Multilevel Model Building

August 22, 2017

Objectives

This lab exercise has two parts—a model *building* exercise and a model *coding* exercise. The material covered here is important and broadly useful, and it will be worthwhile to dig in deeply to understand it. I suggest that you go back and forth between the model writing exercises here and the coding exercises (provided as `MultilevelModelCodingExercise.html`). These exercises will reinforce the following:

- Diagraming and writing hierarchical models
- Using data to "model parameters"
- JAGS coding
- Creating index variables, a critically important and useful skill
- Posterior predictive checks

The big picture is to demonstrate the flexibility that you gain as a modeler by understanding basic principles of Bayesian analysis.

Nitrous oxide emissions from agricultural soils

Ecological data are often collected at multiple scales or levels of organization in nested designs. "Group" is a catchall term for the upper level in many different types of nested hierarchies. Groups could logically be composed of populations, locations, species, treatments, life stages, and individual studies, or really, any sensible category. We have measurements within groups on individual organisms, plots, species, time periods, and so on. We may also have measurements on the groups themselves, that is, covariates that apply at the upper level of organization or spatial scale or the category that contains the measurements. Multilevel models represent the way that a quantity of interest responds to the combined influence of observations taken at the group level and within the group.

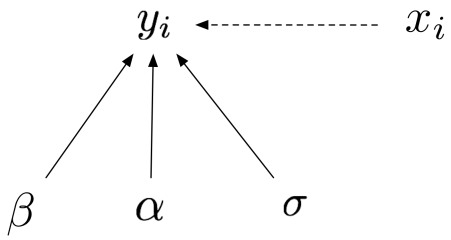
Nitrous oxide, a greenhouse gas roughly 300 times more potent than carbon dioxide in forcing atmospheric warming, is emitted when synthetic nitrogenous fertilizers are added to soils. Qian and colleagues (2010) conducted a Bayesian meta-analysis of such additions (gN $\cdot$ ha$^{-1}\cdot$ d^{-1}) using data from a study conducted by Carey (2007), who reviewed 164 relevant studies. Studies occurred at different locations, forming a group-level hierarchy[1]. Soil carbon content (g $\cdot$ organic C $\cdot$ g^{-1} soil dry matter) was measured as a group-level covariate and is assumed to be measured without error. Observations of N_2O emission are also assumed to be measured without error and were paired with measurements of fertilizer addition (kgN$\cdot$ ha^{-1}). The effect of different types of fertilizer was also studied.

1. Begin by ignoring the data on soil carbon, site, and fertilizer type so that all observations are drawn from a single pool. Draw a Bayesian network and write out the posterior and joint distribution for a linear regression model of N_2O emission on fertilizer addition for a single observation. Use a linearized power function for your deterministic model of emissions:

[1]We will use only sites that have both nitrogen and carbon data, which reduces the number of sites to 107 in the analysis here.

$$
\begin{aligned}
\mu_i &= \gamma x^{\beta} \\
\alpha &= \log\big(\gamma\big) \\
\log\big(\mu_i\big) &= \alpha + \beta\big(\log(x_i)\big) \\
g\big(\alpha, \beta, \log(x_i)\big) &= \alpha + \beta\big(\log(x_i)\big).
\end{aligned}
$$

Start by using generic [ ]. Use σ^2 to represent the uncertainty in your model realizing, of course, that you might need moment matching when you choose a specific distribution. Finish by choosing specific distributions for likelihoods and priors. Use the math in the answer as a template to code your model, although you might have made some other correct choices of distributions. You will build on this code in the subsequent problems.

