Linear models

Liliana M. Dávalos
Ecology & Evolution
9 July 2013

Stony Brook University

Today's lecture

- Summarizing Bayesian results

- Linear models
    - Examples linear models
    - The Bayesian linear model

- Spatial linear models
    - The problem of space
    - How these work

- Hierarchical models
    - Why hierarchy
    - Pooling and not pooling
    - Group-level (sometimes called random) effects
   - Exercise

Summarizing Bayesian results

- Bayesian inference is typically about the **density** or **distribution** of the parameters of interest (rather than just a point estimate) we need new tools

- Using just the mean and variance of a parameter does not summarize the results well in all cases (e.g., multi-modality)

- Small sample inferences may be different

- Typically want to summarize intervals around the mode of the posterior with a given probability coverage

From: Brandt 2013

For example

- Examine the density plotted

- Would the mean and variance be enough to summarize this distribution?

- Why or why not?

- You have 3 min

A Marginal Posterior Distribution
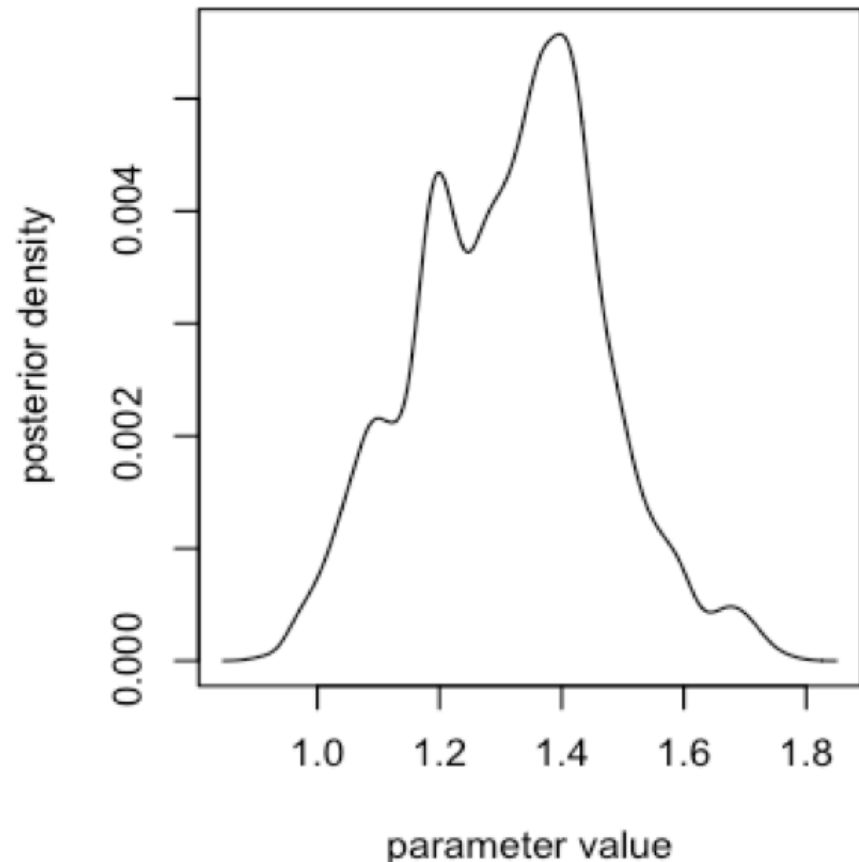

From: Merow & Silander

Credible intervals

- Often computed the same way as a confidence interval

- Need not be contiguous

- Definition is different

- **Credible set**: A $100(1 - \alpha)\%$ credible interval gives the region of the parameter space where the probability of covering θ is equal to $1 - \alpha$
  - Remember θ?

Example credible interval

- Is the credible interval of this posterior distribution contiguous?

- Why or why not
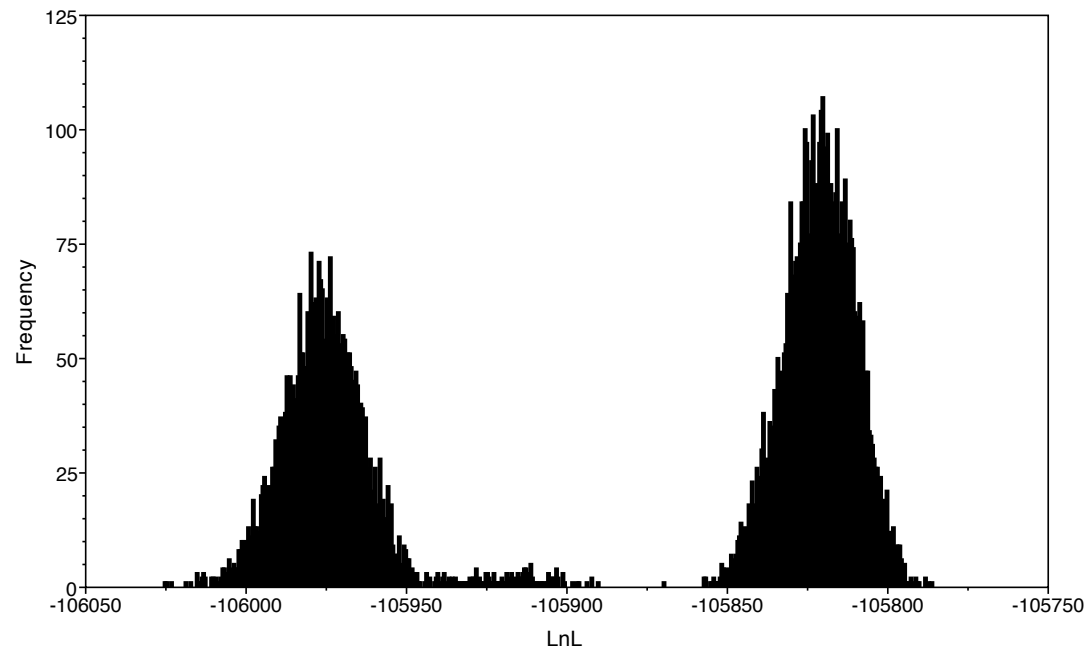
Credible sets versus confidence regions

- Confidence intervals are inherently frequentist:
  - Require repeated **sampling** or **replication** of an experiment to get the region that covers the <u>true</u> parameter $(1 - \alpha)$% of the time <u>on average</u>

- Credible sets are a true probability statement over the parameter space of interest

- Applying the credible set definition to a confidence interval means that the coverage of the confidence interval is either zero or one

From: Brandt 2013

Formal definition

- A credible set, C is the subset of the posterior parameter space Θ such that the $1 - \alpha$ region satisfies

$$1 - \alpha = \int_C p(\theta \mid X) \partial\theta$$

- Note that this is a probability zero or one event with respect to a confidence interval

- This is not a unique quantity unless we center it around some measure of central tendency (mean, mode, median, etc.)

From: Brandt 2013

Notice

- Yesterday we used the notation

$$\Pr[\theta \mid D] = \frac{\Pr[D \mid \theta] \Pr[\theta]}{\Pr[D]}$$

- Today we are using the notation

$$p(\theta \mid X)$$

- To mean the posterior distribution of □ given the data *D*

- To mean the posterior distribution of □ given some observed variables *X*

- **They are the same thing**
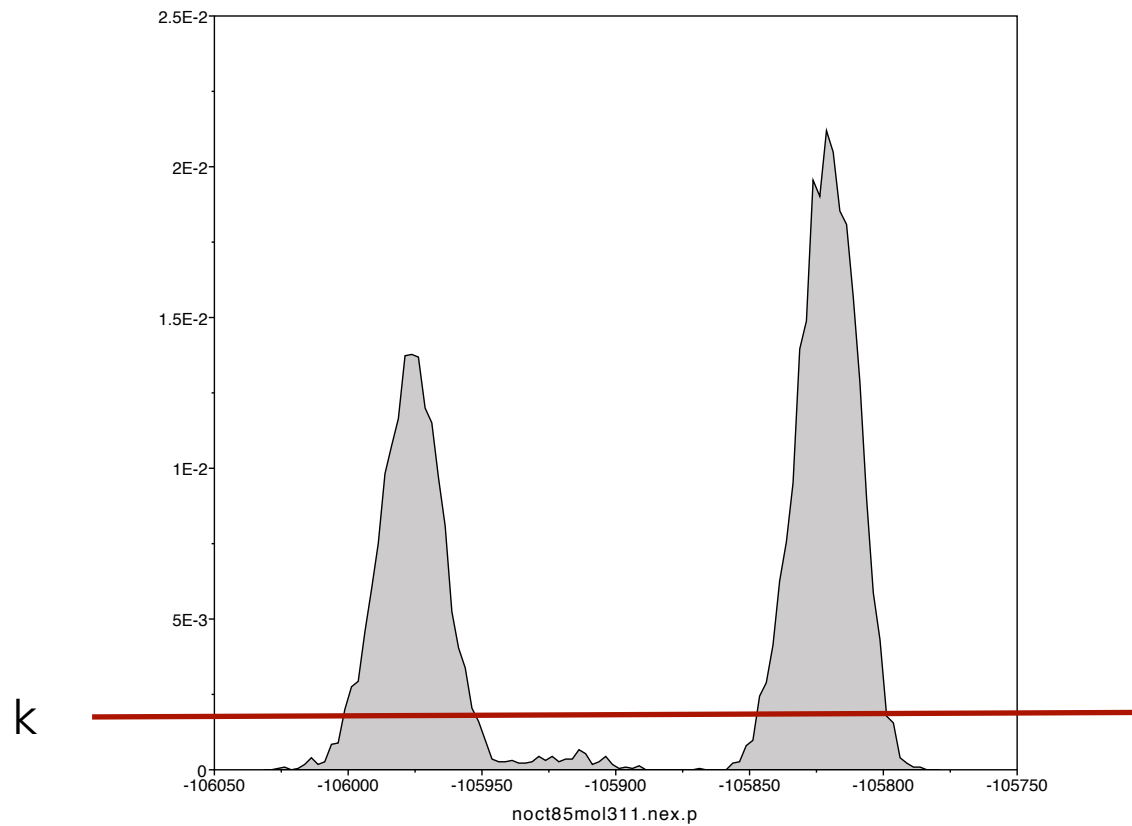
Highest Posterior Density Regions

