

# Bayesian generalised mixed models with MCMCglmm

Ferran Sayol

6th October 2022

*Previous Course editions:* 8th September 2021, Gothenburg (Sweden) 5th November 2020, Gothenburg (Sweden) 20th December 2018, Erlangen (Germany)

*Summary* This is a short guide of running mixed models under a bayesian framework using the MCMCglmm R package (Hadfield 2010).

The course is structured in three parts: 1) See the basics of MCMCglmm models (run and check the output); 2) Adding random effects; 3) Correct for phylogenetic effects.

## Used packages

First we need to install the MCMCglmm package. We also install an additional packages to do plots (ggplot2) and work with phylogenetic trees (phytools). Next, we load the packages we have just installed from the library.

```
library(MCMCglmm)
library(phytools)
library(ggplot2)
```

## PART 1: Introduction to MCMCglmm models

First we will load some data as an example. We will use data on morphological measurements and ecology of pigeons & doves (Columbidae).

```
getwd() #Check the Working directory use "setwd()" if necessary
```

```
## [1] "/Users/fsayol/My Drive/1_ResearchOrganization/5_Teaching/2022/221006_SpatialR"
```

```
cdata <- read.table("data/ColumbidaeTraits.txt",h=T)
```

This data file is based on a subset of the data used in an analysis on the relation between foraging behaviour and the evolution of morphological adaptations (Lapiedra et al. 2013).

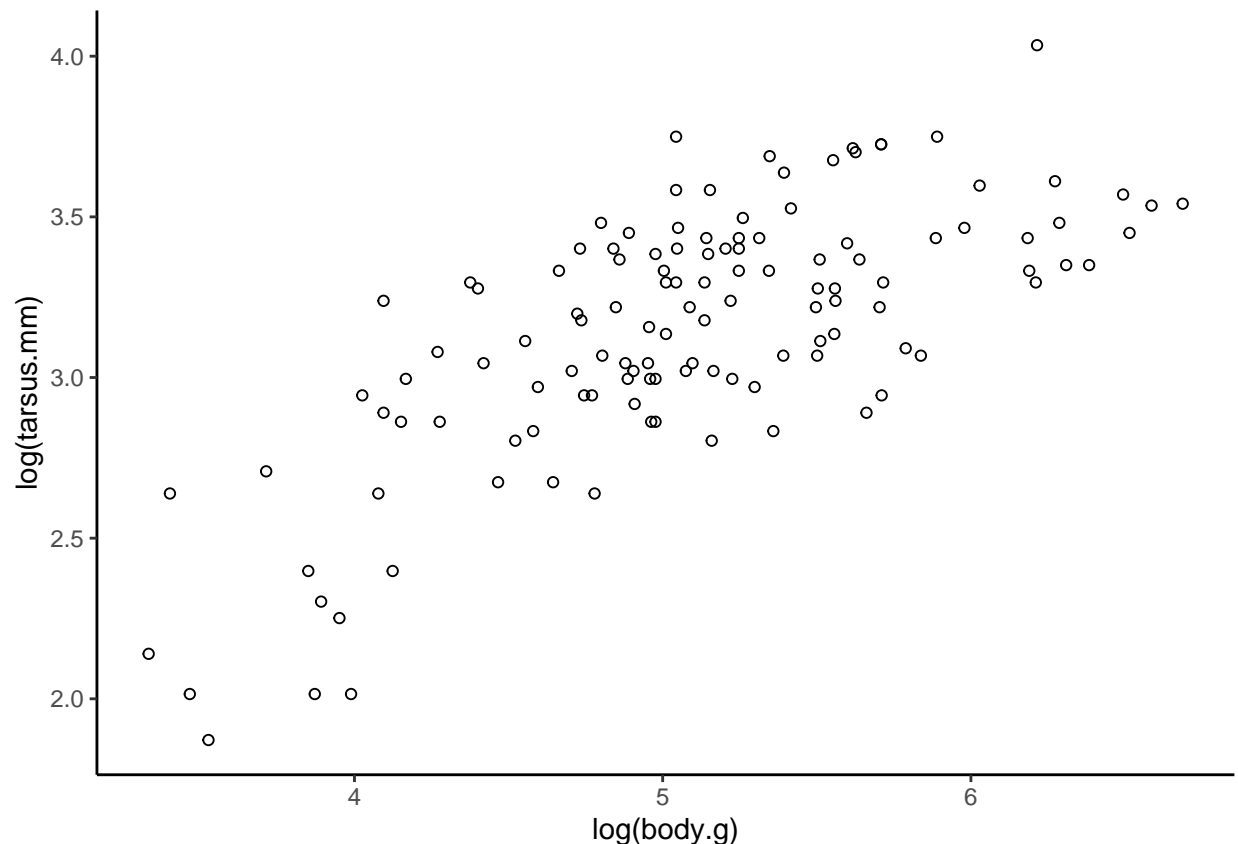
```
head(cdata)
```

```
##           species tarsus.mm tail.mm wing.mm body.g   foraging   region
## 1 Uropelia_campestris      8.5   55.5   64.0    28 terrestrial Americas
## 2  Geopelia_cuneata     14.0  102.5   91.0    30 terrestrial Australasia
## 3 Columбина_passerina      7.5   30.0   78.0    32 terrestrial  Americas
```

```
## 4   Columbina_minuta      6.5   29.5   64.5   34 terrestrial Americas
## 5   Oena_capensis       15.0  140.5  107.0   41 terrestrial  Africa
## 6   Columbina_cruziana   11.0   29.5   83.0   47 terrestrial Americas
##      measure
## 1 ResearcherA
## 2 ResearcherC
## 3 ResearcherC
## 4 ResearcherB
## 5 ResearcherB
## 6 ResearcherA
```

Let's pretend our goal is to study the relation between morphology and ecology. For instance, whether the tarsus length is related to foraging behaviour. But we will first start by exploring the relation between tarsus length and body size.

```
ggplot(cdata, aes(x=log(body.g), y=log(tarsus.mm))) +
  geom_point(shape=21)+theme_classic()
```



Now let's run a simple model with tarsus.mm as response of body size (body.g).

