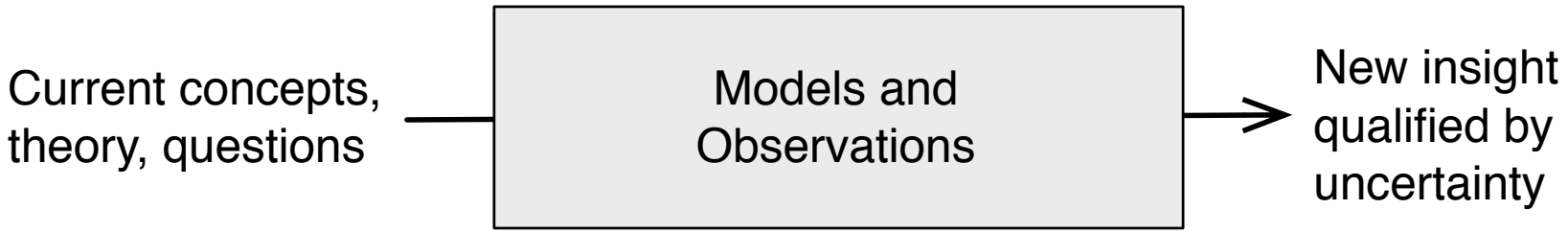
Bayesian Regression Continued

ESS 575 Models for Ecological Data

N. Thompson Hobbs

March 26, 2019

Where are we?


Current concepts,
theory, questions
Models and
Observations
New insight
qualified by
uncertainty

Bayesian regression models are often key compoents of more complex models

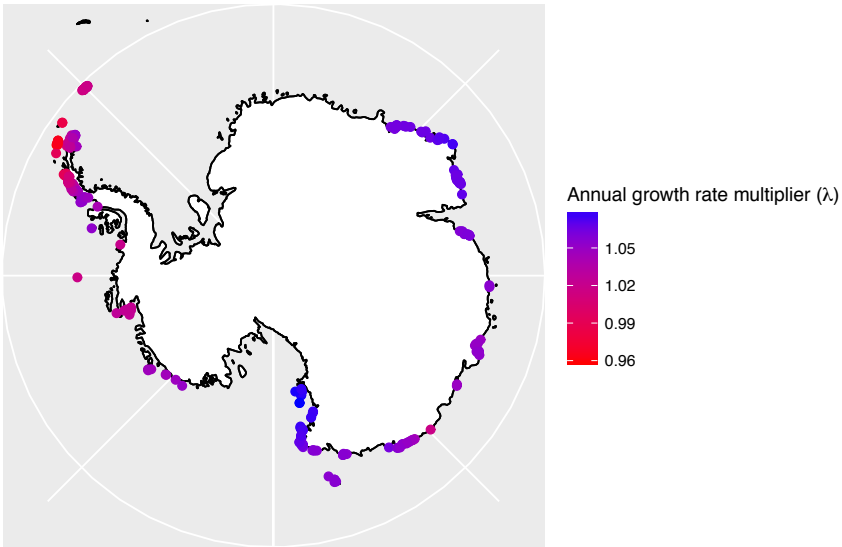
Continent-wide Adélie penguin population dynamics

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

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

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

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



Lecture roadmap

Before break:

- Overview of simple Bayesian models
- Exercise in writing regression models
- Multi-level models
- Introduction to hierarchical models

Subsequent two weeks: Practice writing models, learning JAGS

Today:

- A bit of review
- Multilevel models for intercepts and slopes
- Some nuts and bolts: Standardizing, interpreting coefficients, matrix notation

Thursday: Model building problems

Next week: Missing data

Lab roadmap: Multi-lelvel models (aka random effects)

This week and next week:

- More practice writing models and codinng them
- Random intercepts
- Random intercept for groups with anicallary data on groups
- Random intercepts and slopes for groups
- Nuts and bolts
  - How to code subscripts
  - Predictions (and plotting) across groups

Learning outcomes for Bayesian regression

- Be able to write proper Bayesian, regression models for all types of data.
- 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.
- Understand options for dealing with missing data
- Understand how to formulate multi-level models for slopes and intercepts (aka "random effects")

Review: 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...