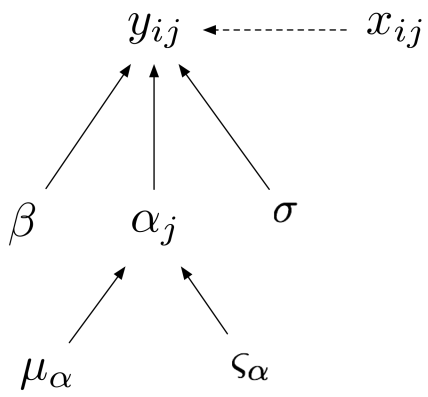


$$
\begin{aligned}
\mu_i &= \gamma x^{\beta} \\
\alpha &= \log\big(\gamma\big) \\
\log\big(\mu_i\big) &= \alpha + \beta \\
&big(\log(x_i)\big) \\
g\big(\alpha, \beta, \log(x_i)\big) &= \alpha + \beta \\
&big(\log(x_i)\big) \\
\big[\alpha, \beta, \sigma \mid y_i\big] &\propto \big[\log(y_i) \mid g\big(\alpha, \beta, \log(x_i)\big), \sigma^2\big][\alpha]\big[\beta\big]\big[\sigma\big]
\end{aligned}
$$

$$
\begin{aligned}
[\alpha, \beta, \sigma \mid \mathbf{y}] \propto & \prod_{i=1}^{n} \text{normal}\big(\log(y_i) \mid g\big(\alpha, \beta, \log(x_i), \sigma^2\big) \\
& \times \text{normal}(\alpha \mid 0, 10000)\text{normal}(\beta \mid 0, 10000) \\
& \times \text{uniform}(\sigma \mid 0, 200)
\end{aligned}
$$

What are assuming about the distribution of the untransformed μ_i? It is lognormal. Be sure you understand this.

2. Now treat the intercept in your model as a group level effect (aka, random effect). The model of the process remains a linearized power function for your deterministic model of emissions but two subscripts are required, i indexes the measurement within sites and j indexes site. Assume that the intercepts are drawn from a distribution with mean μ_α and variance ς_α^2. Draw the DAG; write an expression for the posterior and joint distributions for a single observation using [] notation, then write an expression for the posterior and joint distributions including all observations including specific choices of distributions.



$$
\begin{aligned}
g\big(\alpha_j, \beta, \log(x_{ij})\big) &= \alpha_j + \beta \log\big(\log(x_{ij})\big) \\
\big[\alpha_j, \beta, \mu_\alpha, \sigma, \varsigma_\alpha \mid y_{ij}\big] &\propto \big[\log(y_{ij}) \mid g\big(\alpha_j, \beta, \log(x_{ij}\big), \sigma^2\big]\big[\alpha_j \mid \mu_\alpha, \varsigma_\alpha^2\big]\big[\beta\big]\big[\sigma\big]\big[\mu_\alpha\big]\big[\varsigma_\alpha\big]
\end{aligned}
$$

$$
\begin{aligned}
[\boldsymbol{\alpha}, \beta, \sigma, \mu_\alpha, \varsigma_\alpha \mid \mathbf{y}] \propto & \prod_{i=1}^{n_j}\prod_{j=1}^{J} \text{normal}(\log(y_{ij}) \mid g\big(\alpha_j, \beta, \log(x_{ij}), \sigma^2\big) \\
& \times \text{normal}(\alpha_j \mid \mu_\alpha, \varsigma_\alpha) \\
& \times \text{normal}(\mu_\alpha \mid 0, 10000)\text{uniform}(\varsigma(0, 100) \\
& \times \text{normal}(\beta \mid 0, 10000)\text{uniform}(\sigma \mid 0, 100)
\end{aligned}
$$

where y_{ij} is the i^{th} observation of N_2O emissions in study j; x_{ij} is a paired measurement of fertilizer addition and β is the change in N_2O emissions per unit change in fertilizer addition. The model $g\big(\alpha_j, \beta, x_{ij}\big)$ represents the hypothesis that the log of emissions increases in direct proportion to the log of fertilizer additions and that this increase is the same for all sites and fertilizer types[2]. The intercept α_j varies among studies as a random variable drawn from a distribution with parameters[3] μ_α and ς_α^2. The fact that we explicitly represent variation among studies using the distribution of the α_j is what sets this analysis apart from conventional, single level regression that could be done separately for each of the individual sites or by pooling all of the data across sites to estimate a single

[2]We will remedy this brave assumption later.