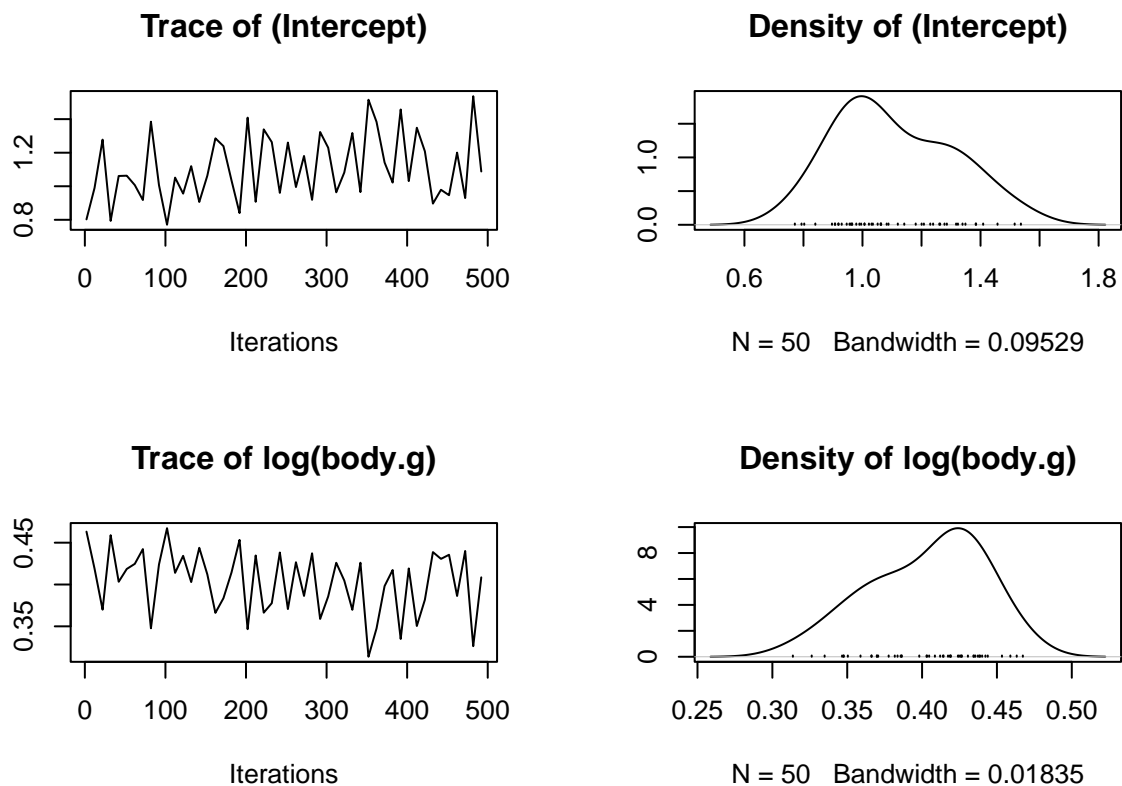
When running an MCMCglmm, we need to specify some parameters of the mcmc chain: How many iterations we want to run the chain for (nitt), the burnin we want to discard at the start of the chain (burnin) and also how often we want to sample and store from the chain (thin). We discard a burnin as is normal practice in Bayesian analyses.

```
prior1 <- list(R=list(V = 1,nu = 0.002)) #We will see this later

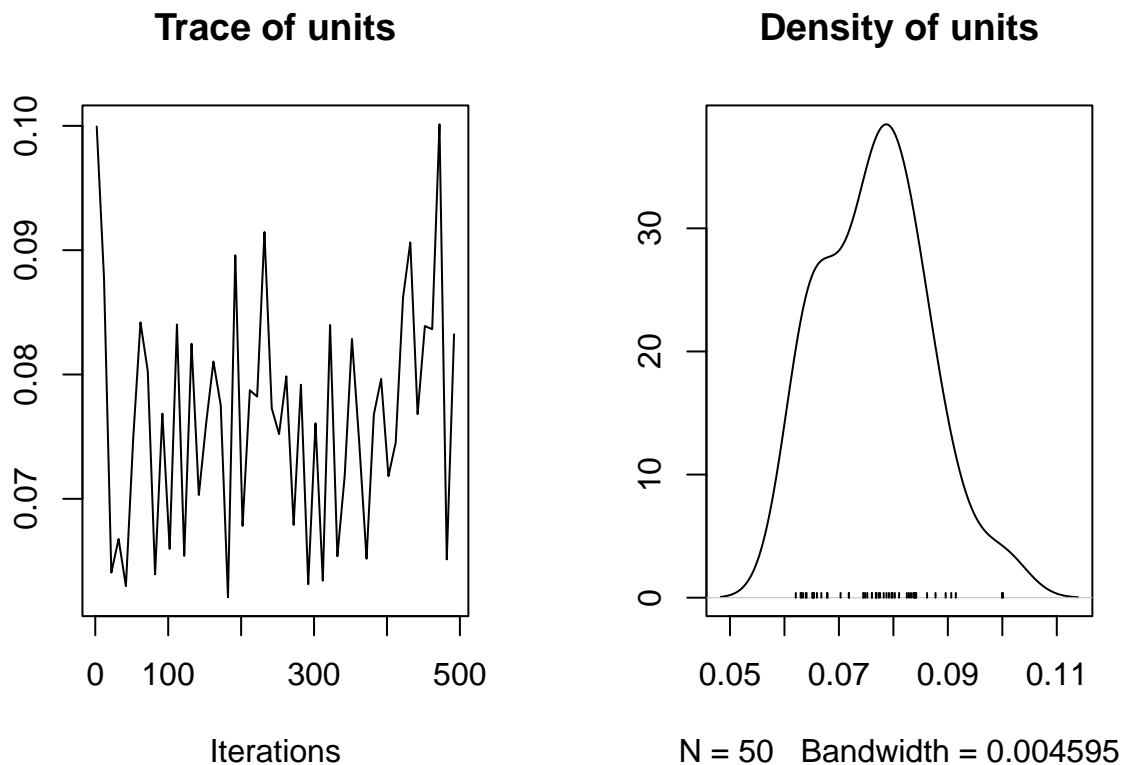
mod1.1 <- MCMCglmm(log(tarsus.mm) ~ log(body.g),
  data = cdata, prior = prior1,verbose=F,
  nitt = 500, thin=10, burnin = 1)
```

Before we even look at our model output we can check if the model ran appropriately (e.g. if the model converged and if there is any issue in chain mixing). We can do this by visually inspecting the chains. We can extract the full chains using *modelSol \* for the fixed effects and \* modelVCV* for the variance terms. So *Sol[,1]* will give you the first fixed term, in this case the intercept, and *VCV[,1]* will give you the first random term, which is just the residual term here. As our model is an mcmc object when we use the *plot()* function we get a trace plot.

```
#plot the first fixed term, the intercept.
plot(mod1.1$Sol)
```



```
#plot the first variance term, the residual error term.
plot(mod1.1$VCV)
```



On the right hand side of the plots is the posterior distributions for each of the terms. On the left side of these plots are the traces of the mcmc chain for each estimate. What we want to see in these trace plots has an apparent random pattern. That is a trace with no obvious trend that is bouncing around some stable point.