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

Review: Simple Bayesian regression models

As always, we start with a deterministic model,

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

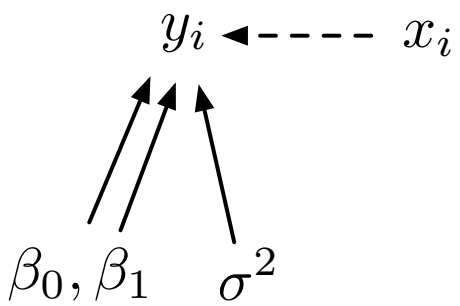
where β is a vector of regression coefficients and $\mathbf{x}_i$ is a vector or predictor variables corresponding to observation y_i. We use likelihood to connect the predictions of our model to data:

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

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

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

Review: Building a simple regression



$$g(\beta_0, \beta_1, x_i) = \beta_0 + \beta_1 x_i$$

$$[\beta_0, \beta_1, \sigma^2 \mid y_i] \propto [\beta_0, \beta_1, \sigma^2, y_i]$$

factoring rhs using DAG:

$$[\beta_0, \beta_1, \sigma^2 \mid y_i] \propto [y_i \mid g(\beta_0, \beta_1, x_i), \sigma^2][\beta_0], [\beta_1][\sigma^2]$$

joint for all data :

$$[\beta_0, \beta_1, \sigma^2 \mid \boldsymbol{y}] \propto \prod_{i=1}^{n} [y_i \mid g(\beta_0, \beta_1, x_i), \sigma^2][\beta_0][\beta_1][\sigma^2]$$

choose specific distributions:

$$\begin{aligned}[\beta_0, \beta_1, \sigma^2 \mid \boldsymbol{y}] \propto & \prod_{i=1}^{n} \text{normal}(y_i \mid g(\beta_0, \beta_1, x_i), \sigma^2) \\ & \times \text{normal}(\beta_0 \mid 0, 10000)\text{normal}(\beta_1 \mid 0, 10000) \\ & \times \text{uniform}(\sigma^2 \mid 0, 500)\end{aligned}$$

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

or

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

Bernoulli, data 0 or 1 (aka logistic)

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

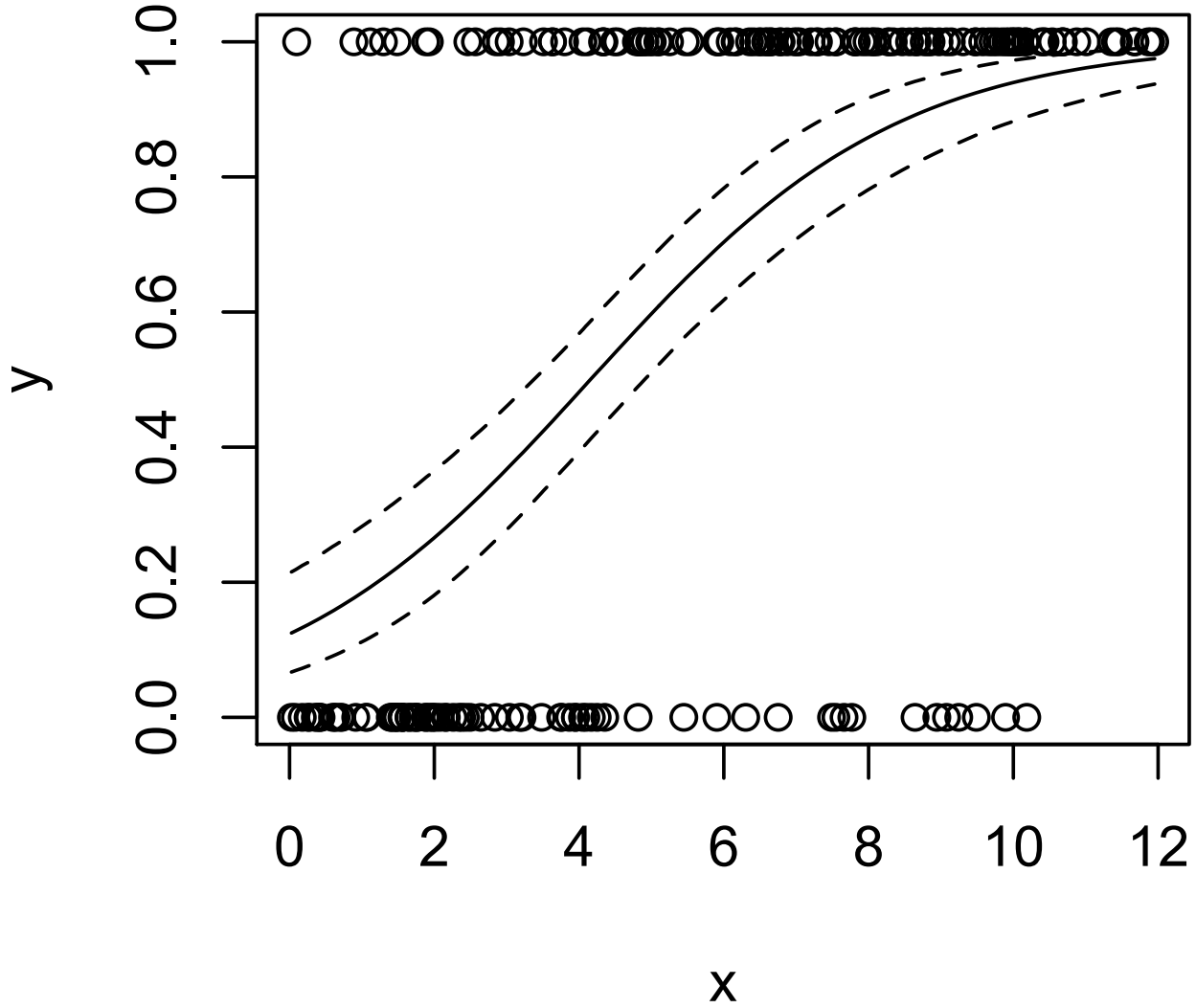
```
b0 ~ dnorm(0, .5)
b1 ~ dnorm(0, .5)
for(i in 1:length(y)){
  p[i] <- inv.logit(b0 + b1 * x[i])
  y[i] ~ dbern(p[i])

 }
```