[3]Again, these parameters will *not* be means and variances for any distributions except the normal and Poisson, but it turns out for this example, they are because we will use normals. Remember moment matching for other distributions.

intercept and slope. The σ^2 represents the uncertainty about N_2O emissions[4] and the ς_α^2 represents the uncertainty that arises as a result of variation among sites.

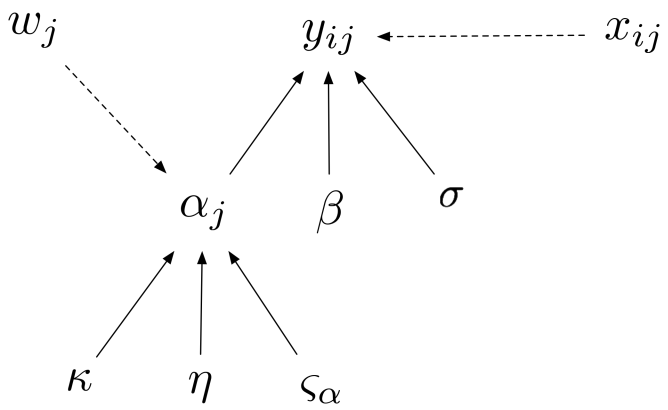
[4]We could have estimated separate σ for each site, but this was not done by Carey (2007) and I wanted to maintain consistency with their analysis as much as possible.

3. In the previous example, we assumed that there was variation the intercept that was attributable to spatial variation among sites. We did not try to explain that variation, we simply acknowledged that it exists. Now we are going to "model a parameter" using data at the group-level to explain variation in the intercepts among sites. Modify the previous model to represent the effect of soil carbon on the intercept using the deterministic model

$$g_2\big(\kappa, \eta, \text{logit}(w_j)\big) = \kappa + \eta \text{logit} w_j$$

Why do we logit transform the carbon data? Hint–they are proportions ranging from 0 to 1.

Draw a Bayesian network and write out the posterior and joint distributions for the full dataset, which means you must include the proper products. Use ij notation. You will need to notate that there are n_j measurements of N_2O emissions paired with fertilizer additions from study j. Chose appropriate distributions for each random variable.



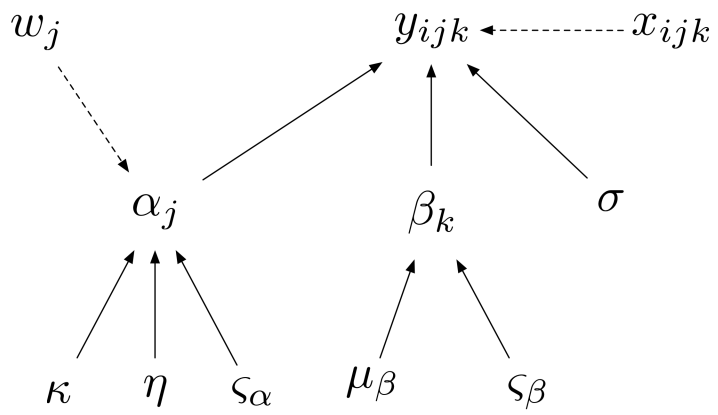
$$
\begin{aligned}
g_1\big(\alpha_j, \beta, \log(x_{ij})\big) &= \alpha_j + \beta \log\big(\log(x_{ij})\big) \\
g_2\big(\kappa, \eta, \text{logit}(w_j)\big) &= \kappa + \eta \text{logit}\big(w_j\big) \\
\big[\boldsymbol{\alpha}, \beta, \sigma, \varsigma_\alpha, \kappa, \eta \mid \mathbf{y}\big] &\propto \prod_{j=1}^{J} \prod_{i=1}^{n_j} \text{normal}\big(\log(y_{ij}) \mid g_1\big(\alpha_j, \beta, \log(x_{ij})\big), \sigma^2\big) \\
&\quad \times \text{normal}\big(\alpha_j \mid g_2\big(\kappa, \eta, \text{logit}(w_j)\big), \varsigma_\alpha^2\big) \\
&\quad \times \text{normal}\big(\beta \mid 0, 1000\big) \\
&\quad \times \text{normal}\big(\eta \mid 0, 1000\big) \\
&\quad \times \text{normal}\big(\kappa \mid 0, 1000\big) \\
&\quad \times \text{uniform}\big(\sigma \mid 0, 200\big) \\
&\quad \times \text{uniform}\big(\varsigma_\alpha \mid 0, 200\big)
\end{aligned}
$$