Another thing we also want to check is the level of auto-correlation in the chain traces. We can do this using `autocorr.diag()` which gives the level of correlation along the chain between some lag sizes. (Note that this is the same thing we were looking at for another purpose yesterday when we discussed temporal autocorrelation for CO2 and temperature)

Let's see some diagnosis (Autocorrelation)

```
autocorr.diag(mod1.1$Sol) #Solutions (coeficients)
```

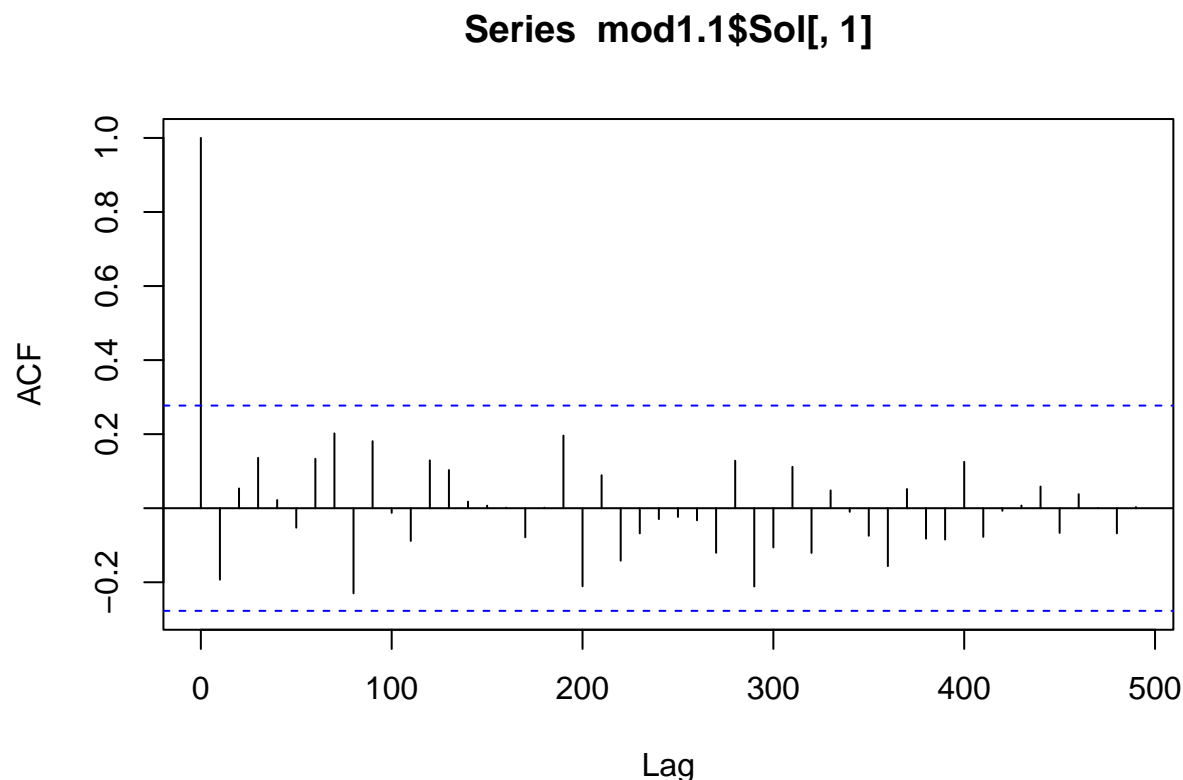
```
##          (Intercept) log(body.g)
## Lag 0      1.00000000  1.00000000
## Lag 10    -0.19294194 -0.17980641
## Lag 50    -0.05260378 -0.02140413
## Lag 100   -0.01245856 -0.01290855
```

```
autocorr.diag(mod1.1$VCV) #Variance
```

```
##          units
## Lag 0      1.00000000
## Lag 10    -0.09807331
## Lag 50    -0.10942511
## Lag 100   -0.04972382
```

Another way is to look at autocorrelation plots for each of the traces. For example, let's check the autocorrelation in the intercept chain using the *acf* function

```
#acf plot for the first fixed estimate in our model (the intercept)
acf(mod1.1$Sol[,1],lag.max =100)
```



Ideally, we have to make sure that the autocorrelation is low (i.e. less than 0.1 is recommended). The thinning is used to reduce autocorrelation in our sample, how much you use often depends on how much autocorrelation you find and we can reduce autocorrelation by increasing the thinning interval. As a result, we might have to increase the total number of iterations as well to have a sample of at least 1000. We can also set a Burn-in (normally 5-10% of samples) to get rid of the first samples that have not converged yet.

- 
- EXERCISE 1: *Increase the thinning interval and the number of iterations to make sure there is no autocorrelation and to have a sample of >1000.*
- 

Now, let's explore the output of model.

```
summary(mod1.1)
```

```
##
```

```
## Iterations = 1001:100901
## Thinning interval = 100
## Sample size = 1000
##
## DIC: 37.37207
##
## R-structure: ~units
##
##      post.mean l-95% CI u-95% CI eff.samp
## units  0.07856 0.06243 0.1007    1000
##
## Location effects: log(tarsus.mm) ~ log(body.g)
##
##      post.mean l-95% CI u-95% CI eff.samp pMCMC
## (Intercept)  1.1176 0.7679 1.4866    1000 <0.001 ***
## log(body.g)  0.4002 0.3300 0.4694    1000 <0.001 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can see the estimates for the fixed factor. Each parameter has a measure of the effect size under `post.mean` and a lower and higher 95% credible interval (CI).

Another way to directly look at the posterior means and confidence intervals for the factors is with the following commands.

Posterior mean of fixed factors:

```
posterior.mode(mod1.1$Sol)
```

```
## (Intercept) log(body.g)
## 1.0811903 0.4114074
```

Posterior mean of fixed factors:

```
HPDinterval(mod1.1$Sol)
```

```
##      lower      upper
## (Intercept) 0.7679241 1.4865714
## log(body.g) 0.3299767 0.4694338
## attr(,"Probability")
## [1] 0.95
```

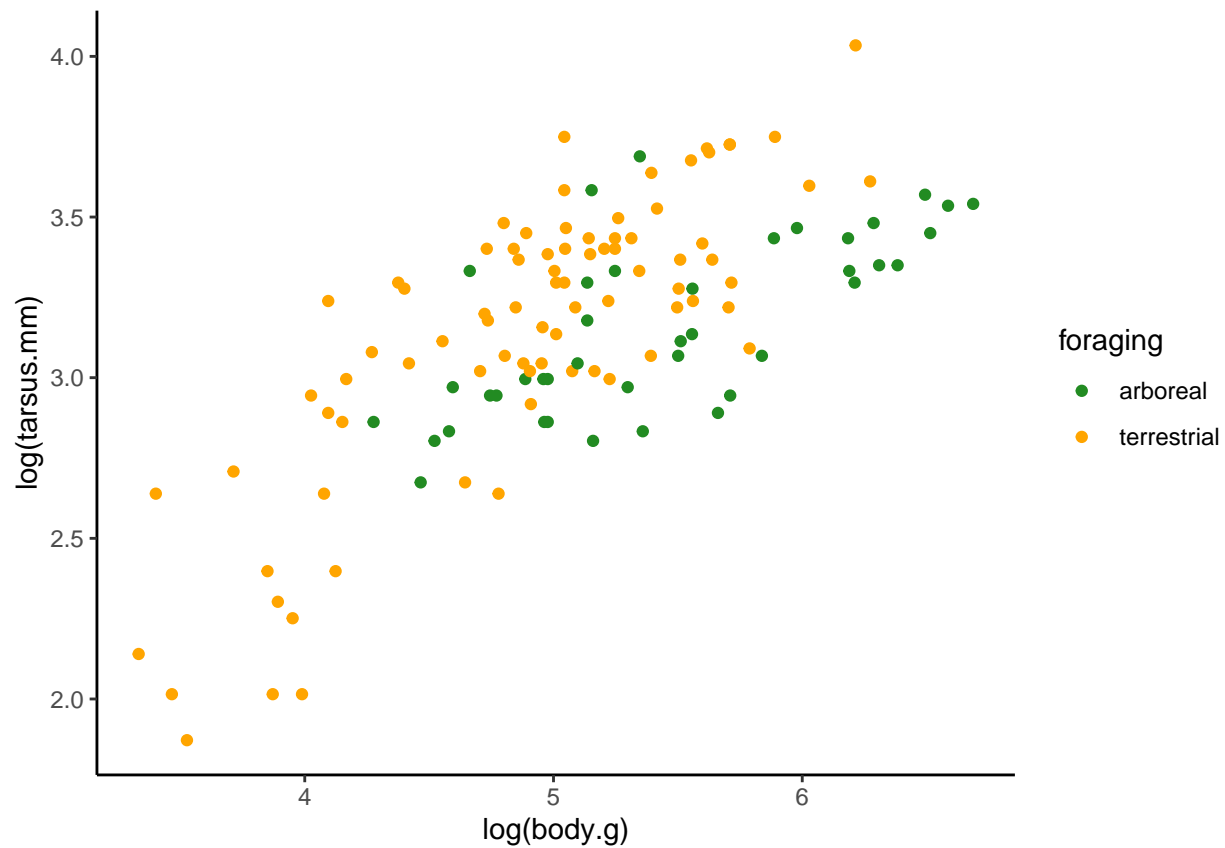
We also have the effective sample size (`eff.samp`).

Finally, the *pMCMC* is an equivalent of the famous *p-value* in Frequentist statistics. In MCMCglmm models, this is calculated as two times the probability that the estimate is either  $>$  or  $<$  0, using whichever one is smaller. However, since our data has been mean centred and expressed in units of standard deviation we can simply look at what proportion of our posterior is on either side of zero.

To evaluate the fit of the model, we have the parameter *DIC*, which is a Bayesian version of *AIC*. Like *AIC* it is a measure of the trade-off between the “fit” of the model and the number of parameters, with a lower number better.

Coming back to the initial question, we want to see if foraging behaviour can explain tarsus length while accounting for body size. Let's do a plot first:

```
ggplot(cdata, aes(x=log(body.g), y=log(tarsus.mm), color=foraging)) +
  geom_point(shape=19)+theme_classic()+scale_color_manual(values=c("forestgreen","orange"))
```



It looks like relative length of the tarsus is influenced by foraging behaviour. But we need to formally test this:

- 
- EXERCISE 2: Run a new model (*mod1.2*) including also the foraging ecology together with body size as predictor and compare the DIC with *mod1.1* / Which of the models is better?
- 

```
prior1 <- list(R=list(V = 1,nu = 0.002)) #We will see this later
```

```
names(cdata)
```

```
## [1] "species" "tarsus.mm" "tail.mm" "wing.mm" "body.g" "foraging"
## [7] "region" "measure"
```

```
mod1.2 <- MCMCglmm(log(tarsus.mm) ~ log(body.g)+foraging,
  data = cdata, prior = prior1,verbose=F,
  nitt = 101000, thin=100, burnin = 1000)
# We can remove the intercept:
```

```
mod1.3 <- MCMCglmm(log(tarsus.mm) ~ log(body.g)+foraging-1,
                  data = cdata, prior = prior1, verbose=F,
                  nitt = 101000, thin=100, burnin = 1000)

summary(mod1.2)
```

```
##
## Iterations = 1001:100901
## Thinning interval = 100
## Sample size = 1000
##
## DIC: 20.17921
##
## R-structure: ~units
##
##      post.mean l-95% CI u-95% CI eff.samp
## units  0.06755 0.05089 0.08457      983
##
## Location effects: log(tarsus.mm) ~ log(body.g) + foraging
##
##      post.mean l-95% CI u-95% CI eff.samp pMCMC
## (Intercept)      0.6680   0.3392   1.0841   1000 <0.001 ***
## log(body.g)       0.4573   0.3914   0.5244   1000 <0.001 ***
## foragingterrestrial 0.2384   0.1337   0.3446   1000 <0.001 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