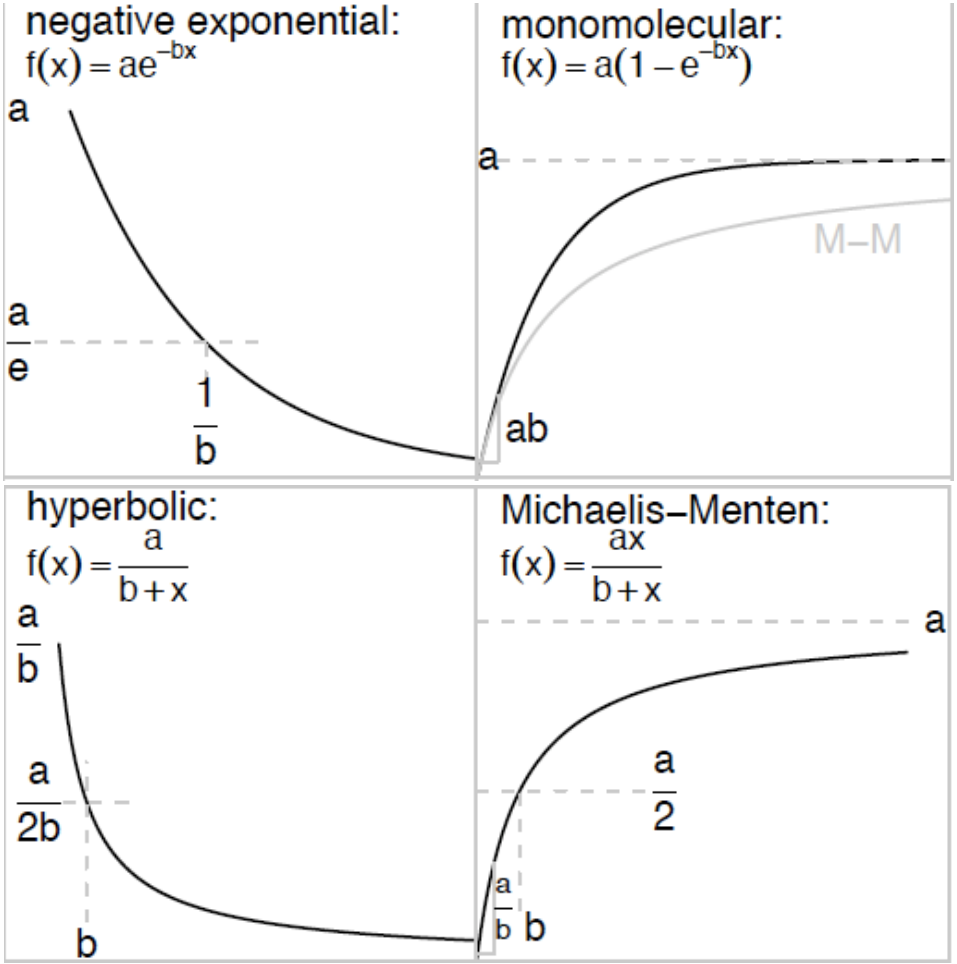
or

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

Bernoulli, data 0 or 1 (aka logistic)