- Would like to find credible sets with the highest density — the height of the posterior pdf (**probability density function**) should be "highest" inside our credible set

- This is a generalization of the credible set defined in a previous slide

- A $100(1 - \alpha)\%$ HPD is the subset of the support of the posterior for θ where

$$C = 1 - \alpha = \{\theta : \Pr(\theta \mid X) \geq k\}$$

$$1 - \alpha = \int_{\theta : \Pr(\theta \mid X) \geq k} \Pr(\theta \mid X) \partial \theta$$

From: Brandt 2013

Example: bimodal HPD

noct85mol311.nex.p

Computing credible intervals

- If you know the posterior pdf of the parameter, this can be found analytically (most of the time, we don't so best avoid)

- Should consider dropping a horizontal line down the density of θ

- Can be done in R using functions in the **coda** or **hdrcde** packages

- Outside R, download Tracer application, visualizes posteriors, pdf, HPD

From: Brandt 2013

Today's lecture

- Summarizing Bayesian results

- Linear models
  - Examples linear models
  - The Bayesian linear model

- Spatial linear models
  - The problem of space
  - How these work

- Hierarchical models
  - Why hierarchy
  - Pooling and not pooling
  - Group-level (sometimes called random) effects
  - Exercise

Linear regression models

- Linear regression is, perhaps, *the* most widely used statistical modeling tool

- It addresses the following question: How does a quantity of primary interest, *y*, vary as (depend upon) another quantity, or set of quantities, **x**?
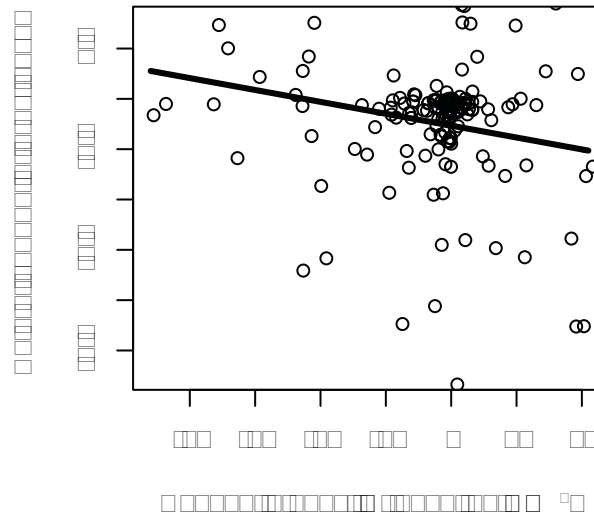
Figure 3

From: Banerjee & Finley 2013

X and Y

- The quantity *y* is called the *response* or *outcome variable*
  - Some people simply refer to it as the *dependent variable*

- The variable(s) **x** are called *explanatory variables, covariates* or simply *independent variables*

- In general, we are interested in the conditional distribution of *y*, given **x,** parameterized as
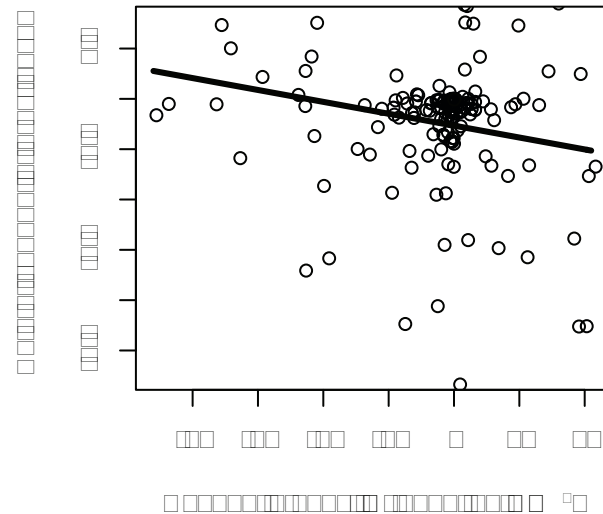 - $p(y \mid \Box, \mathbf{x})$

Figure 3

From: Banerjee & Finley 2013

How linear regression works

- Typically, we have a set of units or experimental subjects i = 1,2,...,n

- For each of these units we have measured an outcome y_i and a set of explanatory variables $\mathbf{x}'_i$ = (1, x_{i1}, x_{i2}, . . . , x_{ip})

- The first element of $\mathbf{x}'_i$ is often taken as 1 to signify the presence of an "intercept"

- In this example we have several explanatory variables up to p

Our variables

- We collect the outcome and explanatory variables into an $n\times1$ vector and an $n\times(p+1)$ matrix:

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}; \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} = \begin{pmatrix} \mathbf{x}_1' \\ \mathbf{x}_2' \\ \vdots \\ \mathbf{x}_n' \end{pmatrix}$$

From: Banerjee & Finley 2013

Importance linear model

- The linear model is the most fundamental of all serious statistical models underpinning:
    - ANOVA: y_i is continuous, x_{ij}'s are all categorical
    - REGRESSION: y_i is continuous, x_{ij}'s are continuous
    - ANCOVA: y_i is continuous, x_{ij}'s are continuous for some j and categorical for others
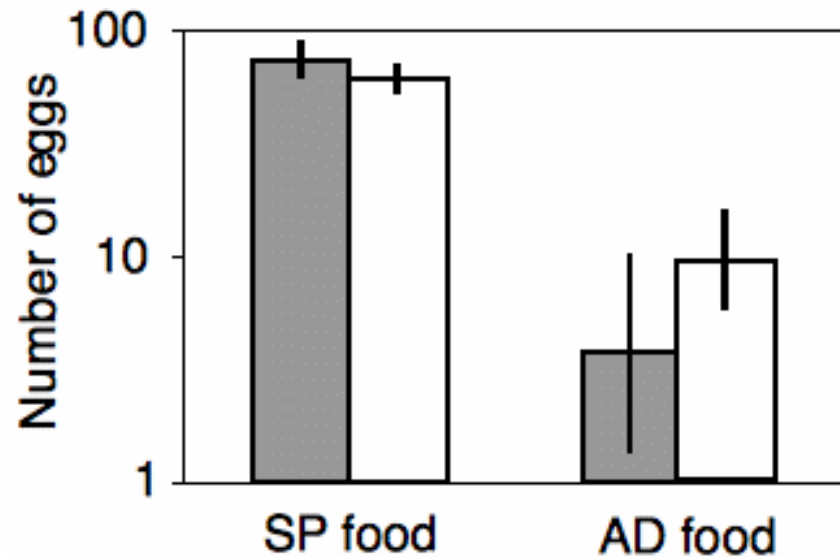
From: Banerjee & Finley 2013

Example one-way anova

- Length of the anterior adductor muscle scar divided by total length in *Mytilus trossulus*