Wrong! To make data more spread out to eliminate leverage in a few observations

We logit transform the carbon data so they include all real numbers, which is necessary for normal distribution that we use to model the intercepts.

4. Now we are interested in the effect of carbon *and* fertilizer type on N_2O emissions. Model the effect of carbon as above, but include a group level effect of fertilizer type on the slope of the emission vs fertilizer addition model. This is to say that the slopes of the regressions are drawn from a distribution of fertilizer types. Index plot with i, site with j, and fertilizer type with k. Thus, there will be K slopes, one for each fertilizer type, drawn from a distribution with mean μ_β and variance ς^2_β. Modify the carbon model you built in the previous step to incorporate effect of fertilizer type.

   Be careful here because the group level effects are formed for two *separate* groups, site and fertilizer type. You might be tempted (or perhaps terrified) to think that you need to model the covariance in this problem, which is not the case. This is required only if you are modeling slope and intercept as group level effects for the *same* grouping variable, for example, site. You will see how this is done in the next problem. Think about it. Covariance between the slope and intercept is only important if they are being estimated from data within the same group. There is only a singe fertilizer type with each group, so it cannot covary with the intercept.



$$
\begin{aligned}
g_1\big(\alpha_j, \beta_k, \log(x_{ijk})\big) &= \alpha_j + \beta_k \log(x_{ijk}) \\
g_2\big(\kappa, \eta, \text{logit}(w_j)\big) &= \kappa + \eta \text{logit}(w_j) \\
\big[\boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma, \varsigma_\alpha, \kappa, \eta, \mu_\beta, \varsigma_\beta \mid \mathbf{y}\big] &\propto \prod_{j=1}^{J} \prod_{k=1}^{K_j} \prod_{i=1}^{n_j} \text{normal}\big(\log\big(y_{ijk}\big) \mid g_1\big(\alpha_j, \beta_k, \log(x_{ijk})\big), \sigma^2\big) \\
&\quad \times \text{normal}\big(\alpha_j \mid g_2\big(\kappa, \eta, \text{logit}(w_j)\big), \varsigma_\alpha^2\big) \\
&\quad \times \text{normal}\big(\beta_k \mid \mu_\beta, \varsigma_\beta^2\big) \\
&\quad \times \text{normal}\big(\eta \mid 0, 1000\big) \\
&\quad \times \text{normal}\big(\kappa \mid 0, 1000\big) \\
&\quad \times \text{uniform}\big(\sigma \mid 0, 100\big) \\
&\quad \times \text{uniform}\big(\varsigma_\alpha \mid 0, 200\big) \\
&\quad \times \text{normal}\big(\mu_\beta \mid 0, 1000\big) \\
&\quad \times \text{uniform}\big(\varsigma_\beta \mid 0, 200\big)
\end{aligned}
$$

5. Now return to problem 2 where you assumed that different groups had the different intercepts but the same slope, which is to say that individual sites had emission responses to fertilizer that were parallel. This seems unreasonable (particularly when you look at lattice plots of the data), representing the need to model group effects on intercepts *and slopes.* The idea is that both the slope and the intercept are random variables drawn from a distribution of slopes and intercepts where variation in the values of the random variable is attributable to unspecified spatial differences among sites. It is tempting to simply add a distribution for the slopes in the same way you modeled the intercepts, and you will see papers where this is done (wrongly). However, when you seek to understand group effects on multiple parameters you must account for the way that the parameters *covary.* Write a model for group effects on slope *and* intercepts. Exploit the following hints.