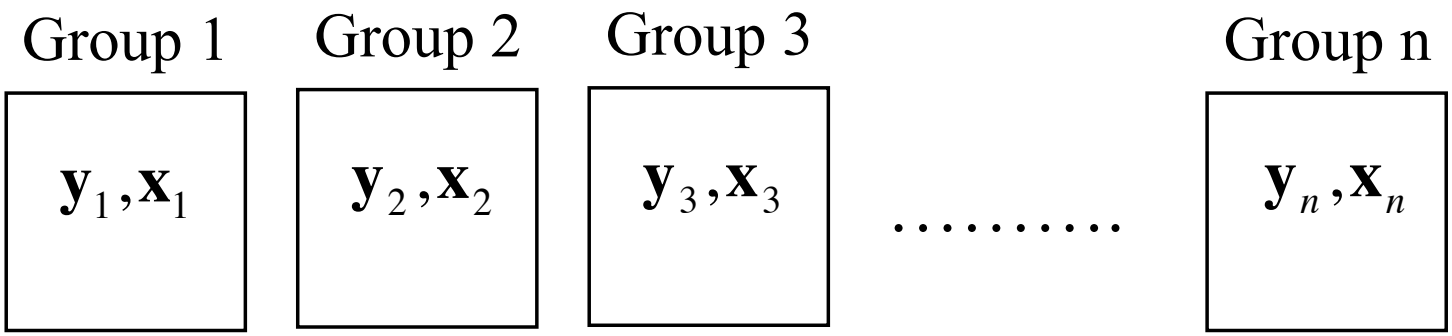


Nonlinear regression

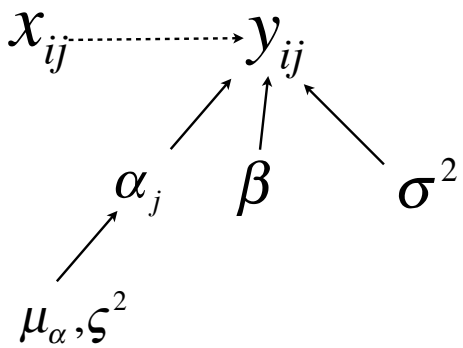


Figures c/o Bolker, B. 2008. *Ecological Models and Data in R*. Princeton University Press, Princeton, NJ. USA.

