Introduction to Bayesian statistics with R 5. Case studies and GLMMs

Olivier Gimenez

last updated: 2025-03-10

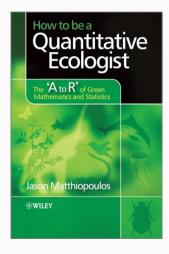
Introduction

Introduction

- GLMs avec les 3 composantes
- Intuition des GLMMs

GLMM with Poisson response

From Jason Matthiopoulos' book



Longitudinal study on coral reef

A survey of a coral reef uses 10 predefined linear transects covered by divers once every week. The response variable of interest is the abundance of a particular species of anemone as a function of water temperature. Counts of anemones are recorded at 20 regular line segments along the transect. The following piece of code will generate a data set with realistic properties according to the above design. Make sure you understand what it is doing. You might want to explain the script to the colleague next to you. Also, to try and make sense of the code of others, it is always good to plot and/or run small sections of the code.

Data generation

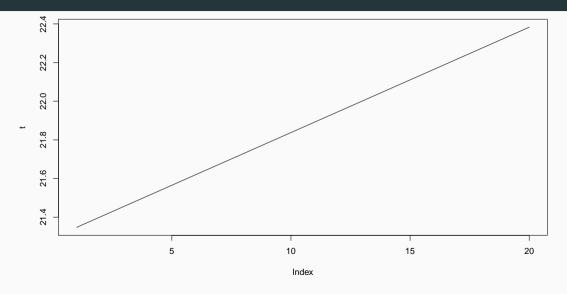
```
transects <- 10
data <- NULL
for (tr in 1:transects){
  # random effect (intercept)
  ref <- rnorm(1,0,.5)
  # water temperature gradient
  t \leftarrow runif(1, 18,22) + runif(1,-.2,0.2)*1:20
  # Anemone gradient (expected response)
  ans \leftarrow exp(ref -14 + 1.8 * t - 0.045 * t^2)
  # actual counts on 20 segments of the current transect
  an <- rpois(20, ans)
  data <- rbind(data, cbind(rep(tr, 20), t, an))</pre>
```

Our job

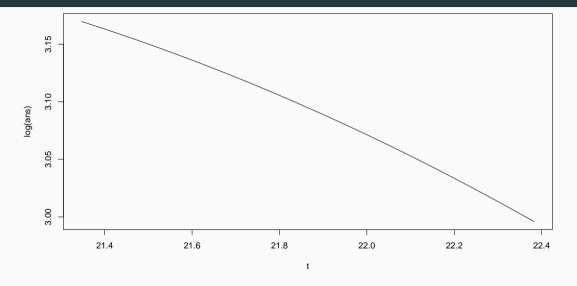
- Generate a data set using the anemone code.
- Using NIMBLE and brms, fit a GLMM with quadratic effect of temperature and a random intercept.
- Fit the same model to the same data in a Frequentist framework using function lme4::glmer().
- Compare the estimates.

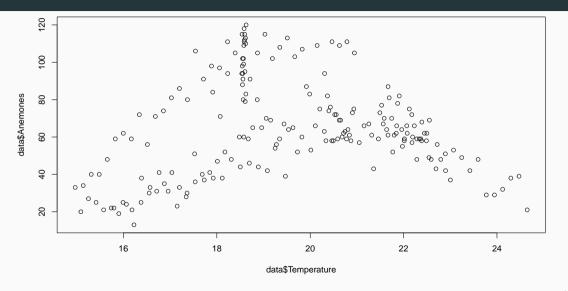
- Always difficult to make sense of the code of others.
- Good to plot and/or run small sections of the code.

```
# random effect (intercept)
ref <- rnorm(1,0,.5)
# water temperature gradient
t <- runif(1, 18,22) + runif(1,-.2,0.2)*1:20
plot(t,type='l')</pre>
```



```
# Anemone gradient (expected response)
ans <- exp(ref -14 + 1.8 * t - 0.045 * t^2)
plot(t,log(ans),type='l')</pre>
```





Write down model

NIMBLE

Load package

library(nimble)