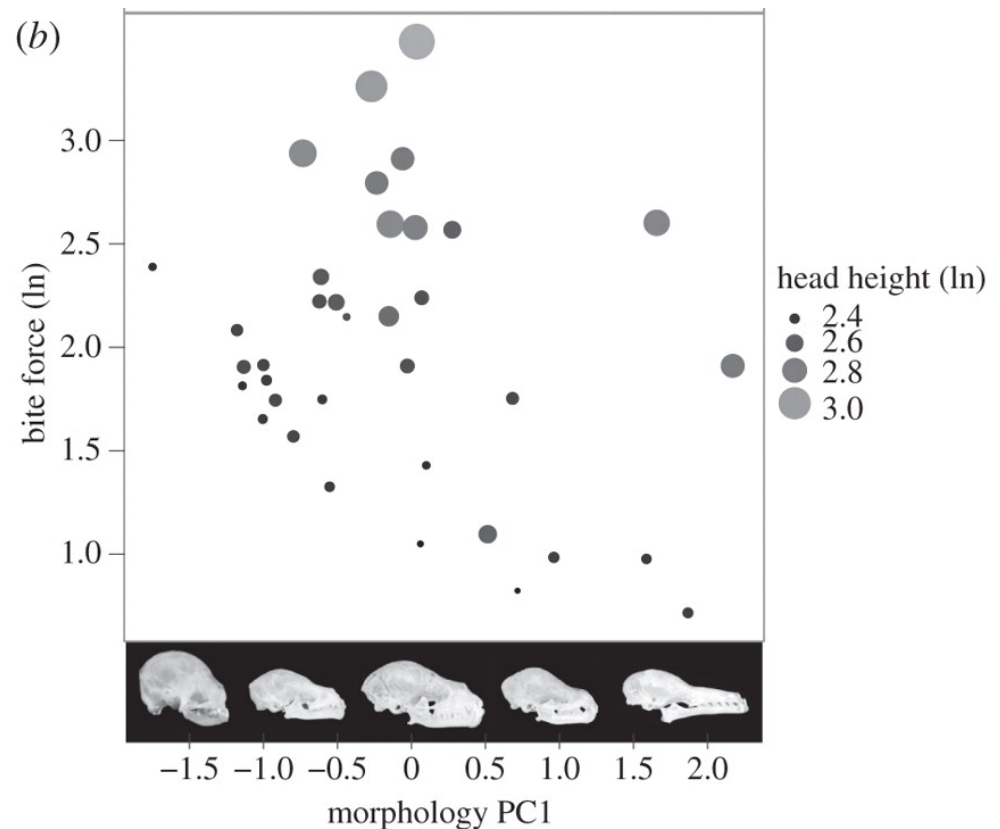- Means ± one standard error are shown for five locations.
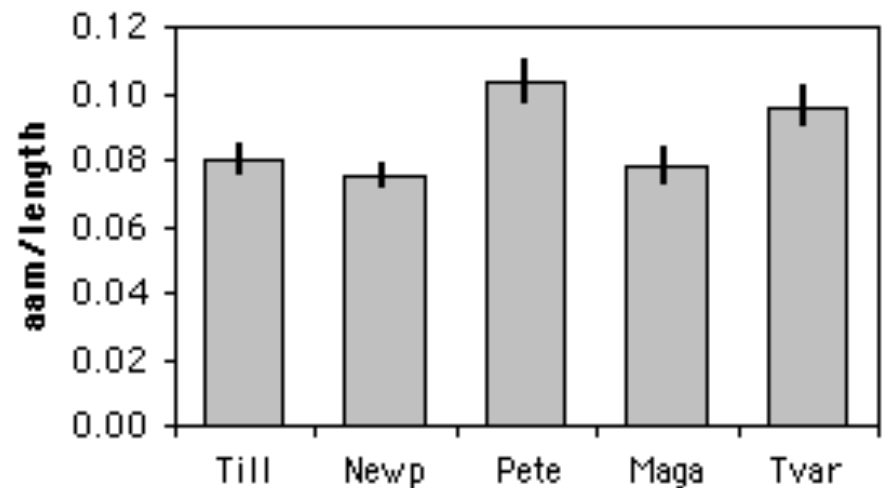
Two-way anova

- Mean total numbers of eggs of females from the SP strain (gray bars) and AD strain (white bars)

- Values are mean ± SEM

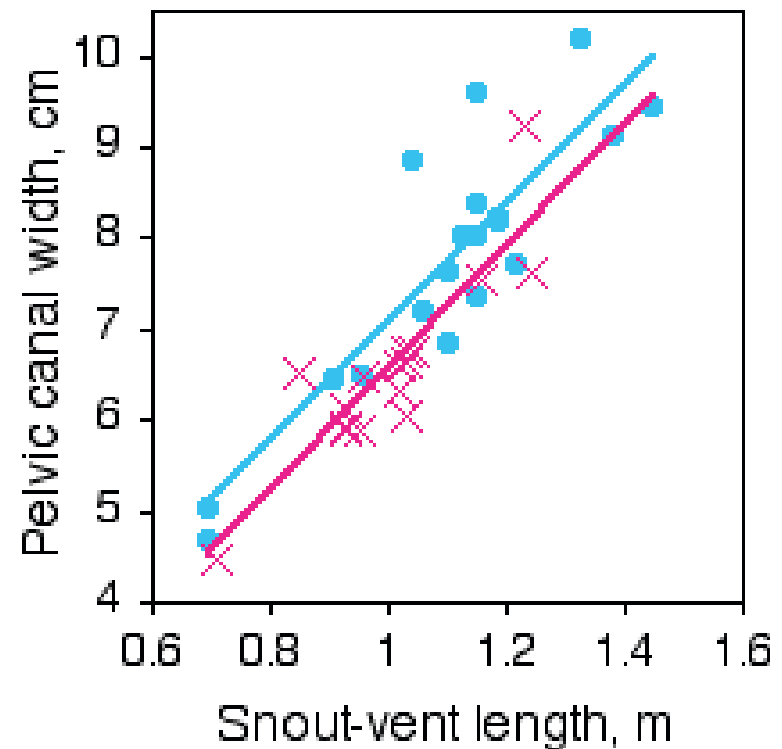- Adapted from Fig. 4 of Shimoji and Miyatake [2002]

Example regression

- Morphology PC1 as a predictor of ln bite force (*b*, mean coefficient = −0.327 ± 0.007)

- From Dumont, Dávalos et al 2012

Example ancova

- Pelvic canal width vs. snout-vent length in the American alligator

- Blue circles and line are males; pink X's and line are females

- From Prieto-Marquez et al. 2007

Think, pair, share

- Let's discuss for 2 min with neighbor
  - Propose a **linear** model you want to build using Bayesian methods
  - What is the **dependent** variable?
  - What is the **independent** variable(s)?
  - Does it involve anova, regression, or ancova?

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}; X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} = \begin{pmatrix} x_1' \\ x_2' \\ \vdots \\ x_n' \end{pmatrix}$$

Bayesian linear model

- The Bayesian linear model is given by:

$$y_i \mid \mu_i, \sigma^2, X \overset{ind}{\sim} N(\mu_i, \sigma^2); i = 1, 2, \dots n;$$

$$\mu_i = \beta_0 + \beta_1 x_i + \dots + \beta_p x_{ip} = x_i' \beta; \beta = (\beta_0, \beta_1, \dots \beta_p);$$

$$\beta, \sigma^2 \mid X \sim p(\beta, \sigma^2 \mid X)$$

- Unknown parameters include the regression parameters and the variance, i.e. $\theta = \{\beta, \sigma^2\}$

- $p(\beta, \sigma^2 \mid \mathbf{X}) \equiv p(\theta \mid \mathbf{X})$ is the joint *prior* on the parameters. We assume $\mathbf{X}$ is observed *without error* and all inference is conditional on $\mathbf{X}$

- We suppress dependence on $\mathbf{X}$ in subsequent notation

From: Banerjee & Finley 2013

Bayesian linear model

- Specifying $p(\beta, \sigma^2)$ completes the model

- All inference proceeds from $p(\beta, \sigma^2 | \mathbf{y})$

- With no prior information, we specify

  Or its equivalent

$$p(\beta, \sigma^2) \propto \frac{1}{\sigma^2} \qquad\qquad p(\beta) \propto 1; \, p(\log(\sigma^2)) \propto 1$$

- The above is **NOT** a probability density (they do not integrate to any finite number)

- So why is it that we are even discussing them?

- Even if the priors are *improper,* as long as the resulting posterior distributions are valid we can still conduct legitimate statistical inference on them

From: Banerjee & Finley 2013

Practical questions

- Look at yesterday's **jags** model

- What is the prior on σ^2?

- What is the prior on β ?

- Was the prior specified as in:

$$p(\beta, \sigma^2) \propto \frac{1}{\sigma^2}$$

- Or as in:

$$p(\beta) \propto 1; p(\log(\sigma^2)) \propto 1$$

Computing the posterior distribution

- Strategy: Factor the **joint** posterior distribution for β and σ^2 as:

$$p(\beta, \sigma^2 \mid y) = p(\beta \mid \sigma^2, y) \times p(\sigma^2 \mid y)$$

- The *conditional posterior* distribution of β, given σ^2:

$$\beta \mid \sigma^2, y \sim N(\hat{\beta}, \sigma^2 V_\beta)$$

- where, using some algebra, one finds: $\hat{\beta} = (X'X)^{-1}X'y$

- and $V_\beta = (X'X)^{-1}$

From: Banerjee & Finley 2013

Computing the posterior distribution

- The *marginal posterior* distribution of σ^2: Let $k = (p + 1)$ be the number of columns of **X**

$$\sigma^2 \mid \mathrm{y} \sim IG\left(\frac{n-k}{2}, \frac{(n-k)s^2}{2}\right),$$

- Where

$$s^2 = \frac{1}{n-k}(\mathrm{y} - \mathrm{X}\hat{\beta})'(\mathrm{y} - \mathrm{X}\hat{\beta})$$

- is the classical unbiased estimate of σ^2 in the linear regression model

Questions

- What are the columns of **X**?

- Why are we interested in the marginal posterior? (As opposed to the joint posterior)?

- Given that $$s^2 \propto \frac{1}{n-k}$$

- what can we do to reduce variance in our models?