Review: The multi-level problem

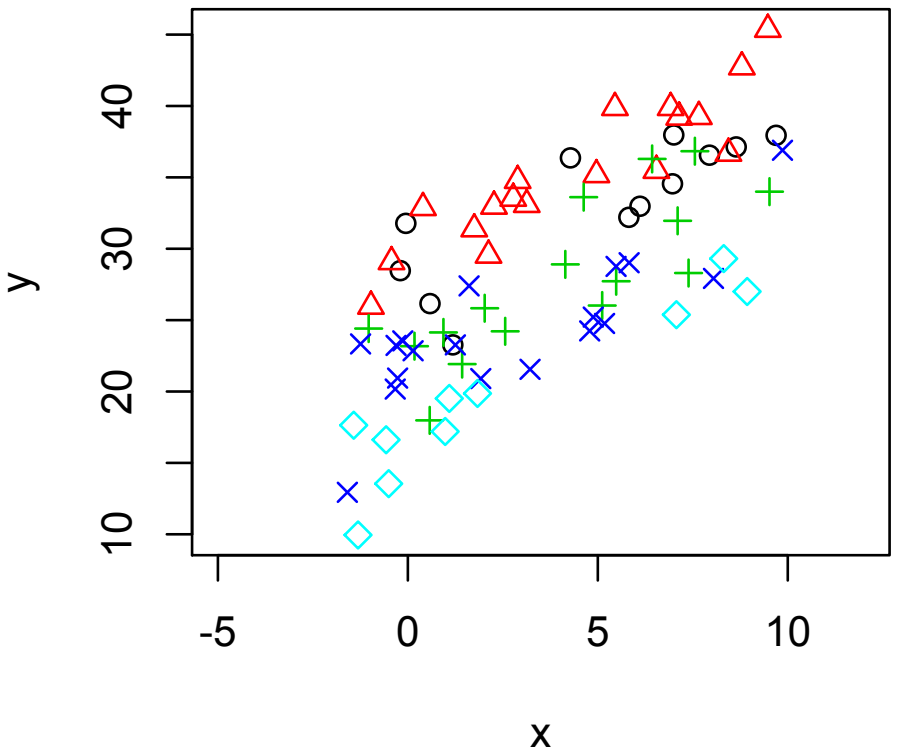


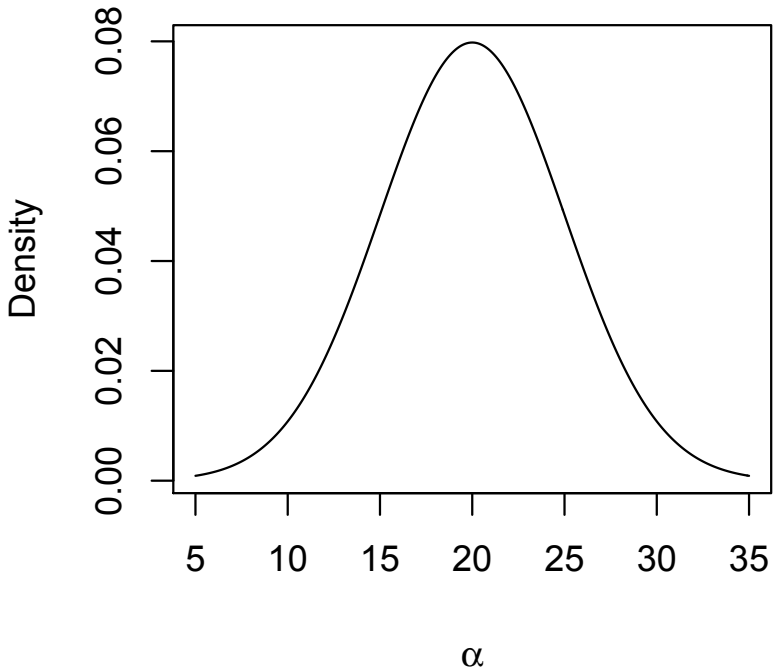
We can model intecepts (or slopes) for each group



$$[\beta, \boldsymbol{\alpha}, \sigma^2, \mu_\alpha, \varsigma^2, \mid \mathbf{y}] \propto \prod_{i=1}^{n_j} \prod_{j=1}^{J} \text{normal}(\text{y}_{\text{ij}} \mid \alpha_{\text{j}} + \beta \text{x}_{\text{ij}}, \sigma^2)$$
$$\times \text{normal}(\alpha_j \mid \mu_\alpha, \varsigma^2)$$
$$\times \text{normal}(\beta \mid 0, 10000)\text{normal}(\mu_\alpha \mid 0, 1000)$$
$$\times \text{uniform}(\sigma^2 \mid 0, 100)\text{uniform}(\varsigma^2 \mid 0, 100)$$

We seek to find the marginal posterior distribution of intercepts





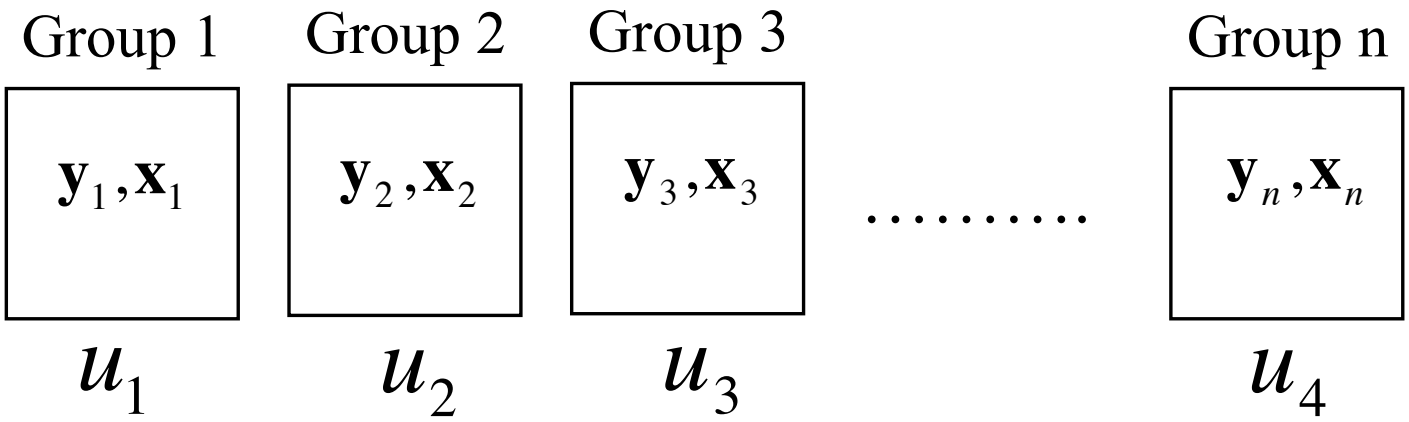
Some notation that might be confusing

$$\mu_{ij} = \beta_0 + \beta_1 x_{ij} + \alpha_j$$
$$y_{ij} \sim \text{normal}(\mu_{ij}, \sigma^2)$$
$$\alpha_j \sim \text{normal}(0, \varsigma^2)$$

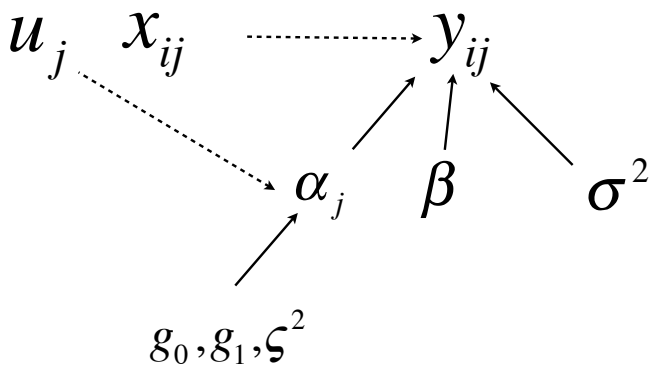
is identical to:

$$\mu_{ij} = \alpha_j + \beta_1 x_{ij}$$
$$y_{ij} \sim \text{normal}(\mu_{ij}, \sigma^2)$$
$$\alpha_j \sim (\mu_\alpha, \varsigma^2)$$