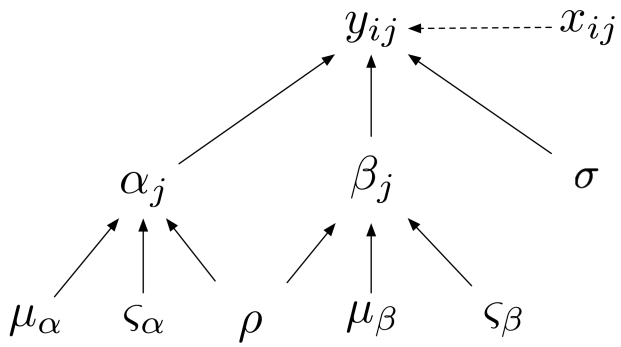
   - Represent the slope and intercept as a two element vector for each group and use a multivariate normal distribution in the same way you used a normal for the intercept. So the individual slopes (α_j) and intercept (β_j) will be drawn from

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

   - The second parameter in the multivariate normal is a variance-covariance matrix representing how the slope and intercept covary. It has variance terms on the diagonal, and covariance terms on the off-diagonal, i.e.,

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

     The covariance terms are defined as $\text{Cov}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \rho \varsigma_\alpha \varsigma_\beta$ where ρ is the coefficient of correlation between $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$.



$$
\begin{aligned}
g\big(\alpha_j, \beta_j, \log(x_{ij})\big) &= \alpha_j + \beta_j \log\big(x_{ij}\big) \\
\big[\boldsymbol{\alpha}, \boldsymbol{\beta}, \mu_\alpha, \mu_\beta, \sigma, \varsigma_\alpha, \varsigma_\beta, \rho \mid \mathbf{y}\big] &\propto \prod_{j=1}^{J} \prod_{i=1}^{n_j} \text{normal}\big(\log(y_{ij}) \mid g\big(\alpha_j, \beta_j, \log(x_{ij})\big), \sigma^2\big) \\
&\quad \times \text{multivariate normal}\left(\begin{pmatrix} \alpha_j \\ \beta_j \end{pmatrix} \,\middle|\, \begin{pmatrix} \mu_\alpha \\ \mu_\beta \end{pmatrix}, \begin{pmatrix} \varsigma_\alpha^2 & \rho\varsigma_\alpha\varsigma_\beta \\ \rho\varsigma_\alpha\varsigma_\beta & \varsigma_\beta^2 \end{pmatrix}\right) \\
&\quad \times \text{normal}\big(\mu_\alpha \mid 0, 1000\big) \\
&\quad \times \text{normal}\big(\mu_\beta \mid 0, 1000\big) \\
&\quad \times \text{uniform}\big(\sigma \mid 0, 100\big) \\
&\quad \times \text{uniform}\big(\varsigma_\alpha \mid 0, 200\big) \\
&\quad \times \text{uniform}\big(\varsigma_\beta \mid 0, 200\big) \\
&\quad \times \text{uniform}\big(\rho \mid 0, 1\big)
\end{aligned}
$$

References

Carey, K. 2007. Modeling N_2O emission from agricultural soils using a multilevel linear regression. (Doctoral dissertation, Duke University).

Qian, S. S., Cuffney, T. F., Alameddine, I., McMahon, G., & Reckhow, K. H. 2010. On the application of multilevel modeling in environmental and ecological studies. Ecology, 91(2), 355-361.