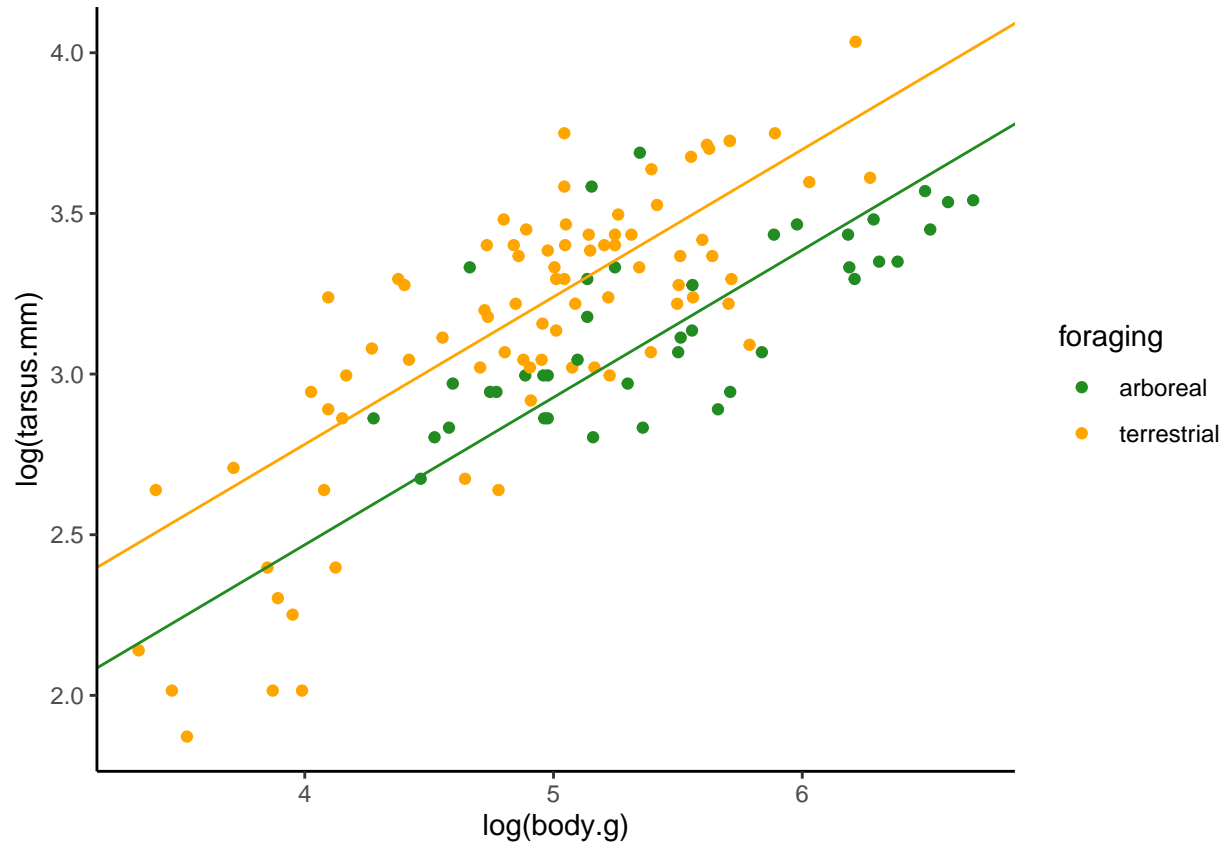
```
summary(mod1.3)
```

```
##
## Iterations = 1001:100901
## Thinning interval = 100
## Sample size = 1000
##
## DIC: 20.19541
##
## R-structure: ~units
##
##      post.mean l-95% CI u-95% CI eff.samp
## units  0.06734 0.05177 0.08442      1000
##
## Location effects: log(tarsus.mm) ~ log(body.g) + foraging - 1
##
##      post.mean l-95% CI u-95% CI eff.samp pMCMC
## log(body.g)       0.4558   0.3876   0.5221   1000 <0.001 ***
## foragingarboreal   0.6766   0.3046   1.0652   1000 0.002 **
## foragingterrestrial 0.9130   0.5783   1.2385   1000 <0.001 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Now we can plot the output of the models:
```



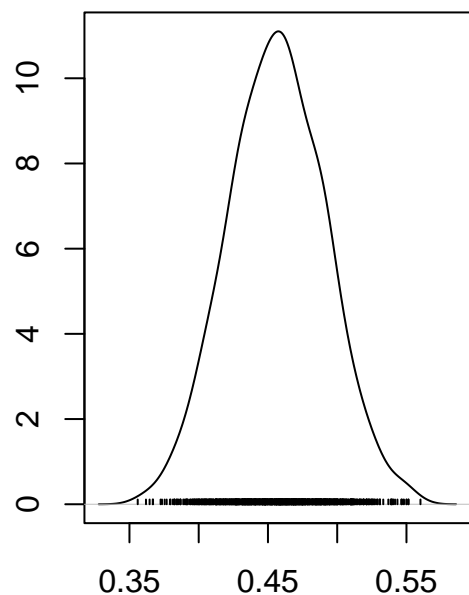
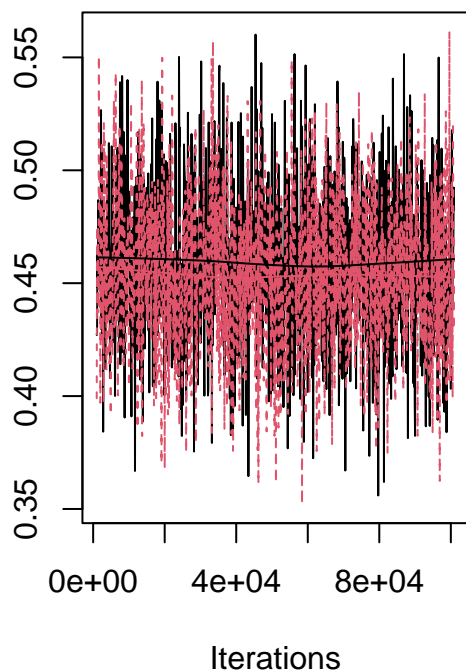
```
ggplot(cdata, aes(x=log(body.g), y=log(tarsus.mm), color=foraging)) +
  geom_point(shape=19) + theme_classic() + scale_color_manual(values=c("forestgreen","orange")) +
  geom_abline(intercept = posterior.mode(mod1.3$Sol)[2], slope = posterior.mode(mod1.3$Sol)[1], color="forestgreen") +
  geom_abline(intercept = posterior.mode(mod1.3$Sol)[3], slope = posterior.mode(mod1.3$Sol)[1], color="orange")
```



## Model convergence

One last thing to check is that our MCMC chain has properly converged and that our estimate is not the result of some type of transitional behaviour. That is have our chains “found” the optimum or do we need to let them run longer before they settle around some estimate. To check this we will run a second model and see if it converges on the same estimates as our first model.

```
mod1.3a <- MCMCglmm(log(tarsus.mm) ~ log(body.g)+foraging,
  data = cdata, prior = prior1, verbose=F,
  nitt = 101000, thin=100, burnin = 1000)
mod1.3b <- MCMCglmm(log(tarsus.mm) ~ log(body.g)+foraging,
  data = cdata, prior = prior1, verbose=F,
  nitt = 101000, thin=100, burnin = 1000)
#They have reached the same solution?
plot(mcmc.list(mod1.3a$Sol[,2], mod1.3b$Sol[,2]))
```



N = 1000 Bandwidth = 0.008093

```
summary(mod1.3b)
```

```
##
## Iterations = 1001:100901
## Thinning interval = 100
## Sample size = 1000
##
## DIC: 20.18055
##
## R-structure: ~units
##
##      post.mean l-95% CI u-95% CI eff.samp
## units  0.06764  0.05041  0.08557    1000
##
## Location effects: log(tarsus.mm) ~ log(body.g) + foraging
##
##      post.mean l-95% CI u-95% CI eff.samp pMCMC
## (Intercept)      0.6857   0.2763   1.0309   1000 <0.001 ***
## log(body.g)      0.4539   0.3823   0.5169   1000 <0.001 ***
## foragingterrestrial 0.2375   0.1309   0.3400   1000 <0.001 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

## PART 2: Modify priors and add random factors

Since we are using a Bayesian approach we will need to set up the priors. In most cases we want to use a non-informative prior that doesn't influence the estimated posterior distribution. We are basically saying that we don't know anything about the expected values for our parameters. That is we have no prior information.

To give priors for MCMCglmm we need to make an object that is in a list format that includes terms B (fixed effects), R (residual terms) and G (random effects).

In our model we have 3 fixed terms B (1 intercept + 2 factors) and the residual term R.

For fixed effects (B), MCMCglmm uses a normal distribution. The terms mu and V give the mean and variance of this normal distribution. Here we set mu as 0 and the variance as a large number to make these priors effectively uninformative (These are called non-informative priors).

Since we have three fixed terms (two intercepts and one slope) we can use the *diag()* function to create a matrix to store a prior for each.