We can model the intercepts using data on the groups

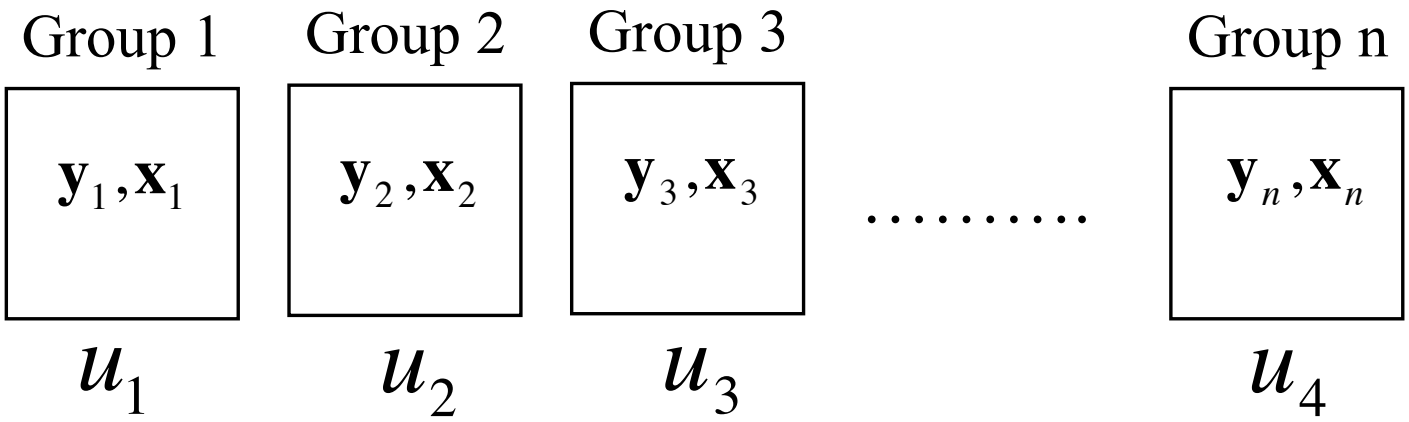


We can model the intercepts using data on the groups



$$
\begin{aligned}
\left[\boldsymbol{\alpha}, \beta, \sigma^2, \mathbf{g}, \varsigma^2, \mid \mathbf{y}\right] \propto & \prod_{i=1}^{n_j} \prod_{j=1}^{J} \text{normal}\left(y_{ij} | \alpha_j + \beta x_{ij}, \sigma^2\right) \\
& \times \text{normal}\left(\alpha_j | g_0 + g_1 u_j, \varsigma^2\right) \\
& \times \text{normal}\left(\beta | 0, .001\right) \text{normal}\left(g_0 | 0, 1000\right) \text{normal}\left(g_1 | 0, 1000\right) \\
& \times \text{inverse gamma}\left(\sigma^2 | .001, .001\right) \text{inverse gamma}\left(\varsigma^2 | .001, .001\right)
\end{aligned}
$$

Borrowing strength



Modeling the intecepts *and* slopes: correlation matrix

Correlations

	Weight in kg	Hours of Sleep	Exposure while Sleeping	Life Span
Weight in kg	1	-.307	.338	.302
Hours of Sleep	-.307	1	-.642	-.410
Exposure while Sleeping	.338	-.642	1	.360
Life Span	.302	-.410	.360	1

1

[1] http://www.theanalysisfactor.com/covariance-matrices/

Correlation and covariance

$$\rho_{x,y} = \frac{\mathrm{cov}(x,y)}{\sigma_x \sigma_y} = \frac{\mathrm{E}((x-\mu_x)(y-\mu_y))}{\sigma_x \sigma_y}$$

Therefore:

$$\mathrm{cov(x,y)} = \rho_{x,y} \sigma_x \sigma_y$$

Modeling intercepts and slopes: covariance matrix

Imagine a vector of 3 random variables, $(z_i, z_2, z_3)'$. The covariance between any two of these random variables is simply an unstandardized version of the correlation between them– it is correlation measured in the units of the random variables. The covariance matrix (aka variance covariance matrix) of the random variable is:

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_1^2 & \text{Cov}_{1,2} & \text{Cov}_{1,3} \\ \text{Cov}_{2,1} & \sigma_2^2 & \text{Cov}_{2,3} \\ \text{Cov}_{3,1} & \text{Cov}_{3,2} & \sigma_3^2 \end{pmatrix}$$

Generalizing, a $m \times m$ covariance matrix has the variances of the random variable on the diagonal and the covariance on the off diagonal. The covariance between random variable i and j is $\text{Cov}_{ij} = \rho\sigma_i\sigma_j$ where ρ is the correlation coefficient, which takes on values between -1 and 1. Covariance can take on values between $-\infty$ and $+\infty$.

