Seeds example

This is based on the “Seeds: Random effect logistic regression” example in OpenBUGS Examples, Volume I. It is somewhat similar to the “Moth example,” in that it involves a logistic regression. The response data are counts for the number of seeds that germinated (*r*) out of the total number of seeds “planted” (*n*). The goal is to determine if germination probability (*p*) differs between two plant species and if it is affected by application of two different types of root extract to the media in which the seeds are placed. Dummy variables are created for these two experimental factors (species/seed and extract), both of which have 2 levels. For this problem, read through the OpenBUGS example documentation (next page) and work through the model. Program the model in JAGS and run it via R, as done for Activity 1 (Day 1). This documentation describes the model.

Things to consider as you work through and run the model:

1. Can you identify the likelihood, mean model, and prior components? For example, what priors are used for the regression coefficients (α’s)?
2. The OpenBUGS example provides initials for 2 MCMC chains; modify this to include initials for 3 chains.
3. Run the model via jags.model and coda.samples; evaluate the chains for mixing, convergence, autocorrelation, and burn-in. Compute posterior statistics for parameters of interest based on the converged samples.
4. Does the probability of germination depend on species or extract type?
5. How do you interpret α_{12} and is it statistically significant?
6. Is an observation level random effect appropriate?
7. Is there an alternative way of modeling over-dispersion that doesn’t require explicit inclusion of b_i? For example, one could also implement:

$$\text{logit}(p_i) \sim \mathit{Normal}\left(\alpha_0 + \alpha_1 x_{1i} + \alpha_2 x_{2i} + \alpha_{12} x_{1i} x_{2i},\ \tau\right)$$

8. Try modify the model to include the alternative over-dispersion formulation (above). Does this have any notable impact on the behavior of the MCMC chains or the posterior statistics for the parameters of interest?

Seeds: Random effect logistic regression

This example is taken from Table 3 of Crowder (1978), and concerns the proportion of seeds that germinated on each of 21 plates arranged according to a 2 by 2 factorial layout by seed and type of root extract. The data are shown below, where r_i and n_i are the number of germinated and the total number of seeds on the *i* th plate, *i* =1,...,N. These data are also analysed by, for example, Breslow: and Clayton (1993).

	seed O. aegyptiaco 75						*seed O. aegyptiaco 73*				
	Bean			Cucumber			Bean			Cucumber	
r	n	r/n	r	n	r/n	r	n	r/n	r	n	r/n
10	39	0.26	5	6	0.83	8	16	0.50	3	12	0.25
23	62	0.37	53	74	0.72	10	30	0.33	22	41	0.54
23	81	0.28	55	72	0.76	8	28	0.29	15	30	0.50
26	51	0.51	32	51	0.63	23	45	0.51	32	51	0.63
17	39	0.44	46	79	0.58	0	4	0.00	3	7	0.43
			10	13	0.77						

The model is essentially a random effects logistic, allowing for over-dispersion. If p_i is the probability of germination on the *i* th plate, we assume

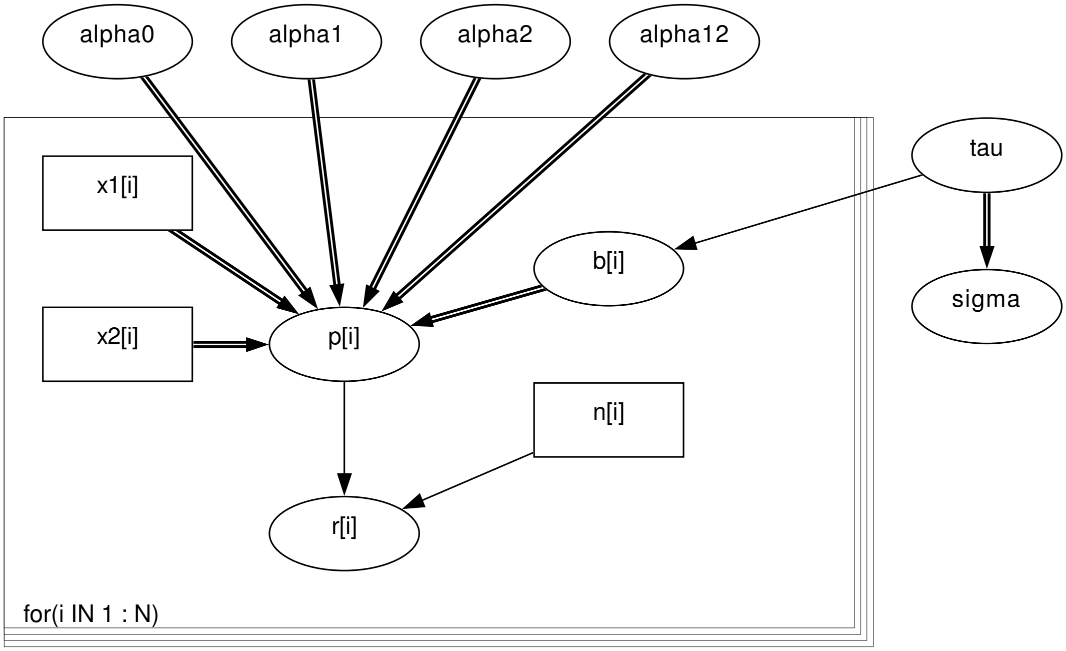
$$r_i \sim \text{Binomial}(p_i, n_i)$$

$$\text{logit}(p_i) = \alpha_0 + \alpha_1 x_{1i} + \alpha_2 x_{2i} + \alpha_{12} x_{1i} x_{2i} + b_i$$

$$b_i \sim \text{Normal}(0, \tau)$$

where x_{1i} , x_{2i} are the seed type and root extract of the *i* th plate, and an interaction term $\alpha_{12} x_{1i} x_{2i}$ is included. α_0 , α_1 , α_2 , α_{12} , τ are given independent "noninformative" priors.