Standardize temperature covariate

```
boo <- data$Temperature
data$Temp <- (boo - mean(boo)) / sd(boo)</pre>
head(data)
    Transect Temperature Anemones
                                   Temp
#> 1
         1 19.46887 39 -0.05470958
#> 2
             19.27560 56 -0.13851042
#> 3
             19.08233 42 -0.22231126
             18.88906 44 -0.30611209
#> 4
#> 5
         1 18.69579 46 -0.38991293
#> 6
             18.50253 44 -0.47371377
```

```
model <- nimbleCode({</pre>
  for (i in 1:n){
    count[i] ~ dpois(lambda[i])
    log(lambda[i]) <- intercept[transect[i]] + slope[1] * x[i] + slope[2] * pow(x[i],2)
  for (j in 1:nbtransects){
    intercept[j] ~ dnorm(mu.a, sd = sigma.a)
  mu.a \sim dnorm(0, sd = 1.5)
  sigma.a ~ dunif(0, 10)
  slope[1] \sim dnorm(0, sd = 1.5)
  slope[2] \sim dnorm(0, sd = 1.5)
})
```

Initial values

Parameters to monitor (and save)

```
parameters.to.save <- c("slope", "mu.a", "sigma.a")</pre>
```

Set up MCMC details

```
n.iter <- 5000
n.burnin <- 1000
n.chains <- 2
```

Run model

Post-process results

```
library (MCMCvis)

MCMCsummary (object = mcmc.output, round = 2)

#> mean sd 2.5% 50% 97.5% Rhat n.eff

#> mu.a 4.34 0.12 4.10 4.34 4.57 1.00 6789

#> sigma.a 0.36 0.10 0.21 0.34 0.61 1.00 1115

#> slope[1] 0.09 0.03 0.03 0.09 0.16 1.09 131

#> slope[2] -0.26 0.02 -0.29 -0.26 -0.22 1.00 262
```

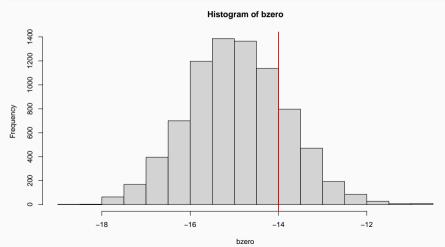
Convert regression coefficients from scaled to non-scaled and compare to values used to generate data (from https://stats.stackexchange.com/questions/361995/how-to-convert-coefficients-from-quadratic-function-from-scaled-to-not-scaled-co) Pull two chains together:

```
samples <- rbind(mcmc.output$chain1,mcmc.output$chain2)</pre>
```

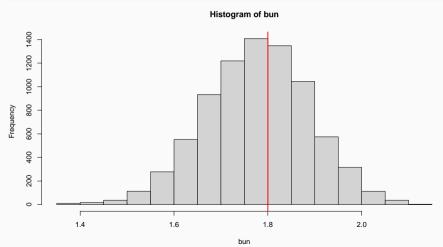
```
sbzero <- samples[,'mu.a']
sbun <- samples[,'slope[1]']
sbdeux <- samples[,'slope[2]']

mu <- mean(boo)
sg <- sd(boo)</pre>
```

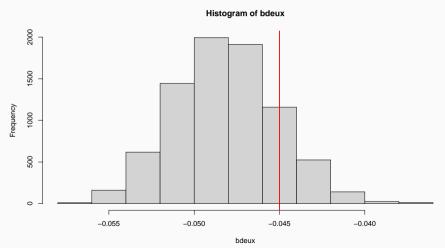
```
bzero <- sbzero - sbun*mu/sg + sbdeux*mu^2/(sg^2)
hist(bzero)
abline(v = -14, col = "red", lwd = 2)</pre>
```



```
bun <- sbdeux/sg - 2 * sbdeux * mu / (sg^2)
hist(bun)
abline(v = 1.8, col = "red", lwd = 2)</pre>
```



```
bdeux <- sbdeux/(sg^2)
hist(bdeux)
abline(v = -0.045, col = "red", lwd = 2)</pre>
```