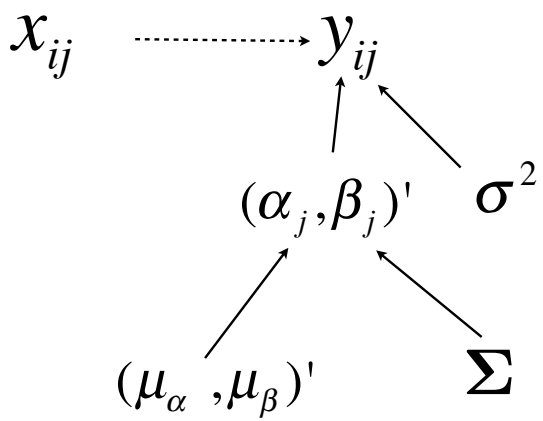
Covariance matrix for two parameter model

Imagine that we have $j = 1, ..., J$ groups with multiple observations within groups and we fit a two parameter linear model to each group, finding J intercepts and slopes. We denote the vector of intercepts as $\boldsymbol{\alpha}$ and the vector of slopes as $\boldsymbol{\beta}$. We can calculate the variance for each vector $(\sigma^2_\alpha, \sigma^2_\beta)$ as well as the correlation between the vectors ρ. The variance covariance matrix is:

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma^2_\alpha & \text{Cov}(\boldsymbol{\alpha}, \boldsymbol{\beta}) \\ \text{Cov}(\boldsymbol{\beta}, \boldsymbol{\alpha}) & \sigma^2_\beta \end{pmatrix},$$

where $\text{Cov}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \text{Cov}(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \rho\sigma_\alpha\sigma_\beta$

Modeling the intercepts and slopes



$$\begin{pmatrix} \alpha_j \\ \beta_j \end{pmatrix} \sim \text{multivariate normal} \left(\begin{pmatrix} \mu_\alpha \\ \mu_\beta \end{pmatrix}, \mathbf{\Sigma} \right)$$

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_\alpha^2 & \rho\sigma_\alpha\sigma_\beta \\ \rho\sigma_\alpha\sigma_\beta & \sigma_\beta^2 \end{pmatrix}$$

Modeling the intercepts and slopes for a two parameter model

$$\begin{aligned}\left[\boldsymbol{\alpha}, \boldsymbol{\beta}, \mu_\alpha, \mu_\beta, \sigma^2_{\text{reg}}, \sigma^2_\alpha, \sigma^2_\beta, \rho | \mathbf{y}\right] &\propto \prod_{j=1}^{J} \prod_{i=1}^{n_j} \text{normal}(y_{ij} | \alpha_j + \beta_j x_{ij}, \sigma^2_{\text{reg}}) \\ &\times \text{MVN}\left(\begin{pmatrix} \alpha_j \\ \beta_j \end{pmatrix} \middle| \begin{pmatrix} \mu_\alpha \\ \mu_\beta \end{pmatrix}, \boldsymbol{\Sigma} \right) \\ &\times \text{priors on } \mu_\alpha, \mu_\beta, \sigma^2_{\text{reg}}, \sigma^2_\alpha, \sigma^2_\beta, \rho\end{aligned}$$

Note that the likelihood does not need to be normal. It could be Poisson, Bernoulli, gamma, lognormal etc.

Modeling the intercepst and slopes for >1 slope

See Gelman and Hill 2009 [2], pages 376-380

[2] Gelman, A., and J. Hill. 2009. Data analysis using regression and multilevel / hierarchical models. Cambridge University Press, Cambridge, UK.

Standardizing the x's, (aka, covariates, predictor variables)

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.

Standardizing

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

Why complicate things?

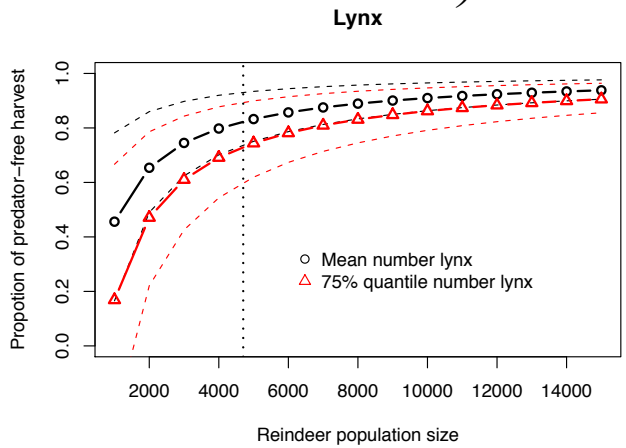
- To reduce autocorrelation in MCMC chain and speed convergence.
- To make the intercept more easily interpretable.
- To make slopes comparable in magnitude by putting them on the same scale.
- To allow interpretation of slopes in the presence of interactions.

