# Hierarchical Bayesian Modeling

## Overview and Model Specifications

Hierarchical models (also known as multilevel models or mixed effects models) are an extension of linear modeling which account for a hierarchical structure in the data such as ecosubsections within ecosections or even provinces. This thesis implements Bayesian hierarchical models to accomplish small area estimation on forest attributes, meaning that these models do not only take into account the hierarchical structure of the data, but they also incorporate prior information about the explanatory variables in the model. To best understand a Bayesian hierarchical model, I will first begin with the standard OLS regression form. We have

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon_i \\ &= \vec{\beta} X_i + \epsilon_i \end{aligned}$$

where $Y_i$ is the response variable for the $i$th observation and all observations are assumed to be independent. In this case, $\beta_0$ is the intercept term, $\epsilon_i$ is the unexplained error, and $x_1, \ldots, x_p$ are predictors with coefficient estimates $\beta_1, \ldots, \beta_p$ fit with the methods of Least Squares. The method of Least Squares estimates the intercept and slopes of these predictors by fitting a hyperplane which minimizes sum of squared residuals, where each residual is defined as $e_i = Y_{observed} - Y_{expected}$.

Bayesian hierarchical modeling does not use the method of Least Squares to estimate parameter values, instead we specify priors on the explanatory variables and then sample from the posterior distribution using Markov Chain Monte Carlo (MCMC) methods. The most common MCMC method for hierarchical modeling is the Gibbs Sampler. The Gibbs Sampler is useful because it uses the conditional distributions for each variable (which we will specify in our model) to estimate the joint distribution of our response variable.

Before we dig much deeper into how to use the Gibbs Sampler in this case, we need to specify the model. I will first specify a very simple hierarchical model work up from there. With a hierarchical model, we can choose to vary to slope(s), the intercept, or both. First, let's consider a varying intercept model with $j$ levels:

$$\begin{aligned} Y_i &\sim N(\alpha_j + \vec{\beta}\vec{X}_i,\ \sigma^2) \\ \alpha_j &\sim N(\mu_\alpha,\ \sigma_\alpha^2) \\ \mu_\alpha &\sim N(a,\ b) \end{aligned}$$

In this model, we have the response variable, $Y$, which is modeled to have a Gaussian posterior distribution with mean $\alpha_j + \vec{\beta}\vec{X}_i$ which can change intercept based on the level that a given observation is in. Note that $\mu_\alpha$ is given a hyperprior distribution where $a$ and $b$ are honest-to-god numbers that often specify a weakly informative prior.

Another type of Bayesian Multilevel Model is the varying slopes model, where the coefficient of each parameter estimate can change based on the level of the observation. We specify that model as follows:

$$\begin{aligned} Y_i &\sim N(\alpha + \vec{\beta}_j\vec{X}_i,\ \sigma^2) \\ \vec{\beta}_j &\sim N(\mu_{\vec{\beta}},\ \sigma_\beta^2) \\ \mu_{\vec{\beta}} &\sim N(c,\ d) \end{aligned}$$

In this model, the intercept of each level is set to be the same, however the coefficients ($\beta$'s) can change between groups. Note again that in order to gain information from other groups that the prior for $\mu_{\vec{\beta}}$ has hyperparmeters, which again have priors with true numbers specified in their distribution.

Neither of these models are too common, as it often makes sense to allow both the slope and intercept to vary across levels. From these first two models, it is easy to imagine what this varying-intercept-and-slopes

model will look like:

$$
\begin{aligned}
Y_i &\sim N(\alpha_j + \vec{\beta}_j \vec{X}_i,\ \sigma^2) \\
\alpha_j &\sim N(\mu_\alpha,\ \sigma_\alpha^2) \\
\vec{\beta}_j &\sim N(\mu_{\vec{\beta}},\ \sigma_\beta^2) \\
\mu_\alpha &\sim N(a,\ b) \\
\mu_{\vec{\beta}} &\sim N(c,\ d)
\end{aligned}
$$

This multilevel model allows for both the slopes and intercept to vary between groups by placing prior distributions on each which will update simulataneously between groups when we use the Gibbs sampler to sample from the posterior distribution. While these groups are specified to allow for different parameter estimates and intercept, they are also not unaware of each other. The model borrows strength from other levels and uses this information to help create the posterior distribution as well. The model borrows strength by setting the prior distribution of $\alpha_j$ and $\vec{\beta}_j$ as functions of hyperparameters, $\mu_\alpha$ and $\mu_{\vec{\beta}}$ which have their own priors that are specified with actual numbers: $a,\ b,\ c$, and $d$. Note that if we had specified exact values for the parameters of $\mu_\alpha$ and $\mu_{\vec{\beta}}$ we would not be able to share information that we are able to with the above model.