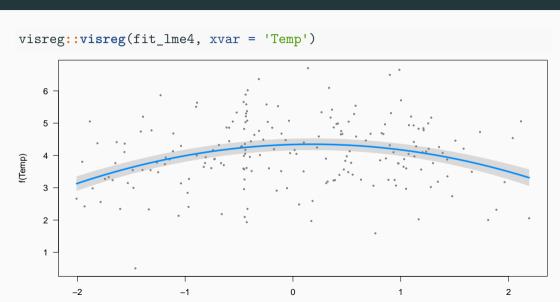
28

Frequentist approach

1me4 fit

```
library(lme4)
fit lme4 <- glmer(Anemones ~ Temp + I(Temp^2) + (1 | Transect), data = data, family = poisson)
fit 1me4
#> Generalized linear mixed model fit by maximum likelihood (Laplace
#> Approximation) [glmerMod]
#> Family: poisson (log)
#> Formula: Anemones ~ Temp + I(Temp^2) + (1 | Transect)
     Data: data
        ATC
                 BTC
                      logLik deviance df.resid
#> 1443.3781 1456.5714 -717.6891 1435.3781
#> Random effects:
#> Groups Name
                     Std.Dev.
#> Transect (Intercept) 0.2878
#> Number of obs: 200, groups: Transect, 10
#> Fixed Effects:
#> (Intercept) Temp I(Temp^2)
   4.33936 0.09095
                          -0.25586
```

Visualize



brms

Implementation

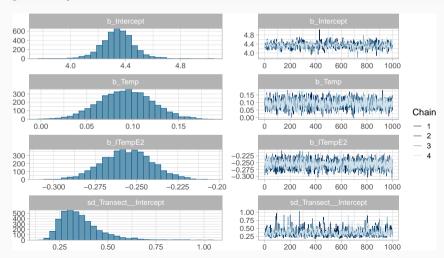
```
library(brms)
bayes.brms <- brm(Anemones ~ Temp + I(Temp^2) + (1 | Transect),
                  data = data.
                  family = poisson("log"),
                  chains = 2, # nb of chains
                  iter = 5000, # nb of iterations, including burnin
                  warmup = 1000, # burnin
                  thin = 1
```

Numerical summaries

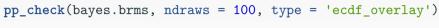
```
summary(bayes.brms)
#> Family: poisson
    l.inks: mu = l.oa
#> Formula: Anemones ~ Temp + I(Temp^2) + (1 | Transect)
     Data: data (Number of observations: 200)
#>
    Draws: 4 chains, each with iter = 2000: warmup = 1000: thin = 1:
#>
#>
           total post-warmup draws = 4000
#>
#> Multilevel Hyperparameters:
#> ~Transect (Number of levels: 10)
#>
                Estimate Est. Error 1-95% CI u-95% CI Rhat Bulk ESS Tail ESS
#> sd(Intercept)
                    0.36
                             0.11
                                      0.21
                                               0.63 1.00
                                                              697
                                                                     1083
#>
#> Regression Coefficients:
            Estimate Est. Error 1-95% CI u-95% CI Rhat Bulk ESS Tail ESS
#>
#> Intercept
               4.34
                        0.12
                                4.11
                                           4.58 1.01
                                                         757
                                                                  981
#> Temp
               0.09 0.03
                               0.04 0.15 1.00
                                                         2110
                                                                  2162
#> ITempE2
               -0.26 0.02
                                 -0.29
                                          -0.23 1.00
                                                         2324
                                                                  2210
#>
#> Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
#> and Tail ESS are effective sample size measures, and Rhat is the potential
#> scale reduction factor on split chains (at convergence, Rhat = 1).
```

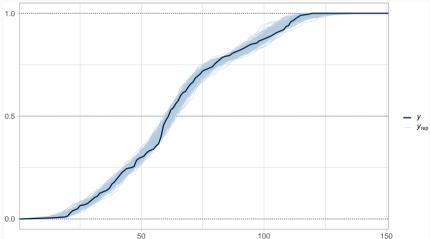
Convergence

plot(bayes.brms)



Quality of fit assessment





Model comparison

- What if we'd like to test the effect of temperature using WAIC?
- We fit a model with no effect of temperature:

Model comparison

Then we compare both models, by ranking them with their WAIC:

GLMM with binomial response

From a paper by Brice, Larsen and MacNulty (2021)

LETTER



Sampling bias exaggerates a textbook example of a trophic cascade



Elaine M. Brice¹ | Eric J. Larsen² | Daniel R. MacNulty¹

Department of Wildland Resources and Ecology Center, Utah State University, Logan, Utah, USA

²Department of Geography and Geology, University of Wisconsin – Stevens Point, Stevens Point, Wisconsin, USA

Correspondence

Elaine M. Brice, Department of Wildland Resources and Ecology Center, Utah State University, Logan, UT 84322, USA. Email: elbrice92@gmail.com

Funding information

National Science Foundation, Grant/ Award Number: DGE-1633756; Yellowstone National Park; University of Wyoming-National Park Service Small Grant Program, Grantl/Award Number: 1003867-USU; Utah State University