Computing the posterior distribution 2

- The *marginal posterior* distribution $p(\beta \mid \mathbf{y})$, averaging over σ^2, is *multivariate t* with $n - k$ degrees of freedom

- But we rarely use this fact in practice

- Instead, we *sample* from the posterior distribution

From: Banerjee & Finley 2013

Algorithm for sampling from the posterior distribution

- We draw samples from $p(\beta, \sigma 2 \mid \mathbf{y})$ by executing the following steps:

- Step 1: Compute $\hat{\beta}$ and $\mathbf{V}_\beta$

- Step 2: Compute s^2

- Step 3: Draw M samples from $p(\sigma^2 \mid \mathbf{y})$:

$$\sigma^{2(j)} \sim IG\left(\frac{n-k}{2}, \frac{(n-k)s^2}{2}\right), j = 1\ldots M$$

- Step 4: For $j = 1,\ldots,M$, draw $\beta^{(j)}$ from $p(\beta \mid \sigma^{2(j)}, \mathbf{y})$:

$$\beta^{(j)} \sim N\left(\hat{\beta}, \sigma^{2(j)}\mathbf{V}\beta\right)$$

From: Banerjee & Finley 2013

Sampling from the posterior distribution

- The marginal distribution of each individual regression parameter β_j is a non-central univariate t_{n-p} distribution

- In fact,

$$\frac{\beta_j - \hat{\beta}_j}{s\sqrt{V_{\beta;jj}}} \sim t_{n-p}$$

- The 95% credible interval for each β_j is constructed from the quantiles of the t-distribution

- This exactly coincides with the 95% classical confidence intervals, but the interpretation is direct: the probability of β_j falling in that interval, given the observed data, is 0.95

From: Banerjee & Finley 2013

Today's lecture

- Summarizing Bayesian results

- Linear models
  - Examples linear models
  - The Bayesian linear model

- Spatial linear models
  - The problem of space
  - How these work

- Hierarchical models
  - Why hierarchy
  - Pooling and not pooling
  - Group-level (sometimes called random) effects
 - Exercise

The first law of geography

- First law of geography:

- *"Everything is related to everything else, but near things are more related than distant things.''*
Tobler 1970

Relief and rivers
— Navigable rivers

B

Quick discussion

- Which one of the variables on the right shows the most aggregated spatial pattern?

- Which one shows the least?

- How important do you think these spatial patterns are for modeling this landscape?
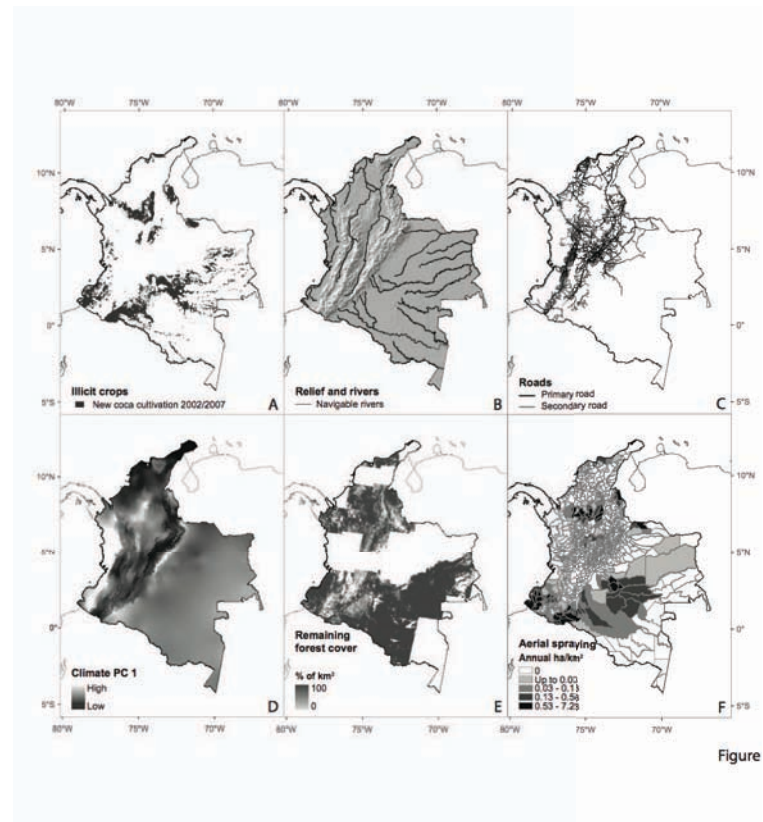

Figure

Spatial models

- Point-level modeling refers to modeling of spatial data collected at locations referenced by coordinates (e.g., lat- long, Easting-Northing)

- **Fundamental concept**: Data from a spatial process $\{Y(\mathbf{s}) : s \in D\}$, where D represents some landscape

- **Example**: $Y(s)$ is a pollutant level at site s

- **Conceptually**: Pollutant level exists at all possible sites

- **Practically**: Data will be a partial realization of a spatial process – observed at $\{s_1, \ldots, s_n\}$

- **Statistical objectives**: Inference about the process $Y(\mathbf{s})$; predict at new locations

From: Merow & Silander

Simple linear model

- $Y(\mathbf{s}) = \square(\mathbf{s}) + \square(\mathbf{s})$

- Assumptions regarding $\square(\mathbf{s})$:

- $\square(\mathbf{s}) \overset{iid}{\sim} N(0, \square^2)$

- $\square(\mathbf{s}_i)$ and $\square(\mathbf{s}_j)$ are uncorrelated for all $i \neq j$
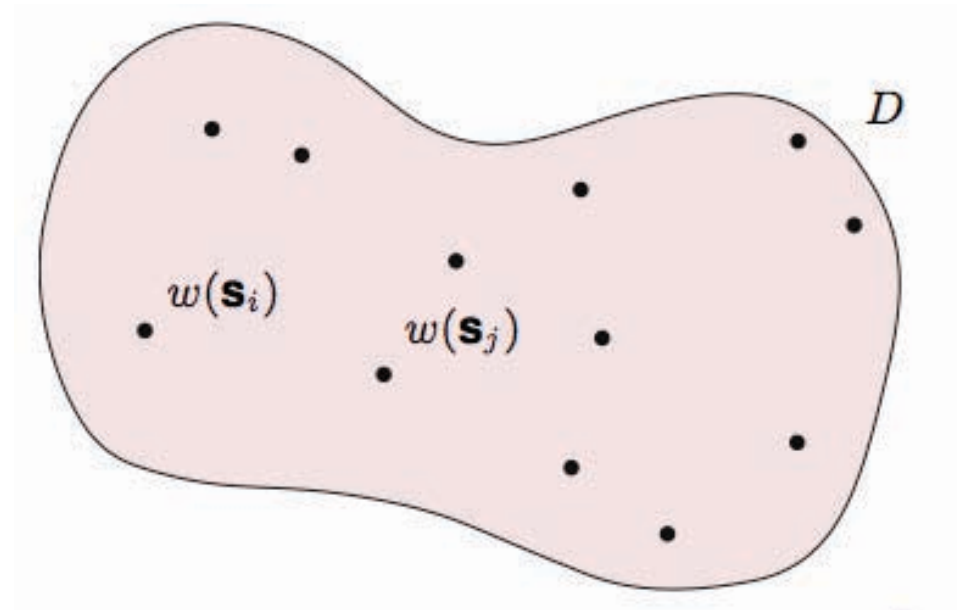