```
fixMu <- rep(0,3)
fixV <- diag(3)*10^8
prior2.1 <- list(B=list(mu=fixMu,V=fixV),R=list(V = 1,nu = 0.002))

mod2.1 <- MCMCglmm(log(tarsus.mm) ~ foraging+log(body.g),
                   data = cdata, prior = prior2.1,verbose=F,family="gaussian",
                   nitt = 1100, thin=10, burnin = 100)

summary(mod2.1)
```

```
##
## Iterations = 101:1091
## Thinning interval = 10
## Sample size = 100
##
## DIC: 19.85891
##
## R-structure: ~units
##
##      post.mean l-95% CI u-95% CI eff.samp
## units    0.06831    0.0546    0.08863      100
##
## Location effects: log(tarsus.mm) ~ foraging + log(body.g)
##
##      post.mean l-95% CI u-95% CI eff.samp pMCMC
## (Intercept)      0.6653    0.2518    1.0058     100 <0.01 **
## foragingterrestrial 0.2487    0.1655    0.3725     100 <0.01 **
## log(body.g)       0.4565    0.3884    0.5224     100 <0.01 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Normally we don't need to set this as MCMCglmm will set non-informative priors automatically for fixed terms. Then, we can set the prior by only specifying the R term:

```
prior2.1 <- list(R=list(V = 1,nu = 0.002))

mod2.1 <- MCMCglmm(log(tarsus.mm) ~ foraging+log(body.g),
                   data = cdata, prior = prior2.1,verbose=F,family="gaussian",
```

```

nitt = 1100, thin=10, burnin = 100)
summary(mod2.1)

```

```

##
## Iterations = 101:1091
## Thinning interval = 10
## Sample size = 100
##
## DIC: 20.54358
##
## R-structure: ~units
##
##      post.mean l-95% CI u-95% CI eff.samp
## units  0.06721  0.04921  0.08499      100
##
## Location effects: log(tarsus.mm) ~ foraging + log(body.g)
##
##      post.mean l-95% CI u-95% CI eff.samp pMCMC
## (Intercept)      0.6452   0.3503   1.0227   186.7 <0.01 **
## foragingterrestrial 0.2397   0.1369   0.3468   100.0 <0.01 **
## log(body.g)        0.4612   0.3963   0.5208   171.4 <0.01 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

MCMCglmm uses inverse-Wishart priors for any of the variance terms (R or G). The variance is described by the parameters  $\nu$  and  $V$ .

$V=1$  and  $\nu=0.002$  is frequently used for variance components.

See some information about inverse-Wishart priors here: [https://en.wikipedia.org/wiki/Inverse-Wishart\\_distribution](https://en.wikipedia.org/wiki/Inverse-Wishart_distribution)

## Mixed models: Adding random factors to our MCMCglmm

Just to remember, mixed models are referred to models that contain both fixed and random factors. Depending in our interest a factor might be considered random or fix. Here, you can see a more detailed explanation.

Let's add a random term of measurement ("measure") in the *cdata* example. This variable states which researcher took the morphological measurements (A,B,C). Like before, we need to set up the prior. To add a random term we now add a  $G$  structure that acts just like the other random variance term and is defined using  $\nu$  and  $V$ .

```

prior2.2 <- list(G = list(G1 = list(nu=0.002, V=1)),
  R = list(nu=0.002, V=1))

```

Here, we will include "measure" as a random effect:

```

table(cdata$measure)

```

```

##
## ResearcherA ResearcherB ResearcherC
##      40      41      42

```

We can do so by including the random variable in the model in the section `random= ~.`

```
table(cdata$measure)

##
## ResearcherA ResearcherB ResearcherC
##          40          41          42

names(cdata)

## [1] "species"  "tarsus.mm" "tail.mm"   "wing.mm"   "body.g"    "foraging"
## [7] "region"    "measure"
```

```
mod2.2 <- MCMCglmm(log(tarsus.mm) ~ foraging+log(body.g),
                  random= ~measure,
                  data = cdata, prior = prior2.2, verbose=F,
                  nitt = 1100, thin=10, burnin = 100)
summary(mod2.2)

##
## Iterations = 101:1091
## Thinning interval = 10
## Sample size = 100
##
## DIC: 17.95901
##
## G-structure: ~measure
##
##      post.mean 1-95% CI u-95% CI eff.samp
## measure  0.02051 0.0003437 0.06454      100
##
## R-structure: ~units
##
##      post.mean 1-95% CI u-95% CI eff.samp
## units  0.06544 0.04926 0.08345      100
##
## Location effects: log(tarsus.mm) ~ foraging + log(body.g)
##
##      post.mean 1-95% CI u-95% CI eff.samp pMCMC
## (Intercept)      0.6468  0.1707  1.0198   145.7 <0.01 **
## foragingterrestrial 0.2583  0.1744  0.3627   139.5 <0.01 **
## log(body.g)       0.4574  0.3799  0.5290   100.0 <0.01 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- 
- EXERCISE 3: Include also the geographical region (“region”) as a random effect. Remember you will need to specify the prior for this new factor as well.

```
##
##      Africa  Americas Australasia  Eurasia Indomalayan
##          17          40          41          8          17
```

```
## [1] "species"    "tarsus.mm" "tail.mm"   "wing.mm"   "body.g"    "foraging"
## [7] "region"     "measure"

##
## Iterations = 101:1091
## Thinning interval = 10
## Sample size = 100
##
## DIC: 18.05949
##
## G-structure: ~measure
##
##           post.mean 1-95% CI u-95% CI eff.samp
## measure  0.01791 0.0005575 0.09667      100
##
##           ~region
##
##           post.mean 1-95% CI u-95% CI eff.samp
## region  0.009005 0.0002262 0.04367      100
##
## R-structure: ~units
##
##           post.mean 1-95% CI u-95% CI eff.samp
## units    0.0649 0.04853 0.07832      100
##
## Location effects: log(tarsus.mm) ~ foraging + log(body.g)
##
##           post.mean 1-95% CI u-95% CI eff.samp pMCMC
## (Intercept)      0.6159 0.1838 1.0301      100 <0.01 **
## foragingterrestrial 0.2694 0.1644 0.3895      100 <0.01 **
## log(body.g)       0.4609 0.3989 0.5344      100 <0.01 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

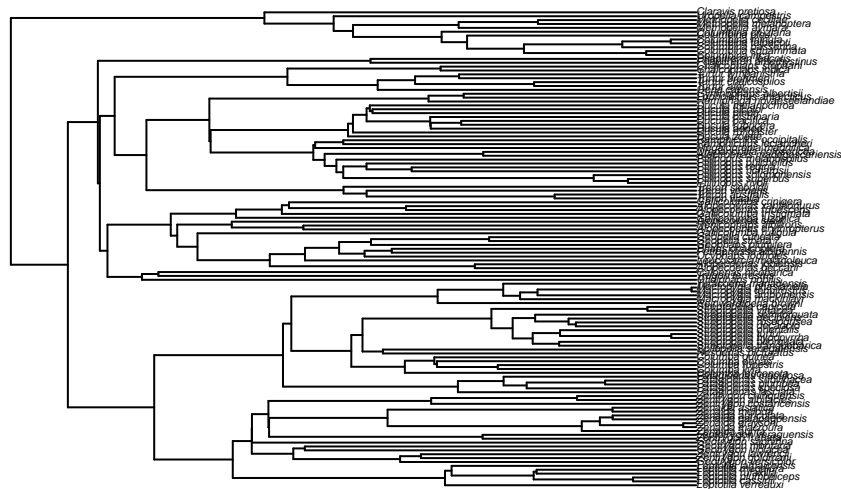
---

### PART 3: Phylogenetic effects and random variance.

As species are not independent of each other due to shared ancestry, we need to take this into account. MCMCglmm allows to include phylogenetic similarity as a random effect. For this, we only need a phylogenetic tree and a column in our data called 'animal' that corresponds to the phylogenetic tips of the tree. Phylogenetic effects can be seen as another type of autocorrelation in addition to spatial and temporal autocorrelation and arises because closely related species are expected to be more similar to each other than random species.

We open the tree and plot it:

```
ctree <- read.tree("data/ColumbidaeTree.tre")
plot(ctree,cex=0.3)
```



Now, we add a column in our data with the tips of the tree. We already have a column “species”, but MCMCglmm need a column names “animal” to associate with the tree.

```
cdata$animal <- cdata$species
prior3.1 <- list(G = list(G1 = list(nu=0.002, V=1), G2 = list(nu=0.002, V=1)),
  R = list(nu=0.002, V=1))
mod3.1 <- MCMCglmm(log(tarsus.mm) ~ 1+log(body.g),
  random= ~animal + measure,
  data = cdata, prior = prior3.1, verbose=F,
  pedigree=ctree,
  nitt = 11000, thin=10, burnin = 100)
summary(mod3.1)
```

```
##
## Iterations = 101:10991
## Thinning interval = 10
## Sample size = 1090
##
## DIC: -145.4256
##
## G-structure: ~animal
##
##          post.mean l-95% CI u-95% CI eff.samp
## animal    0.04832   0.0308  0.06488     870.4
##
##          ~measure
```

```
##
##          post.mean 1-95% CI u-95% CI eff.samp
## measure  0.04411 0.0002535 0.04077 894.3
##
## R-structure: ~units
##
##          post.mean 1-95% CI u-95% CI eff.samp
## units    0.01003 0.004427 0.01589 758.4
##
## Location effects: log(tarsus.mm) ~ 1 + log(body.g)
##
##          post.mean 1-95% CI u-95% CI eff.samp pMCMC
## (Intercept)  1.5871  1.1648  1.9706  1090 0.00183 **
## log(body.g)  0.3072  0.2406  0.3763  1517 < 9e-04 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can see that different random factors explain different proportion of the variance. We can explore the effect of each random factor:

```
posterior.mode(mod3.1$VCV)
```

```
##          animal          measure          units
## 0.0468323702 -0.0000579729 0.0075841073
```

However, it's more useful to report the intraclass correlation in relative terms, as the proportion of variance explained by each random factor. This is because we are interested in the proportion of variance explained by each factor, but not by the absolute variance explained (Which will vary a lot between variables, studies, samples,...).

The proportion of variance explained can be calculated by dividing the variance of factor X by the sum of all variances. So the proportion of variance explained by the phylogeny is:

```
total.variance <- sum(posterior.mode(mod3.1$VCV))
IC.animal <- posterior.mode(mod3.1$VCV)[1]/total.variance #animal variance
IC.animal
```

```
##          animal
## 0.8615463
```

IC.animal is expressed in relation to the total variance (1). If we want the % we can do:

```
round(IC.animal*100,2) #round to 2 decimals and multiple by 100
```

```
## animal
## 86.15
```

This is a useful propriety of the MCMCglmm models, as we can see which is the phylogenetic effect of traits or can calculate the repeatability of measurements, in case we have mutiple measurements for each specimen or species.



## FINAL EXERCISE (4): SpatialR and MCMCglmm into practise.

Now, you can do a final exercise to put into practise what you have learned in this tutorial, in the context of spatial analysis. For the spatial part, you will need to load these R-packages:

```
library(sp)
library(raster)
library(sf)
library(rgeos)
library(Imap)
```

We will continue to use the pigeon's dataset example. Now, you can open an additional dataset with some geographical information: `"data/ColumbidaeGeo.txt"`. This data includes coordinates for the breeding and non-breeding coordinates for each species. You can merge this dataset with the previous one, using `merge()`

```
gdata <- read.table("data/ColumbidaeGeo.txt",h=T,stringsAsFactors = F)
cdata <- merge(cdata,gdata,by="species")
```

Now, you can do two different exercises:

### EXERCISE 4A: Migratory behaviour.

Classify each species into a migratory or non-migratory species. To do so, you can calculate the migratory distance of each species (i.e. the distance between breeding and non-breeding coordinates). For this, you can use the function `gdist()` from the package *imap*.

I will here show how `gdist()` work on. Based on this you should be able to apply it to this specific case.

```
# generate a set of random coordinates
Longitude_1 = c(110,110,110,110)
Latitude_1= c(10,20,30,40)
cords1 <- cbind(Longitude_1,Latitude_1)
# and another set of random coordinates
Longitude_2 = c(100,110,120,130)
Latitude_2= c(15,20,25,30)
cords2 <- cbind(Longitude_2,Latitude_2)
#This line will calculate the distance between the the first set of points, the second set of points e.
gdist(Longitude_1,Latitude_1,Longitude_2,Latitude_2)
```

```
## [1] 658.3227 0.0000 611.2998 1149.4887
```