Editor: Tim Coulson

Abstract

Understanding trophic cascades in terrestrial wildlife communities is a major challenge because these systems are difficult to sample properly. We show how a tradition of non-random sampling has confounded this understanding in a textbook system (Yellowstone National Park) where carnivore [Canis lupus (wolf)] recovery is associated with a trophic cascade involving changes in herbivore [Cerrux canadensis (elk]) behaviour and density that promote plant regeneration. Long-term data indicate a practice of sampling only the tallest young plants overestimated regeneration of overstory aspen (Populus trenuloides) by a factor of 4-7 compared to random sampling because it favoured plants taller than the preferred browsing height of elk and overlooked non-regenerating aspen stands. Random sampling described a tophic cascade, but it was weaker than the one that non-random sampling described. Our findings highlight the critical importance of basic sampling principles (e.g. randomisation) for achieving an accurate understanding of trophic cascades in terrestrial wildlife systems.

KEYWORDS

aspen, carnivore, elk, non-random sampling, predator indirect effects, preferred browsing height, sampling bias, trophic cascade, ungulate, wolf

Read in data

```
rawdat <- read csv("dat/Aspen Data.csv")</pre>
head(rawdat)
#> # A tibble: 6 x 6
#> Plot Year Tree Browse Height Type
#> <dbl> <dbl> <dbl> <dbl> <dbl> <chr>
       1 2008
                       1 115 5T
#> 1
#> 2
    1 2008
                   1 130 5T
#> 3 1 2008
                       1 113 5T
#> 4
   1 2008
                       1 110 5T
#> 5
       1 2008
                       0 132 5T
#> 6
       1 2009
                            200 5T
```

Data collection (1)

We measured browsing and height of young aspen in 113 plots distributed randomly across the study area. Each plot was a 1 x 20 m belt transect located randomly within an aspen stand that was itself randomly selected from an inventory of stands with respect to high and low wolf-use areas. The inventory was a list of 992 grid cells (240 \times 360 m) that contained at least one stand. A 'stand' was a group of tree-size aspen (more than 10 cm diameter at breast height) in which each tree was less than 30 m from every other tree. One hundred and thirteen grid cells were randomly selected from the inventory, one stand was randomly selected from each cell, and one plot was randomly established in each stand.

Data collection (2)

We measured aspen at the end of the growing season (late July-September). focusing on plants less than 600 cm tall and more than 1 year old, which we termed 'young aspen'. For each stand, we measured every young aspen within a plot ('random stems') and each of the five tallest young aspen within the stand ('5T stems'). For all young aspen, we measured browsing status (browsed or unbrowsed) and height of the leader (tallest) stem. A leader was 'browsed' if its growth from the previous growing season had been eaten, which we identified by a sharp, pruned edge at the base of the current year's growth. Most plots were measured nearly every year since 1999 and our analysis focused on data from 10 years (2007-2014, 2016-2017) in which sampled stands included measurements of random and 5T stems.

Data collection (3)

We combined measurements of 5T and random stems into one data set of all stems (N = 18,623) across all years (N = 10 years).

Our job

- Read in the data and get familiar with it.
- Using NIMBLE and brms, fit a GLMM with 'Browsed' as the response variable.
 Consider year and type of stems (random or 5T stems) as explanatory variables.
 Use stand identity as a random effect on the intercept.
- Fit the same model to the same data in a Frequentist framework using function lme4::glmer().
- Compare the estimates.

Read in data

```
str(rawdat)
\# spc tbl [18,792 x 6] (S3: spec tbl df/tbl df/tbl/data.frame)
#> $ Plot : num [1:18792] 1 1 1 1 1 1 1 1 1 1 ...
#> $ Year : num [1:18792] 2008 2008 2008 2008 2008 ...
#> $ Tree : num [1:18792] 1 2 3 4 5 1 2 3 4 5 ...
   $ Browse: num [1:18792] 1 1 1 1 0 1 0 0 0 0 ...
   $ Height: num [1:18792] 115 130 113 110 132 200 208 218 240 260 ...
   $ Type : chr [1:18792] "5T" "5T" "5T" "5T" ...
   - attr(*, "spec")=
#>
     .. cols(
#>
     \dots Plot = col double(),
     \therefore Year = col_double(),
#>
#>
     .. Tree = col double().
#>
     ... Browse = col double(),
#>
     \dots Height = col double(),
     .. Type = col_character()
#>
#>
     ..)
   - attr(*, "problems")=<externalptr>
```

Metadata

- lacktriangle Plot: individual identifier for each of 113 plots distributed randomly across the study area. Each plot was a 1 imes 20 m belt transect located randomly within an aspen stand
- Year: year in which aspen was sampled
- Tree: individual identifier for each stem within a plot
- Browse: denotes the browsing status (browsed = 1,unbrowsed = 0) of the leader (tallest) stem. A leader was 'browsed' if its growth from the previous growing season had been eaten
- Height: height (cm) of the leader stem of each individual aspen
- Type: sampling method. Every young aspen within a plot is a "random" stem, and each of the five tallest young aspen within the stand is a "5T" stem.