From: Banerjee & Finley 2013
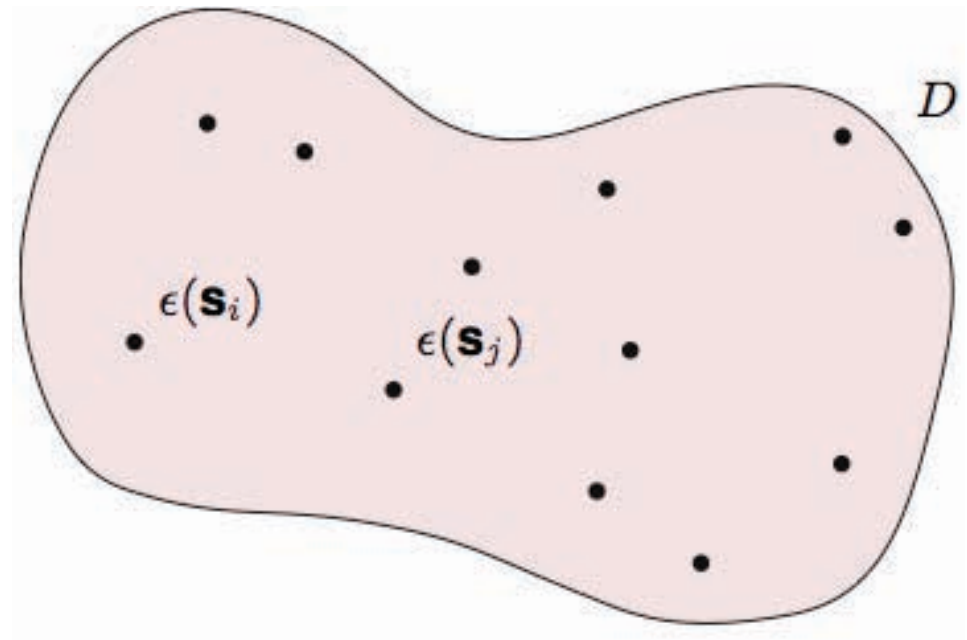
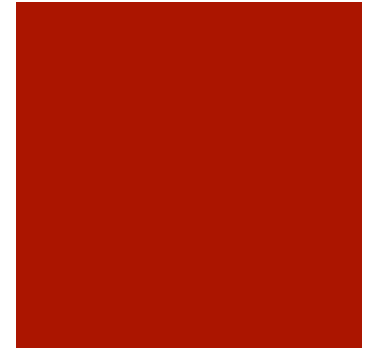# What does spatial autocorrelation look like?
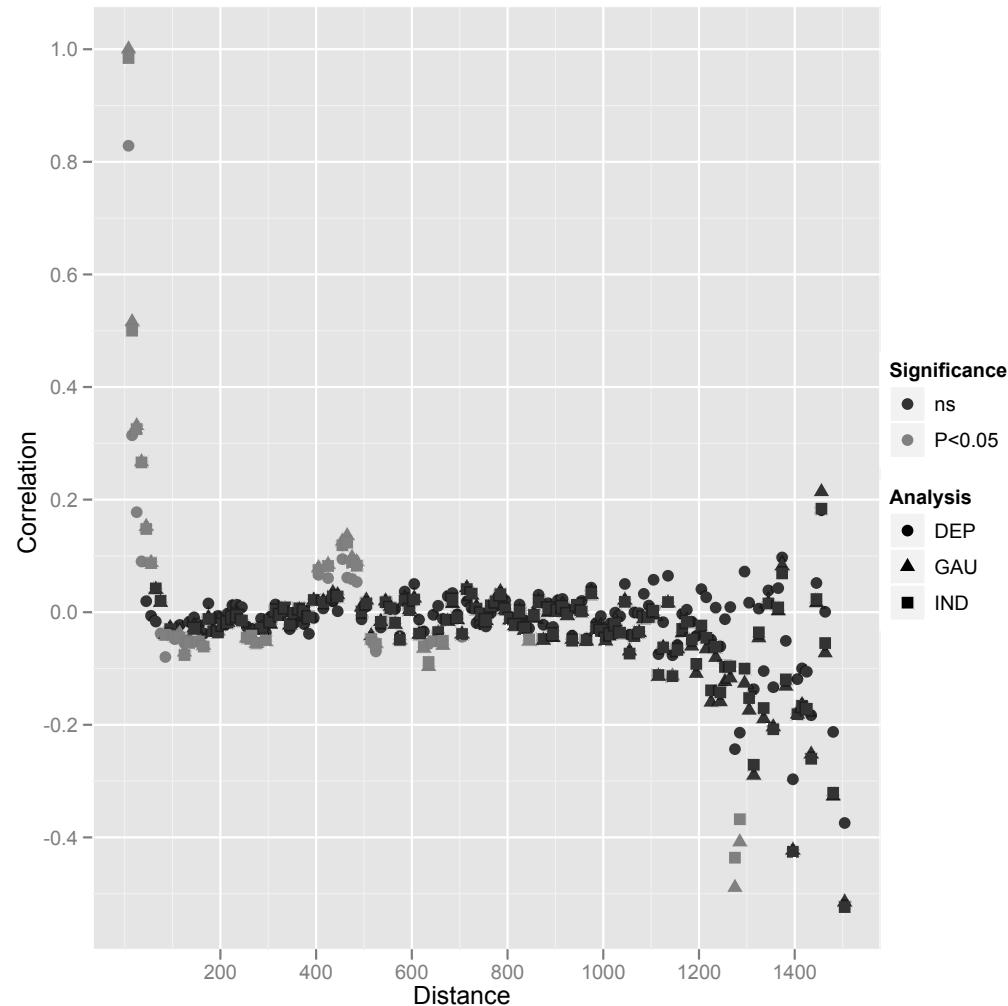
Spatial Gaussian process

- Let $\mathbf{w} = [w(\mathbf{s}_i)]^n_{i=1}$, then

- $\mathbf{w} \sim N(\mathbf{0}, \sigma^2 R(\phi))$,

- Where

- $R(\phi) = [\rho(\phi; |\mathbf{s}_i - \mathbf{s}_j|)]^n_{i,j=1}$

From: Banerjee & Finley 2013

Realization of a Gaussian process

ϕ is the distance

• decay of the spatial process

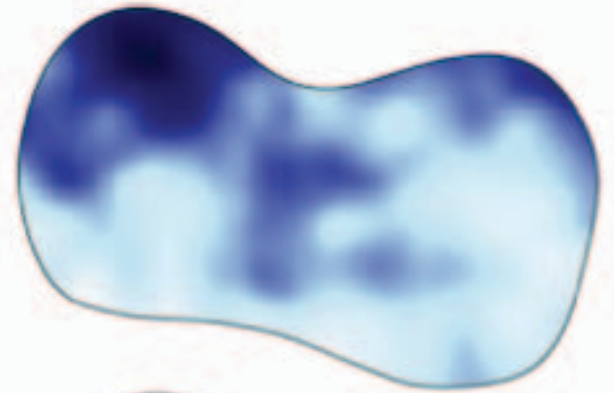- Changing ϕ and holding $\sigma^2 = 1$:

- $\mathbf{w} \sim N(\mathbf{0}, \sigma^2 R(\phi))$,

- Where $R(\phi) = [\rho(\phi; |\mathbf{s}_i - \mathbf{s}_j|)]^n_{i,j=1}$

- Correlation model for $R(\phi)$: e.g., exponential decay

- $\rho(\phi; t) = \exp(-\phi t)$ if $t > 0$

- Other **valid** models e.g., Gaussian, Spherical, Matérn

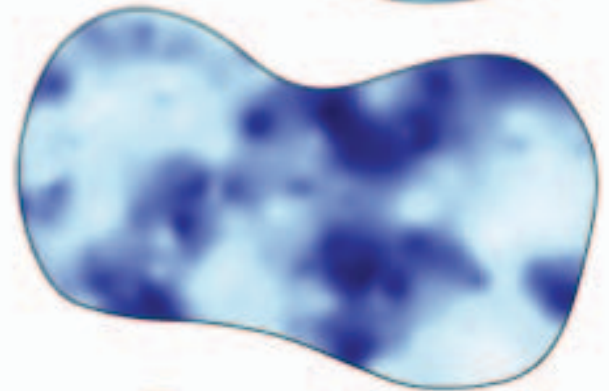- **Effective range**, $t_0 = \ln(0.05)/\phi \approx 3/\phi$

Realization of a Gaussian process

- **Effective range**,
  $t_0 = \ln(0.05)/\square \approx 3/\square$

☐ small

☐ medium

☐ large

From: Banerjee & Finley 2013

Simple linear model + random spatial effects

- $Y(\mathbf{s}) = \mu(\mathbf{s}) + w(\mathbf{s}) + \varepsilon(\mathbf{s})$

- Response: $Y(\mathbf{s})$ at some site

- Mean: $\mu = \mathbf{x}^\top(\mathbf{s})\,\beta$

- Spatial random effects: $w(\mathbf{s}) \sim GP(0,\ \sigma^2 \rho(\phi;\ |\mathbf{s}_1 - \mathbf{s}_2|))$

- Non-spatial variance: $\varepsilon(\mathbf{s}) \overset{iid}{\sim} N(0,\ \tau^2)$

This is the Gaussian process

Non spatial variance is here

From: Banerjee & Finley 2013

Hierarchical modeling

- First stage:

$$y \mid \beta, w, \tau^2 = \prod_{i=1}^{n} N(Y(s_i) \mid xT(s_i)\beta + w(s_i), \tau^2)$$

- Second stage:

$$w \mid \sigma^2, \phi \sim N(0, \sigma^2 R(\phi))$$

- Third stage: priors on $\Omega = (\beta, \tau^2, \sigma^2, \phi)$

From: Banerjee & Finley 2013

To summarize, the parameters we want are:

- β = regression coefficients

- τ^2 = residual variance

- σ^2 = spatial variance

- ϕ = decay of spatial dependence

- Together, σ^2 and ϕ define the spatial weights, **w**

- Usually we want to look at the surface of spatial weights to see if there are any latent patterns
  - (e.g., indications of a missing variable)

From: Merow & Silander

Colorado data illustration

- Modeling temperature: 507 locations in Colorado

- Simple spatial regression model:

$$Y(\mathrm{s}) = \mathrm{x}^{T}(\mathrm{s})\beta + w(\mathrm{s}) + \varepsilon(\mathrm{s})$$

$$w(\mathrm{s}) \sim GP(0, \sigma^{2}\rho(\cdot; \phi, v));$$

$$\varepsilon(\mathrm{s}) \overset{iid}{\sim} N(0, \tau^{2})$$

Parameters	50% (2.5%,97.5%)
intercept	2.827 (2.131,3.866)
[elevation]	-0.426 (-0.527,-0.333)
precipitation	0.037 (0.002,0.072)
σ^2	0.134 (0.051, 1.245)
ϕ	7.39E-3 (4.71E-3, 51.21E-3)
Range	278.2 (38.8, 476.3)
τ^2	0.051 (0.022, 0.092)