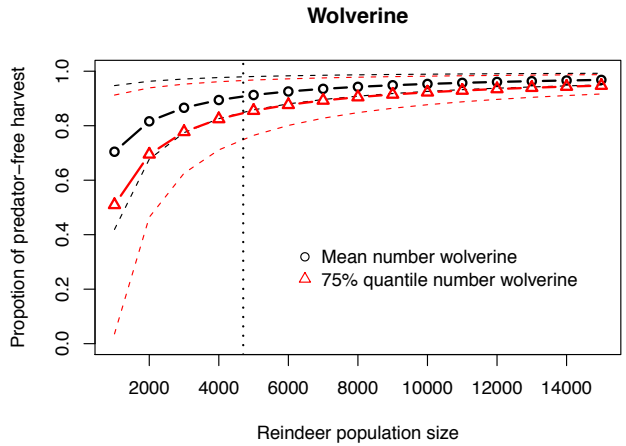
Interpreting the intercept

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

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

```
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 can be algebraically tricky when there are exponents or interactions of x's.
- Generally, I back-transform predictions not parameters for plotting, e.g.

$$\hat{y}_i = \beta_0 + \beta_1\left(\frac{\tilde{x}_i - \bar{x}}{\sigma_x}\right)$$

where $\tilde{x}_i$ are values specified for plotting.

For multiple regression

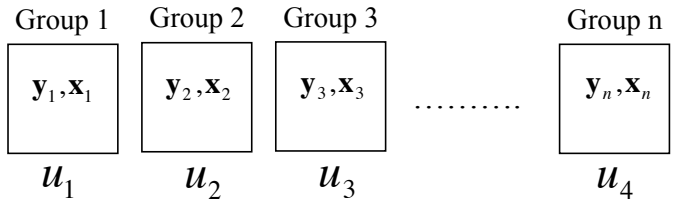
Recovering intercepts:

$$\beta_{0,\text{unstandardized}} = \beta_0 - \sum_{j=1}^{k} \hat{\beta}_j \frac{\bar{x}_j}{\sigma_x}$$

Recovering slopes:

$$\beta_{j,\text{unstandardized}} = \frac{\beta_j}{\sigma_x}$$

Standardizing in multi-level models



- Use with mean and standard deviation across all groups (grand mean centering) to allow intercepts within groups to be defined in terms of means of covariates of all data.
- Use mean and standard deviation within groups when predictions within groups is main objective and you desire borrowing strength provided by multi-level model.

see http://web.pdx.edu/~newsomj/mlrclass/ho_centering.pdf and cited references for details.

Guidance on standardizing

- Most Bayesians standardize predictor variables to speed convergence, to improve interpretation of intercepts, and to allow comparison of slopes.
- It can be useful to know unstandardized intercepts and slopes to allow users to make predictions from data on original scale. This can be done by refitting the model to standardized data or computing back-transforms as derived quantities.

See Hobbs, N. T., H. Andren, J. Persson, M. Aronsson, and G. Chapron. 2012. Native predators reduce harvest of reindeer by Sami pastoralists. Ecological Applications 22:1640-1654.

Interpreting coefficients

What is the interpretation of the coefficients in:

$$\begin{aligned} \mu_i &= \beta_0 + \beta_1 x_i \\ y_i &\sim \text{normal}(\mu_i, \sigma^2) \\ &\times \text{priors} \end{aligned}$$

Interpreting coefficients

What is the interpretation of the coefficients in:

$$\mu_i = \exp(\beta_0 + \beta_1 x_i)$$
$$y_i \sim \text{Poisson}(\mu_i, \sigma^2)$$
$$\times \text{ priors}$$

Interpreting coefficients

What is the interpretation of the coefficients in:

delete sigma^2

$$\mu_i = \text{inverse logit}(\beta_0 + \beta_1 x_i)$$
$$y_i \sim \text{Bernoulli}(\mu_i, \sigma^2)$$
$$\times \text{ priors}$$

Matrix notation for linear models

Remember matrix multiplication?

Example of matrix multiplication for n observations using 2 predictor variables $x_{i,1}$ and $x_{i,2}$ and an intercept.

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

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 (incorrectly, in my view)

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

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])
     }
}
```