Format data

Write down model

$$\begin{split} \mathsf{Brwsed}_i &\sim \mathsf{Bernoulli}(p_i) & \text{[likelihood]} \\ \mathsf{logit}(p_i) &= a_{\mathsf{ID}[i]} + b_1 \; \mathsf{year}_i + b_2 \; \mathsf{type}_i & \text{[linear model]} \\ a_j &\sim \mathsf{Normal}(\bar{a}, \sigma) & \text{[prior for varying intercepts]} \\ \bar{a} &\sim \mathsf{Normal}(0, 1.5) & \text{[prior for population mean]} \\ \sigma &\sim \mathsf{Uniform}(0, 10) & \text{[prior for standard deviation]} \\ b_1, b_2 &\sim \mathsf{Normal}(0, 1.5) & \text{[prior for slopes]} \end{split}$$

NIMBLE

Load package

library(nimble)

```
model <- nimbleCode({</pre>
  for (i in 1:n){
    v[i] ~ dbern(p[i])
    logit(p[i]) \leftarrow a[id[i]] + beta[1] * type[i] + beta[2] * year[i] + beta[3] * year[i] * type[i]
  for (j in 1:nblevels){
    a[j] ~ dnorm(mu.a, sd = sigma.a)
  for (k in 1:3){
    beta[k] \sim dnorm(0, sd = 1.5)
  mu.a \sim dnorm(0, sd = 1.5)
  sigma.a ~ dunif(0,10)
})
```

Initial values

```
init1 <- list(mu.a = -0.5, sigma.a = 0.1, beta = rnorm(3))
init2 <- list(mu.a = 0.5, sigma.a = 0.5, beta = rnorm(3))
initial.values <- list(init1, init2)</pre>
```

Parameters to monitor (and save)

```
parameters.to.save <- c("mu.a", "sigma.a", "beta")</pre>
```

Set up MCMC details

```
n.iter <- 5000
n.burnin <- 1000
n.chains <- 2
```

Run model

Post-process results

```
library (MCMCvis)

MCMCsummary (object = mcmc.output, round = 2)

#> mean sd 2.5% 50% 97.5% Rhat n.eff

#> beta[1] -0.42 0.09 -0.60 -0.43 -0.26 1.00 190

#> beta[2] -0.14 0.01 -0.15 -0.14 -0.12 1.03 192

#> beta[3] -0.08 0.01 -0.11 -0.08 -0.06 1.00 203

#> mu.a 1.57 0.13 1.31 1.57 1.82 1.00 1841

#> sigma.a 1.26 0.09 1.09 1.25 1.45 1.00 1228
```

Visualize (1)

```
# pool two chains together
samples <- rbind(mcmc.output$chain1, mcmc.output$chain2)</pre>
nsim <- nrow(samples)</pre>
# get values sampled in posterior distribution of regression parameters
beta1 <- samples[,'beta[1]']
beta2 <- samples[,'beta[2]']</pre>
beta3 <- samples[,'beta[3]']
intercept <- samples[,'mu.a']</pre>
# predict
logitp 5T <- matrix(NA, nrow = nsim, ncol = 11)</pre>
logitp random <- matrix(NA, nrow = nsim, ncol = 11)</pre>
for (i in 1:nsim){ # loop over simulations
  for (j in 1:11){ # loop over years
    logitp 5T[i,j] <- intercept[i] + beta1[i] * 1 + beta2[i] * j + beta3[i] * 1 * j</pre>
    logitp random[i,i] <- intercept[i] + beta1[i] * 0 + beta2[i] * i + beta3[i] * 0 * i</pre>
# hack-transorm
p_5T <- plogis(logitp_5T)</pre>
p_random <- plogis(logitp_random)</pre>
```

Visualize (2)