From: Banerjee & Finley 2013

Spatial model example

Elevation
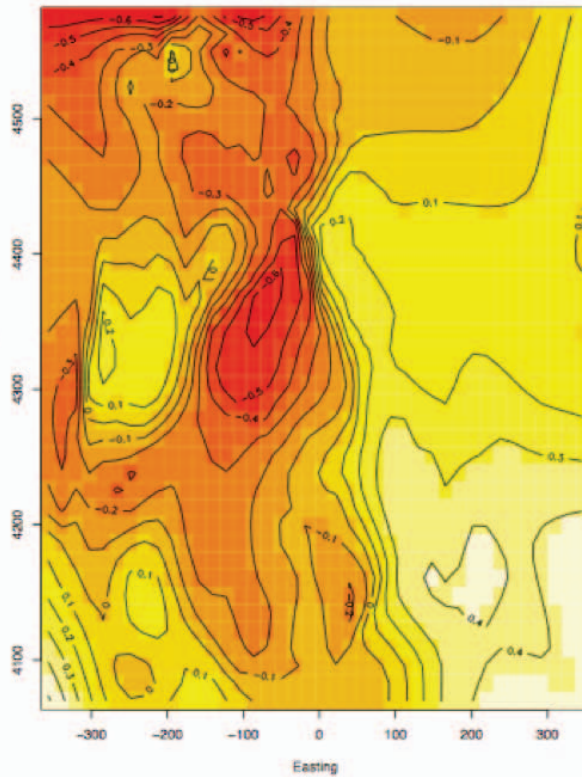
Residual without elevation
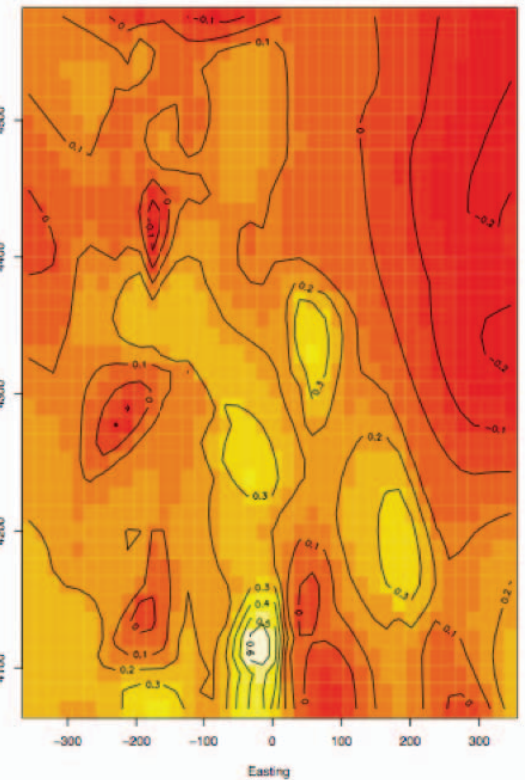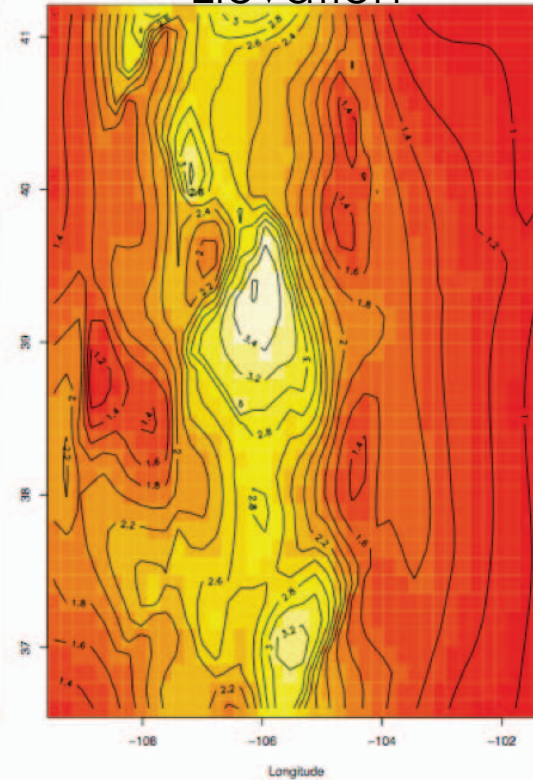
Residual with elevation

From: Banerjee & Finley 2013

Spatial random effects

- Spatial random effects can:
  - Account for spatial autocorrelation (samples aren't independent)
  - Ensure that your regression coefficients aren't contorted to accommodate patterns that aren't actually related to the predictor
  - Identify spatial patterns in residuals that are related to omitted predictors

From: Banerjee & Finley 2013

Today's lecture

- Summarizing Bayesian results

- Linear models
  - Examples linear models
  - The Bayesian linear model

- Spatial linear models
  - The problem of space
  - How these work

- Hierarchical models
  - Why hierarchy
  - Pooling and not pooling
  - Group-level (sometimes called random) effects
  - Exercise

Why use hierarchical models?

- Account for sources of uncertainty/variability
  - Account for individual-level and group-level variation when estimating group-level coefficients
  - Classical models require averaging over individual-level variation

- Borrow strength across groups (minimize effects of small sample sizes in some groups)

From: Merow & Silander

Group-level coefficients

- Group-level effects absorb variation that's NOT related to the fixed effects so you potentially get lower bias for estimates of population-level effects

- Group-level effects pick up unmeasured variation

- It's probably never worse than non-hierarchical models because the worst case is that you find among group variation to be irrelevant

From: Merow & Silander

A note about terms

- Hierarchical models are often called *mixed effects models*
  - *Mixed effects = fixed + random effects*
    - Fixed effects: coefficients that apply to all observations
    - Random effects: coefficients that apply to individual clusters
  - Problem is "fixed" and "random" are used with various meanings in statistics and social science

Note continued

- I follow Andrew Gelman and call these models *hierarchical*
  - Hierarchical models = population-level effects + group-level effects
  - Population-level: affecting all observations
  - Group-level: affecting only observations in a given cluster

- You will see how this works later

What is a hierarchical model?

- Hierarchical models are used when the data are structured in groups. e.g. demographically, temporally, spatially

- Different (but related) parameters are used for each group

- These group-level parameters can be interpreted to describe group-level differences for predictors we DIDN'T measure

From: Merow & Silander

Groups

- Types of groups
  - Clustered data: when modeling individual plant growth and comparing among populations, some observations are made on individuals (e.g. stem length) while some are made on populations (e.g. temperature, soil)
    - E.g., fit a matrix demographic model and use population random effects on probability of seedling survival
  - Repeated measures data: Time effects - e.g. in a mark recapture model survival might vary by year
  - Spatial effects – in a species distribution model, latent, spatially autocorrelated patterns – cells are more like their neighbors
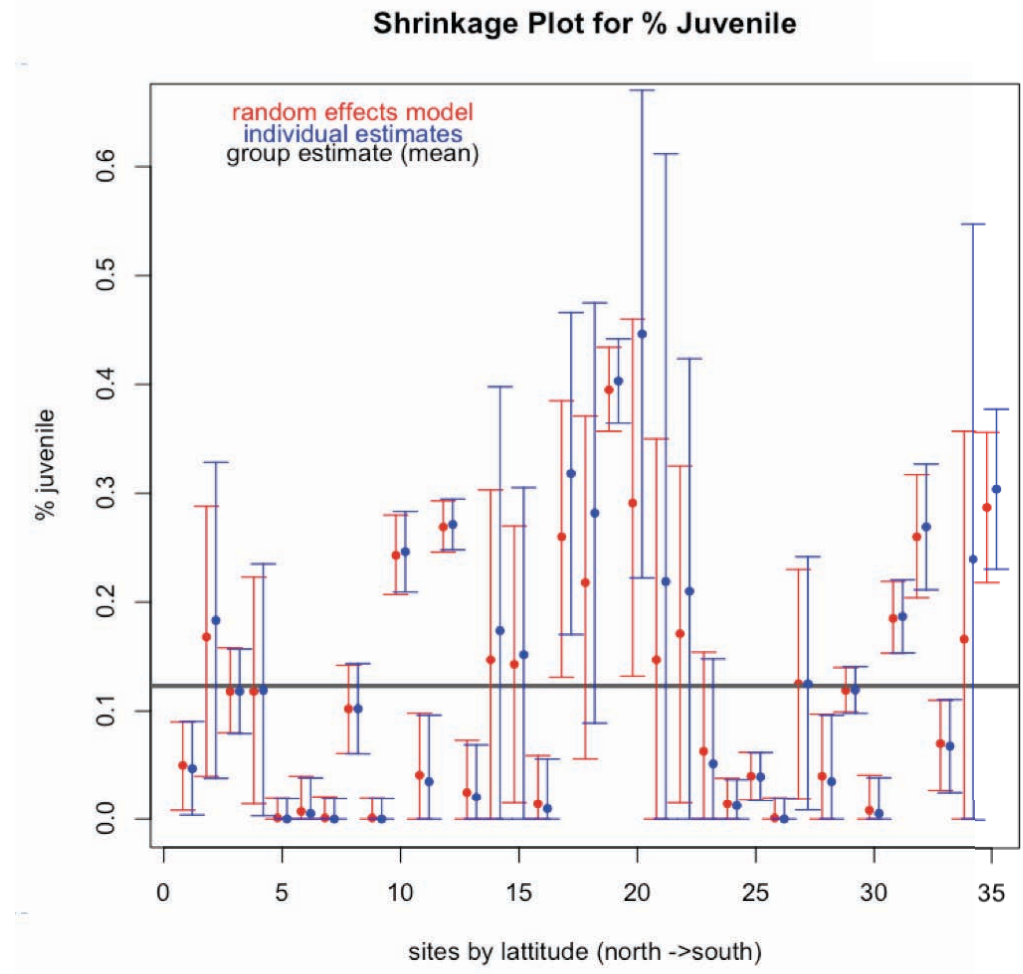
From: Merow & Silander

How to model hierarchical parameters?

