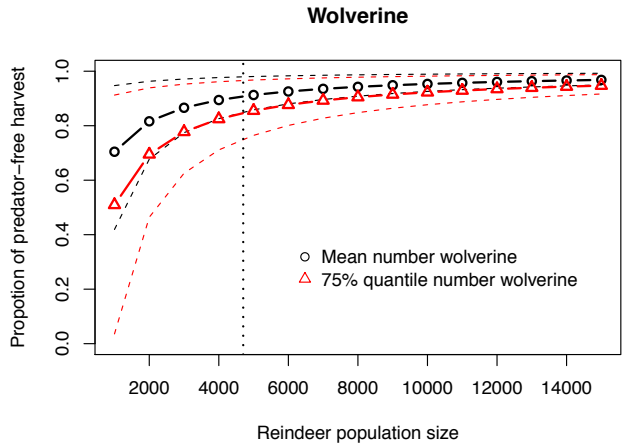
Interpreting the intercept

Reindeer model: example of centering to improve interpretation

$$\lambda_{i,t} = e^{\left(r - \frac{r}{K} N_{t-1} + \beta_1 lynx + \beta_2 wolverine + \beta_3 gradient + \beta_4 NAO \right) \Delta t}$$

$$N_t = \lambda_{i,t} N_{t-1} - H_t$$





Standardizing predictor data

$$
\begin{aligned}
[\beta_0, \beta_1, \sigma \mid \mathbf{y}] &\propto \prod_{i=1}^{n} \text{normal}\big(y_i \mid g(\beta_0, \beta_1, x_i, \bar{x}, \sigma_x), \sigma^2\big) \times \\
&\quad \text{normal}\big(\beta_0 \mid 0, 1000\big)\,\text{normal}\big(\beta_1 \mid 0, 1000\big) \times \\
&\quad \text{uniform}\big(\sigma \mid 0, 100\big) \\
g\big(\beta_0, \beta_1, x_i\big) &= \beta_0 + \beta_1\left(\frac{x_i - \bar{x}}{\sigma_x}\right)
\end{aligned}
$$

```
b0 ~ dnorm(0, .001)
b1 ~ dnorm(0, .001)
sigma ~ dunif(0, 100)
tau <- 1/sigma^2
xBar <- mean(x)
xSD <- sd(x)
for (i in 1:length(y)){
  mu[i] <- b0 + b1 * ((x[i] - xBar)/xSD
  y[i] ~ dnorm(mu[i], tau)
}
```

Recovering unstandardized parameters

$$
\begin{aligned}
y_i &= \beta_0 + \beta_1 \left(\frac{x_i - \bar{x}}{\sigma_x} \right) \\
y_i &= \beta_0 + \frac{\beta_1}{\sigma_x} - \frac{\beta_1 \bar{x}}{\sigma_x} \\
B_0 &= \beta_0 - \frac{\beta_1 \bar{x}}{\sigma_x} \\
B_1 &= \frac{\beta_1}{\sigma_x}
\end{aligned}
$$

- This only works if there are not squared values or interactions.
- Generally, I back-transform predictions not parameters. (How?)

Matrix notation for linear models

Remeber matrix multiplication?

Exaample of n observations for 2 predictor variables $x_{i,1}$ and $x_{i,2}$.

$$\begin{pmatrix} 1 & x_{1,1} & x_{1,2} \\ 1 & x_{2,1} & x_{2,2} \\ 1 & x_{3,1} & x_{3,2} \\ 1 & . & . \\ 1 & . & . \\ 1 & . & . \\ 1 & x_{n,1} & x_{n,2} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} = \begin{pmatrix} \beta_0 + \beta_1 x_{1,1} + \beta_2 x_{1,2} \\ \beta_0 + \beta_1 x_{2,1} + \beta_2 x_{2,2} \\ \beta_0 + \beta_1 x_{3,1} + \beta_2 x_{3,2} \\ . \\ . \\ . \\ \beta_0 + \beta_1 x_{n,1} + \beta_2 x_{n,2} \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ . \\ . \\ . \\ \mu_n \end{pmatrix}$$

Matrix notation for linear models

You will often see models written using something like

$$y_i \sim \text{normal}(\mathbf{x}_i'\beta, \sigma^2)$$

or

$$y_i \sim \text{normal}(\mathbf{x}_i^T\beta, \sigma^2)$$

or

$$y_i \sim \text{normal}(\mathbf{X}_i\beta, \sigma^2)$$

or

$$\mathbf{y} \sim \text{multivariate normal}(\mathbf{X}\beta, \sigma^2 I)$$

Note that $\mathbf{X}$ is a matrix with ones in column 1 and values of covariates in other columns. Thus, $\mathbf{X}\beta$ returns a vector.

Exercise

We want to predict species richness (number of different species) of avian communities in 50 US states based on a set of p predictor variables. Draw the Bayesian network and write the posterior and joint distribution, inducing the specific distributions appropriate for this problem. We assume that the response and predictor variables are measured perfectly. Use matrix notation to specify the deterministic model.

Code for matrix computation of linear model: Prediting bird species diveresity

```
model {
    # PRIORS, p = number of coefficients, including intercept
     for(i in 1:p) {
       beta[i] ~ dnorm(0, 0.01)
     }
   # LIKELIHOOD
     # n = number of states (rows in X)
     # y = number of birds in each state
     # X is a n x p matrix with 1s in column 1
     z <- X %*% beta # the regression model, returns a vector
     #of length n
     for(i in 1:n)   {
       y[i] ~ dpois(lambda[i])
       lambda[i] <- exp(z[i])
     }
}
```