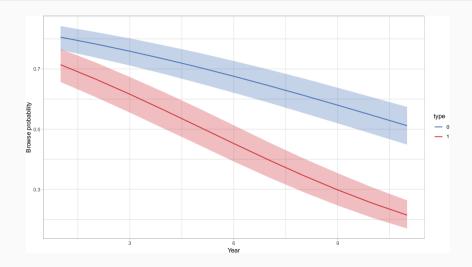
```
# means and credible intervals
p_5T_mean <- apply(p_5T, 2, mean)
p_5T_inf <- apply(p_5T, 2, quantile, probs = 2.5/100)
p_5T_sup <- apply(p_5T, 2, quantile, probs = 97.5/100)

p_random_mean <- apply(p_random, 2, mean)
p_random_inf <- apply(p_random, 2, quantile, probs = 2.5/100)
p_random_sup <- apply(p_random, 2, quantile, probs = 97.5/100)</pre>
```

Visualize (3)

```
data.frame(year = c(1:11, 1:11),
          type = factor(c(rep(1, 11), rep(0, 11))), # Convert type to a factor
          p_mean = c(p_5T_mean,p_random_mean),
          p_inf = c(p_5T_inf, p_random_inf),
          p \sup = c(p 5T \sup, p random sup)) %%
 ggplot() +
 geom ribbon(aes(x = year.
                 ymin = p_inf,
                 ymax = p_sup,
                 group = type,
                 fill = type),
             alpha = 0.3
             show.legend = F) +
    geom line(aes(x = year.
                 v = p mean.
                 group = type,
                 color = type),
             size=0.7) +
 scale_color_manual(values = c("1" = "#CC2529", "0" = "#396AB1")) +
 scale fill manual(values = c("1" = "#CC2529", "0" = "#396AB1")) +
 xlab("Year")+
 vlab("Browse probability")+
 theme light()
```

Visualize (3)



Frequentist approach

```
library(lme4)
fit lme4 <- glmer(y ~ type * year + (1|id), data = dat, family = "binomial")
fit 1me4
#> Generalized linear mixed model fit by maximum likelihood (Laplace
#> Approximation) [glmerMod]
#> Family: binomial (logit)
#> Formula: y ~ type * year + (1 | id)
     Data: dat
        ATC
                 BIC logLik deviance df.resid
#> 21428.86 21468.02 -10709.43 21418.86
                                        18618
#> Random effects:
#> Groups Name
                   Std.Dev.
   id (Intercept) 1.233
#> Number of obs: 18623, groups: id, 113
#> Fixed Effects:
#> (Intercept) type
                             year
                                        type:year
   1.57173 -0.41777 -0.13750
                                      -0.08506
```

Visualize

```
visreg::visreg(fit = fit_lme4, xvar = "year",
                 by = "type", overlay = TRUE, scale = "response")
                                         type: 0 ____ type: 1
  8.0
  0.7 -
  0.6
> 0.5
  0.4
  0.3 -
  0.2 -
                                                                            10
                                             year
```

brms

Implementation

Numerical summaries

```
summary(bayes.brms)
#> Familu: bernoulli
    Links: mu = logit
#> Formula: y ~ type * year + (1 | id)
     Data: dat (Number of observations: 18623)
    Draws: 2 chains, each with iter = 5000: warmup = 1000: thin = 1:
           total post-warmup draws = 8000
#>
#>
#> Multilevel Huperparameters:
#> ~id (Number of levels: 113)
              Estimate Est.Error 1-95% CI u-95% CI Rhat Bulk ESS Tail ESS
#>
#> sd(Intercept) 1.26 0.10
                                 1.08
                                            1.77 1.00
                                                                  1570
#>
#> Regression Coefficients:
           Estimate Est. Error 1-95% CI u-95% CI Rhat Bulk ESS Tail ESS
                                       1.82 1.01
#> Intercept 1.57 0.13
                              1.31
                                                    513
                                                               920
#> tupe
              -0.42 0.09
                              -0.59 -0.25 1.00
                                                   74.57
                                                               6505
         -0.14 0.01
                              -0.15 -0.13 1.00
                                                    10420
#> year
                                                               6334
#> type:year
              -0.09 0.01
                              -0.11
                                      -0.06 1.00
                                                     74.05
                                                               6314
#>
#> Draws were sampled using sampling(NUTS). For each parameter. Bulk ESS
#> and Tail ESS are effective sample size measures. and Rhat is the potential
#> scale reduction factor on split chains (at convergence, Rhat = 1).
```

Convergence

plot(bayes.brms)