- Consider the three alternatives for modeling groups
  - Pool all groups together
    - Ignore latent differences
    - Ignore autocorrelation (observations not independent)

- Model each group separately
  - What about small groups?
  - Many more parameters to estimate
  - Ignores latent similarity

- Model each group hierarchically
  - Groups are similar but different
  - Intermediate number of parameters

From: Merow & Silander

Comparison of pooling alternatives
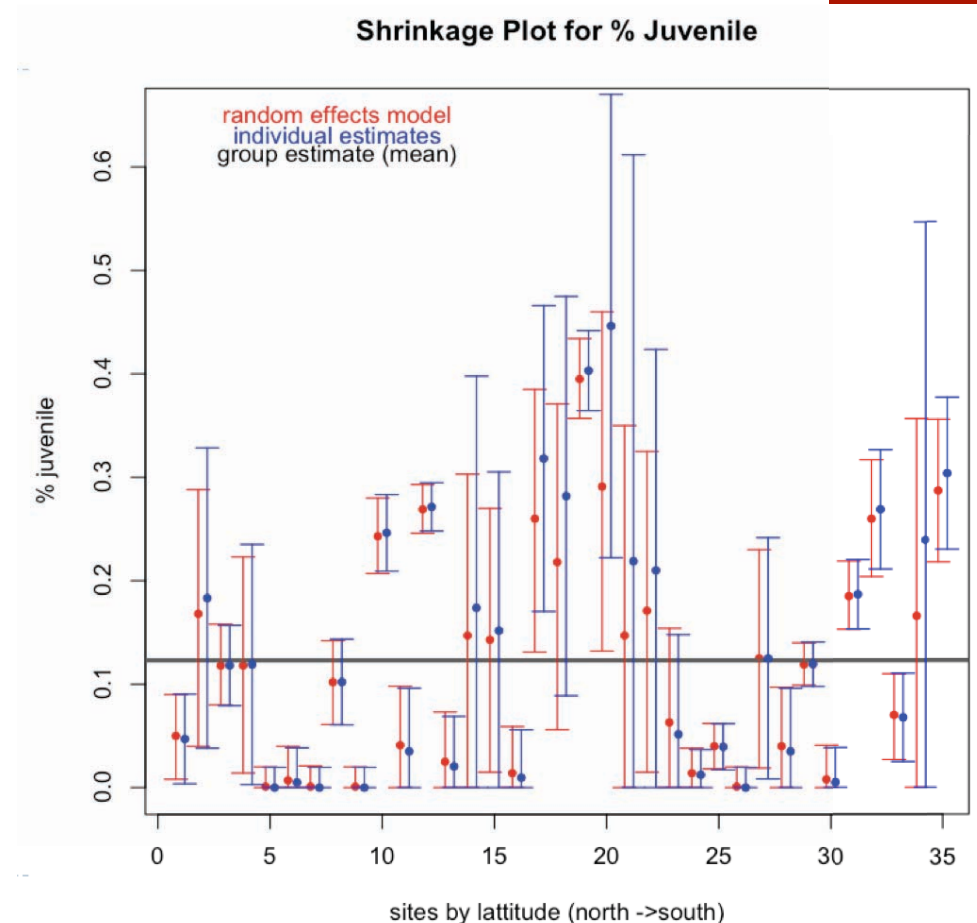
- Proportion of juvenile aloes in 35 populations spanning the species' range

- Multiple samples from each population

- Combined estimates across samples from different years to estimate the variance in each population

Shrinkage Plot for % Juvenile

random effects model
individual estimates
group estimate (mean)

% juvenile

sites by lattitude (north ->south)

From: Merow & Silander

Questions

- Which estimates are closest to the observations?

- Name 1 advantage of the group estimate
  - Name 1 disadvantage

- Name 1 advantage of the individual estimates
  - Name 1 disadvantage

Shrinkage Plot for % Juvenile

random effects model
individual estimates
group estimate (mean)

% juvenile

sites by lattitude (north ->south)

From: Merow & Silander

Hierarchical regression

- Regressions have intercept and slope terms

- Hierarchical regressions have different intercepts or slopes (or both) for each group

- Group level parameters do not vary independently, but are constrained by a distribution – often a normal with a variance that we model

From: Merow & Silander

Example hierarchical regression

- 30 young rats, weights measured weekly for five weeks

- Dependent variable (Y_{ij}) is weight for rat "i" at week "j"

- Hierarchical: weights (observations) within rats (groups)

- Data:

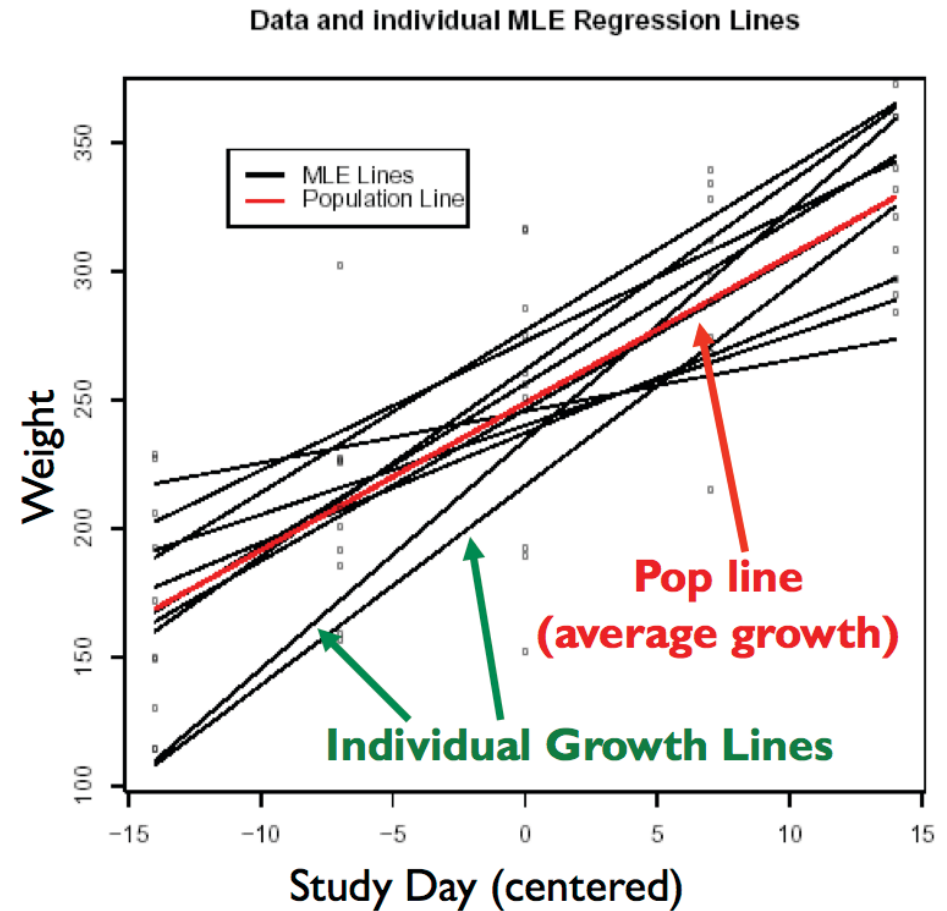| Weights Y_{ij} of rat i on day x_j					
$x_j = 8$	15	22	29	36
Rat 1	151	199	246	283	320
Rat 2	145	199	249	293	354
....					
Rat 30	153	200	244	286	324

From: Gelfand et al. 1990

Options

- Possible analyses
  - Each rat (cluster) has its own line:
    - Intercept=b_{i0}, slope=b_{i1}
  - All rats follow the same line:
    - $b_{i0}=\beta_0$, $b_{i1}=\beta_1$
  - A compromise between these two:
    - Each rat has its own line, BUT…
    - The lines come from a common assumed distribution (a slope and intercept are estimated from each rat)

Rats!

- NON-hierarchical options

- Rat "i" has its own expected growth line:

- $E(Y_{ij}) = b_{0i} + b_{1i}X_j$

- There is also an overall, average population growth line:

- $E(Y_{ij}) = \square_0 + \square_1 X_j$

Data and individual MLE Regression Lines

Pop line (average growth)

Individual Growth Lines

From: Merow & Silander

The hierarchical option

- A compromise between pooling and individual estimates:

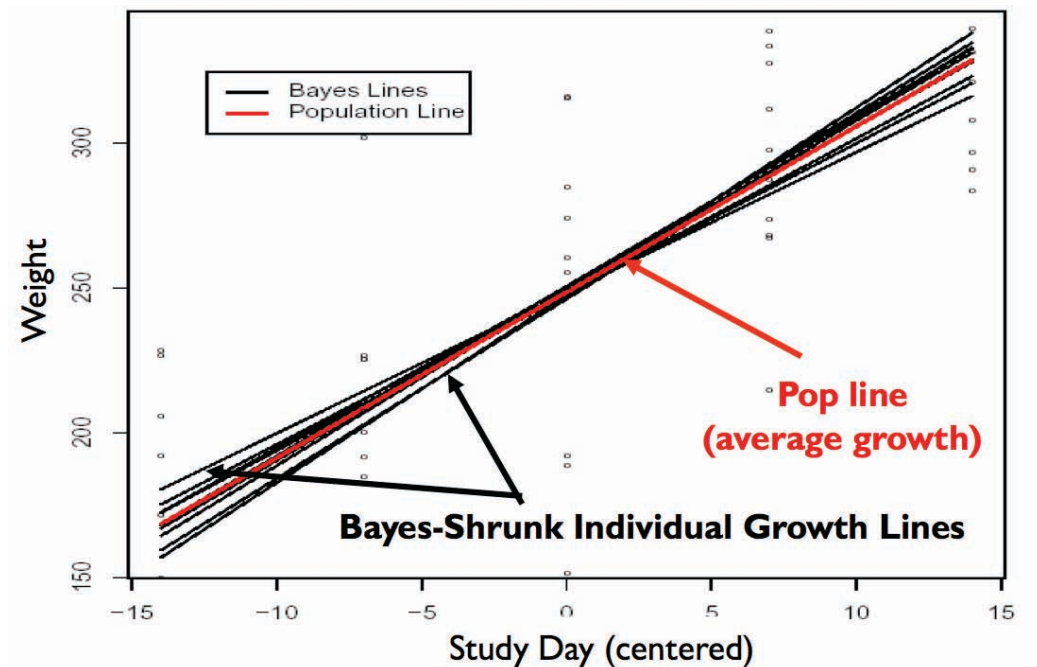$$Y_{ij} \sim \text{Normal}(\alpha_i + \beta_i(x_j - \bar{x}), \tau_c)$$

$$\alpha_i \sim \text{Normal}(\alpha_c, \tau_\alpha)$$

$$\beta_i \sim \text{Normal}(\beta_c, \tau_\beta)$$

- $\alpha_c, \tau_\alpha, \beta_c, \tau_\beta, \tau_c$ are given independent, uninformative priors