There are some great benefits to hierarchical modeling, some of which I have touched on already. However, I will reiterate some of these and mention a few that are a great benefit to forestry estimation. First, we have improved estimation for imbalance in sampling: hierarchical models will automatically deal with issues of over- and under-sampling of clusters by assigning differing uncertainty across clusters. Also, when we are dealing with groups within the data, the hierarhical models will model the variation between groups explicitly. Finally, with hierarchical models we can avoid pre-averaging our data. This allows us to retain the variation in our data to be fully used in the model and we do not lose some valuable insight that this variation could give us. (Source: Statistical Rethinking, edition 2, page 414)

## Example: Application to Forestry Estimation

As noted above, this thesis estimates forest attributes by implementing Bayesian Multilevel Models. To get a concrete sense of Bayesian Multilevel Models and the forestry data, consider a multilevel model estimating mean basal area per acre with one explanatory variable: forest biomass. Mean basal area is the average amount of area that tree stems occupy when we measure their area at breast height while forest biomass is the total biomass of the trees in the forest.

We could naively model this for the entire interior west as a linear function using OLS and get the following output:

$$\hat{Y}_i = 8.54 + 2.13x_i \quad (R^2 = 0.44)$$

This gives us a relatively poor fitting model that would not be any use for us in terms of prediction. Perhaps we want to know about basal area in some given small area, say an ecosubsection. The first thing one might think to do is to use a categorical variable that specifies ecosubsection and then create our linear model, essentially giving us a varying slopes model. While this will give us a different estimate for each ecosubsection, these types of models do not learn anything from the data of surrounding ecosubsections while doing the estimation. In his book *Statistical Rethinking*, Richard McElreath describes these models as having anterograde amnesia: "As the models move from one cluster—individual, group, location—in the data to another, estimating parameters for each cluster, they forget everything about the previous clusters" (McElreath, page 413).

We can do better than a model with anterograde amnesia. Rather than using ecosubsection as an indicator variable, we can use a Bayesian hierarchical model which allows us to help update our priors for a given

ecosubseciton with data from the surrounding area. Let's specify the model:

$$\begin{aligned} Y_i &\sim N(\alpha_j + \beta_j X,\ \sigma^2) \\ \alpha_j &\sim N(\mu_\alpha,\ \sigma_\alpha^2) \\ \beta_j &\sim N(\mu_\beta,\ \sigma_\beta^2) \\ \mu_\alpha &\sim N(0,\ 10) \\ \mu_\beta &\sim N(0,\ 10) \\ \sigma,\ \sigma_\alpha,\ \sigma_\beta &\sim \text{Exp}(1) \end{aligned}$$

Here, $Y_i$ is the mean basal area for a observation with a given ecosubsection, and we are allowing its slope and intercept to vary for each ecosubsection. The slope and intercept are given Gaussion priors with means which are hyperparameters. These hyperparameters have relatively uninformative priors, both Gaussian with mean 0 and variance 10. We must also put priors on the variances of $Y_i$, $\alpha_j$, and $\beta_j$. We specify these as Exponential with rate parameter $\lambda = 1$ which works as a "regularizing prior" to estimate the variation between groups. We do not want a prior that strongly sits on zero because that will assume there is no difference between our groups, and we do not want to allow the estimate of $\sigma$ to blow up towards infinity, as this would give us that the groups are not related in any way. The exponential distribution with rate parameter 1 does a good job at avoiding both of these issues.

Now that we have specified the model and all of our priors, we could use software like `hbsae` or `rstan` or `tidymodels` to see our output:

(Output)