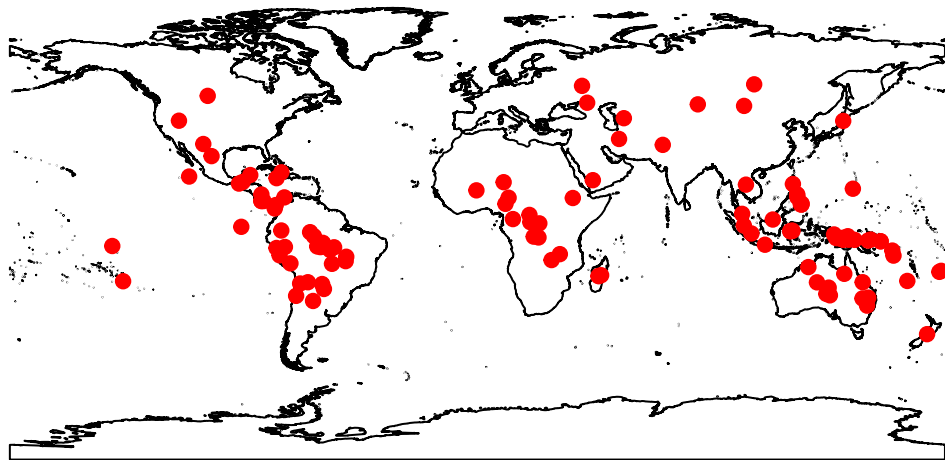
Once you get the migratory distance, you can classify species into resident (distance=0) and migratory (distance>0).

```
##
## no yes
## 114 9
```

### EXERCISE 4B: Insularity.

For this exercise, we will be using a world map with different attributes for islands(1) vs continents(0). This is a **shapefile** called `"MapIslands.shp"`. First, try to open and plot the map and the breeding coordinates.

```
## Reading layer 'MapIslands' from data source
##   '/Users/fsayol/My Drive/1_ResearchOrganization/5_Teaching/2022/221006_SpatialR/data/MapIslands/Map
##   using driver 'ESRI Shapefile'
## Simple feature collection with 4063 features and 1 field
## Geometry type: POLYGON
## Dimension:      XY
## Bounding box:   xmin: -180 ymin: -90 xmax: 180 ymax: 83.6341
## Geodetic CRS:   WGS 84
```



Now, you can classify each species into a island or continental species. We can use the function `over()` to know which breeding coordinates fall inside islands or continents.

```
# Use function over to get pts over polygons.
pts.island <- over(pts,WorldMap)
cdata$insularity <- as.numeric(pts.island$island)
table(cdata$insularity)
```

```
##
##  0  1
## 88 25
```

Note 1: Some coordinates might fall in the sea, with value=NA. To solve that we can try to know which is the closest polygon using the function `gDistance()` and then get the attributes (i.e. island or continent) of the closest polygon. Note 2: Be aware this is an advanced exercise. If you do not succeed after several trials, go directly to Exercise 4C and try to include “migration” in the MCMCglmm model.

```
## [1] 10

## [1] "Alopecoenas_rubescens" "Alopecoenas_stairi"
## [3] "Alopecoenas_xanthonurus" "Leptotila_cassinii"
## [5] "Leptotila_jamaicensis" "Leptotrygon_veraguensis"
## [7] "Phapitreron_leucotis" "Ramphiculus_occipitalis"
## [9] "Zenaida_galapagoensis" "Zentrygon_lawrencii"

## [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [13] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [25] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [37] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [49] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [61] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [73] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [85] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [97] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [109] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [121] FALSE FALSE FALSE

##               species lat.breeding lon.breeding insularity
## 5      Alopecoenas_rubescens    -7.9184600    -140.56860         1
## 6      Alopecoenas_stairi    -17.4582653     178.55336         1
## 7      Alopecoenas_xanthonurus    14.0943535     144.55943         1
## 48      Leptotila_cassinii     11.8815992     -83.15344         0
## 49      Leptotila_jamaicensis    19.3545752     -87.44644         0
## 54      Leptotrygon_veraguensis     6.5088162     -78.01634         0
## 76      Phapitreron_leucotis     11.8447608     123.09723         1
## 86      Ramphiculus_occipitalis    11.8663125     123.09635         1
## 114     Zenaida_galapagoensis    -0.5410993     -90.89218         1
## 123     Zentrygon_lawrencii      9.3668017     -82.34378         0

##
## continent    insular
##           92         31
```

#### EXERCISE 4C: MCMCglmm into practise

Use the information you calculates in exercises **5A** and **5B** (i.e. insularity and migratory behavior) as a predictor in a *MCMCglmm()*.

- Does the tarsus length vary between migratory and resident species?

```
##
## Iterations = 101:99101
## Thinning interval = 1000
## Sample size = 100
##
## DIC: -165.7783
##
## G-structure: ~animal
##
##           post.mean l-95% CI u-95% CI eff.samp
```

```
## animal    0.05067  0.03304  0.07183      100
##
## R-structure: ~units
##
##      post.mean l-95% CI u-95% CI eff.samp
## units  0.008114 0.003734  0.01493      195.3
##
## Location effects: log(tarsus.mm) ~ log(body.g) + migration
##
##      post.mean l-95% CI u-95% CI eff.samp pMCMC
## (Intercept)   1.64200  1.31667  2.06040      100 <0.01 **
## log(body.g)   0.29866  0.21738  0.36185      100 <0.01 **
## migrationyes -0.12323 -0.21657 -0.01277      100  0.04 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- And what about the wing length?

```
##
## Iterations = 101:99101
## Thinning interval = 1000
## Sample size = 100
##
## DIC: -181.1454
##
## G-structure: ~animal
##
##      post.mean l-95% CI u-95% CI eff.samp
## animal  0.01001 0.004588  0.01835      100
##
## R-structure: ~units
##
##      post.mean l-95% CI u-95% CI eff.samp
## units  0.009321 0.005254  0.01337      100
##
## Location effects: log(wing.mm) ~ log(body.g) + migration
##
##      post.mean l-95% CI u-95% CI eff.samp pMCMC
## (Intercept)   3.298527  3.049566  3.527051     134.4 <0.01 **
## log(body.g)   0.339051  0.293699  0.379494     147.1 <0.01 **
## migrationyes  0.077133 -0.004527  0.171790     100.0  0.12
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Does the tarsus length vary between island and continental species?

```
##
## Iterations = 101:99101
## Thinning interval = 1000
## Sample size = 100
##
## DIC: -142.0003
##
```

```
## G-structure: ~animal
##
##      post.mean l-95% CI u-95% CI eff.samp
## animal    0.04181 0.02789 0.05827   141.5
##
## R-structure: ~units
##
##      post.mean l-95% CI u-95% CI eff.samp
## units    0.01034 0.004463 0.01638   178.1
##
## Location effects: log(tarsus.mm) ~ log(body.g) + foraging + insularity
##
##      post.mean  l-95% CI  u-95% CI  eff.samp  pMCMC
## (Intercept)    1.352086  1.041141  1.832363    100.0 <0.01 **
## log(body.g)     0.319369  0.249780  0.387793    100.0 <0.01 **
## foragingterrestrial 0.223772 0.076144 0.368891    138.4 <0.01 **
## insularityinsular  0.084770 -0.005314 0.152632    100.0 0.06 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Can you do a scatterplot to show the data, using *ggplot()*? You can also try to plot the model results, as we did before.