From: Merow & Silander
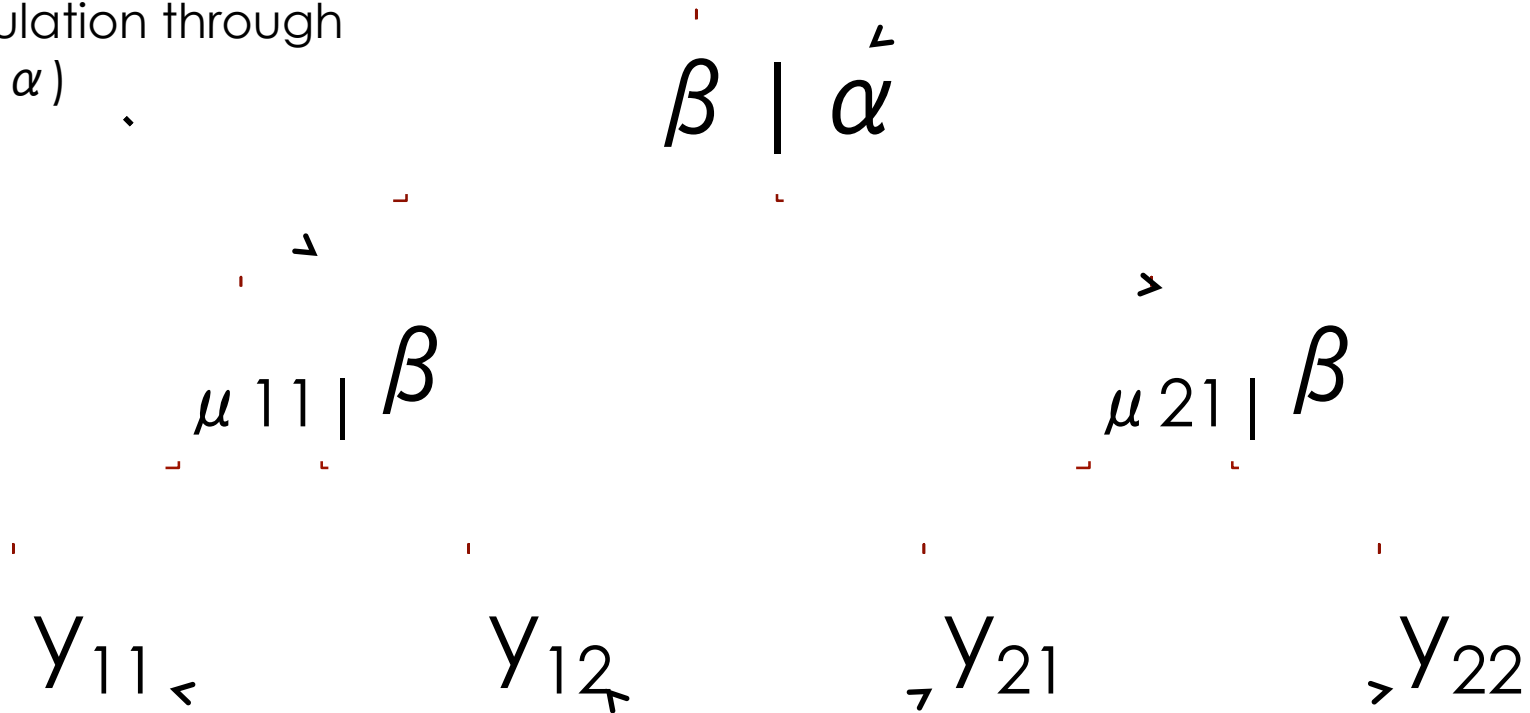
The hierarchical model

- A compromise: Each rat has its own line, but information is borrowed across rats to tell us about individual rat growth

Summary: Visualizing a hierarchical model

Requires a hyperprior
• distribution
$p(\alpha)$

Priors for each sub-population through
$p(\beta | \alpha)$

$$\beta | \alpha$$

$$\mu_{11} | \beta \qquad\qquad \mu_{21} | \beta$$

$$y_{11} \qquad y_{12} \qquad y_{21} \qquad y_{22}$$

Data

From: Merow & Silander

What are the levels?

$\beta \mid \alpha$

$\mu_{11} \mid \beta_1$ $\qquad$ $\mu_{21} \mid \beta_2$

y_{11} $\qquad$ y_{12} $\qquad$ y_{21} $\qquad$ y_{22}

- Which ones are population-level parameters?

- Which ones are group-level parameters?
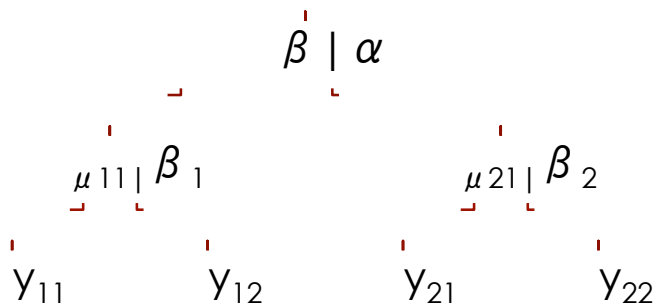
Bayesian inference of hierarchical model

- Non hierarchical model

$$p(\theta_1, \theta_2 \mid x, y) \propto p(y \mid \theta_1, \theta_2, x) \quad \text{Likelihood (data model)}$$

$$\times p(\theta_1) p(\theta_2) \quad \text{Prior (parameter model)}$$

- Hierarchical model

$$p(\theta_1, \theta_2, \theta_3 \mid x, y) \propto p(y \mid \theta_1, \theta_2, x) \quad \text{Likelihood}$$

$$\times p(\theta_1 \mid \theta_3) p(\theta_2) \quad \text{Prior}$$

$$\times p(\theta_3) \quad \text{Hyperprior}$$

From: Merow & Silander

Today's lecture

- Summarizing Bayesian results

- Linear models
  - Examples linear models
  - The Bayesian linear model

- Spatial linear models
  - The problem of space
  - How these work

- Hierarchical models
  - Why hierarchy
  - Pooling and not pooling
  - Group-level (sometimes called random) effects
- Exercise

Exercise: Group means

- Data structure
  - J groups, with means θ_j, each with n_j observations
  - y_{ij} = i^{th} measurement from the jth group variance of each group is σ^2_j (assumed known)

- Model structure
  - Estimate θ_j for each of the J groups
  - Estimate μ, the group mean
  - Estimate τ, the variance of the θ_j

Model structure

- Data structure
  - J groups, with means θ_j, each with n_j observations
  - y_{ij} = i^{th} measurement from the jth group
  - variance of each group is σ^2_j (assumed known)
  - ε_{ij} = error $\varepsilon_{ij} \sim N(0, \sigma^2)$

- Model structure
  - Estimate θ_j for each of the J groups
  - Estimate μ, the group mean
  - Estimate τ, the variance of the θ_j

From: Merow & Silander

Model structure

- All groups the same
  - $y_{ij} = \mu + \varepsilon_{ij}$

- Each group unique
  - $y_{ij} = \theta_j + \varepsilon_{ij}$

- Compromise: Groups similar but different
  - $y_{ij} = \theta_j + \varepsilon_{ij}$, with: $\theta_j \sim N(\mu, \tau^2)$

From: Merow & Silander
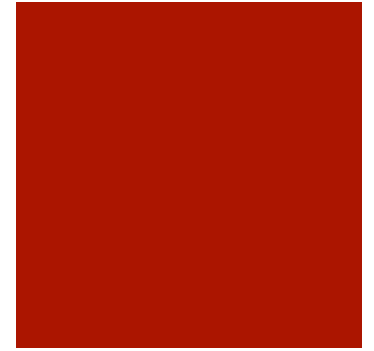
Exercise: Hierarchical Normal Model

- Model from Gelman 2003

- Estimate the proportion of juveniles and dead individuals of *Aloe dichotoma* across their range for a matrix demographic model

- 35 populations, each with a subset of ~100 individuals measured 6 years apart

- Instructions in `hierarchical_exercise.R`

- THE END