Graphical model for seeds example



BUGS *language for seeds example*

```
model
{
   for( i in 1 : N ) {
      r[i] ~ dbin(p[i],n[i])
      b[i] ~ dnorm(0.0,tau)
      logit(p[i]) <- alpha0 + alpha1 * x1[i] + alpha2 * x2[i] +
         alpha12 * x1[i] * x2[i] + b[i]
   }
   alpha0 ~ dnorm(0.0,1.0E-6)
   alpha1 ~ dnorm(0.0,1.0E-6)
   alpha2 ~ dnorm(0.0,1.0E-6)
   alpha12 ~ dnorm(0.0,1.0E-6)
   tau ~ dgamma(0.001,0.001)
   sigma <- 1 / sqrt(tau)
}
```

Data (click to open)

Inits for chain 1   Inits for chain 2   (click to open)

Results

A burn in of 1000 updates followed by a further 10000 updates gave the following parameter estimates:

```
        mean     sd      MC_error val2.5pc median   val97.5pc start  sample
alpha0  -0.5499  0.1965  0.004298 -0.9433  -0.5522  -0.1596   1001   10000
alpha1  0.08902  0.3124  0.005997 -0.5504  0.09795  0.6812    1001   10000
alpha12 -0.841   0.4372  0.008725 -1.736   -0.8265  0.008258  1001   10000
alpha2  1.356    0.2772  0.006133 0.8298   1.351    1.914     1001   10000
sigma   0.2922   0.1467  0.007297 0.04439  0.2838   0.6104    1001   10000
```

We may compare simple logistic, maximum likelihood (from EGRET), penalized quasi-likelihood (PQL) Breslow and Clayton (1993) with the *BUGS* results

	Logistic regression		maximum likelihood		PQL	
variable	β	SE	β	SE	β	SE
α_0	-0.558	0.126	-0.546	0.167	-0.542	0.190
α_1	0.146	0.223	0.097	0.278	0.77	0.308
α_2	1.318	0.177	1.337	0.237	1.339	0.270
α_{12}	-0.778	0.306	-0.811	0.385	-0.825	0.430
σ	---	---	0.236	0.110	0.313	0.121

Heirarchical centering is an interesting reformulation of random effects models. Introduce the variables

$$\mu_i = \alpha_0 + \alpha_1 x_{1i} + \alpha_2 x_{2i} + \alpha_{12} x_{1i} x_{2i}$$

$$\beta_i = \mu_i + b_i$$

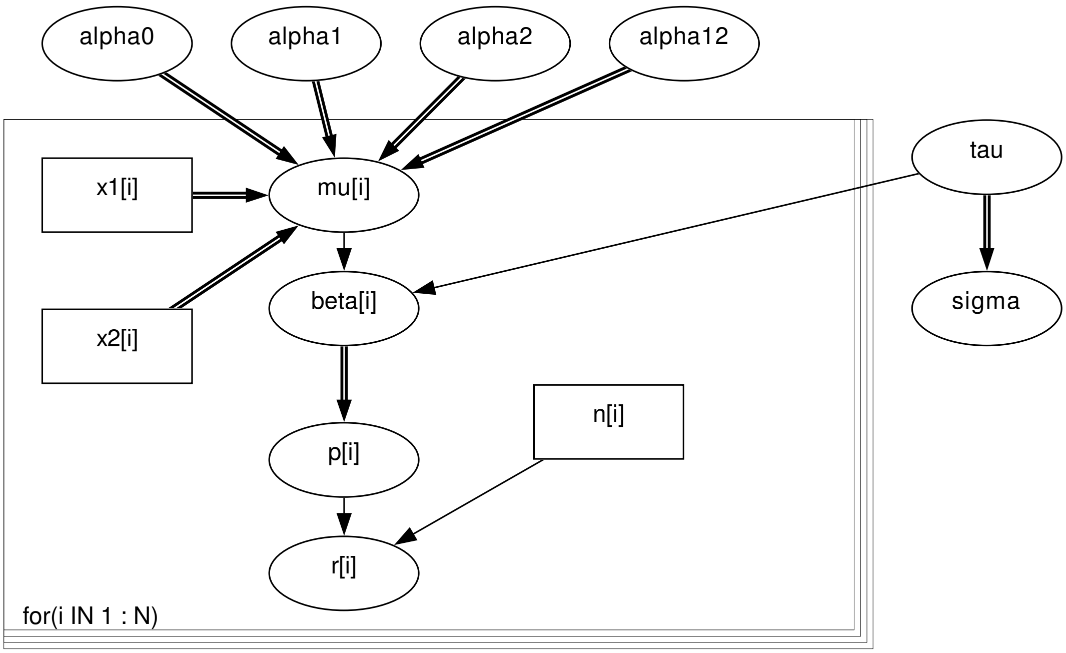
the model then becomes

$$r_i \sim \text{Binomial}(p_i, n_i)$$

$$\text{logit}(p_i) = \beta_i$$

$$\beta_i \sim \text{Normal}(\mu_i , \tau)$$

The graphical model is shown below



This formulation of the model has two advantages: the squence of random numbers generated by the Gibbs sampler has better correlation properties and the time per update is reduced because the updating for the α parameters is now conjugate.