# Exploring the challenges

Amaël Le Squin

2024-12-02

> **!** Available on Gitlab
>
> This note is available on gitlab.

## Introduction

In order to test my models, I first run dummy tests, with generated data. I build models slowly, from the simplest to the most complex. The first challenge is related to the model used by Christine Deleuze *et al.* (scripts given on 11 October 2024):

```
Modele_mai2 = nlme(formTotNew ~ a + b*hdn + d*hsurd, data = Grdata.PV,
    start = c(a = 0.4, 0, b = 1.5, 0, d = 0.0005, 0),
    fixed = list(a + b + d ~ feuil.res), random = a + d ~ 1|nomessence2)
```

Few things raised a flag in my mind:

1. The R package `nlme`, non-linear mixed effect, is for repeated measured data and is based on Lindstrom and Bates (1990). Here, the repetition occurs within the species level, so I am not sure this package is the best adapted… I feel that `lme4` would have been better
2. `hdn` is the hardiness $\sqrt{c}/h$ while `hsurd` is what I call the 'slenderness', $h/c$. Therefore, there is a high (negative) correlation between these two explanatory variables!

The model `Modele_mai2` should be understood as follows, for an individual of height $h$ and circumference at breast height $c$, of species $j$ and functional group (conifer or broadleaf) $i$:

$$\begin{aligned} \mathcal{F} &\sim \boldsymbol{\mathcal{N}}(\mu, \sigma) \\ \mu_{i,j} &= a_{i,j} + b_i \frac{\sqrt{c}}{h} + d_{i,j} \frac{h}{c} \\ a_{i,j} &\sim \boldsymbol{\mathcal{N}}(\alpha_i, \sigma_\alpha) \\ d_{i,j} &\sim \boldsymbol{\mathcal{N}}(\delta_i, \sigma_\delta), \end{aligned} \tag{1}$$

where $\alpha_i$ and $\delta_i$ are the 'group intercepts' (common values within conifers and broadleaves). Therefore, it seems to be a GLMM, and I am not sure `nlme` is relevant here. In the next sections, I will try to reproduce their model with generated data. I will do it step-by-step. Before that, I rewrite the equations with new variable names and without the indices $i$ and $j$, in order to correspond to stan code:

$$\begin{aligned} \mathcal{F} &\sim \mathcal{N}(\mu, \sigma) \\ \mu &= \beta_0 + b_1 \frac{\sqrt{c}}{h} + \beta_2 \frac{h}{c} \\ \beta_0 &\sim \mathcal{N}(b_0, \sigma_0) \\ \beta_2 &\sim \mathcal{N}(b_2, \sigma_2), \end{aligned} \tag{2}$$

## Generating data

### Packages and helpers

First, I load the necessary packages:

```
#### Clear space and load packages
rm(list = ls())
graphics.off()

options(max.print = 500)

library(data.table)
library(cmdstanr)
    register_knitr_engine(override = TRUE)
library(stringi)
library(gt)

setHook(packageEvent("grDevices", "onLoad"),
function(...) grDevices::X11.options(type='cairo'))
options(device='x11')
```

> **Stan engine**
>
> By default, Quarto uses the knitr's built-in stan engine `rstan`. To override it so that all stan chunks are processed with `CmdStanR`, I need to specify:
>
> ```
> register_knitr_engine(override = TRUE)
> ```

Then, I define useful functions:

```r
#### Tool functions
## Tools
source("./toolFunctions.R")

## Generate k integers with constraint they sum to n >= k
generate_random_partition = function(n, k)
{
    if (n < k)
        stop("n should be larger than k")
    parts = c(0, sort(sample(1:(n - 1), k - 1)), n)
    return(diff(parts))
}
```

### Parameters

I define the following parameters, that will be used to generate data according to equation (1):

```r
#### Define parameters
set.seed(1969 - 08 - 18) # Woodstock seed

## Fixed effects
b0 = c(conif = 0.2, broad = 5.3)
b1 = c(conif = 0.27, broad = 1.4)
b2 = c(conif = 4, broad = 0.21)

## Variance for random effects and residuals
sigma_beta0 = 4.2
sigma_beta2 = 6.98
sigma = 1.2

## Number of data and species
n = 3e3
S = 30 # Number of species
```

We aim to recover them later with a statiscal model (the values are stored in Table 1).

### Data

```
rep_species = generate_random_partition(n, S)

lim_broadleaf = sum(rep_species[1:19]) + 1

n_conif = lim_broadleaf - 1
n_broad = n - n_conif

fake_dt = data.table(
    species = rep(paste0("sp_", 1:S), times = rep_species),
    type = c(rep("conif", n_conif), rep("broad", n_broad)),
    fake_hdn = runif(n = n, min = -10, max = 50),
    fake_slenderness = rnorm(n = n, mean = 0, sd = 20),
    b0 = c(rep(b0["conif"], n_conif), rep(b0["broad"], n_broad)),
    b1 = c(rep(b1["conif"], n_conif), rep(b1["broad"], n_broad)),
    b2 = c(rep(b2["conif"], n_conif), rep(b2["broad"], n_broad)))

fake_dt[, beta0 := rnorm(1, unique(b0), sigma_beta0), by = species]
fake_dt[, beta2 := rnorm(1, unique(b2), sigma_beta2), by = species]

b0m = fake_dt[, round(mean(unique(beta0)), 3), by = type]
b0m[, sig := fake_dt[, round(sd(unique(beta0)), 3)]]
setkey(b0m, type)

b2m = fake_dt[, round(mean(unique(beta2)), 3), by = type]
b2m[, sig := fake_dt[, round(sd(unique(beta2)), 3)]]
setkey(b2m, type)

fake_dt[, fake_mu := beta0 + b1*fake_hdn + beta2*fake_slenderness]
fake_dt[, fake_vol := rnorm(.N, fake_mu, sigma)]
fake_dt[, fake_vol_no_randeff := rnorm(.N, b0 + b1*fake_hdn + b2*fake_slenderness, sigma)]

ind_species = fake_dt[, .(start = .I[1], end = .I[.N]), by = .(species)]
ind_species[, n_indiv := end - start + 1, by = species]
ind_species[, sum(n_indiv)] == n
ind_species = merge.data.table(ind_species, unique(fake_dt[, .(species, type)]))
setorder(ind_species, start)

n_sp = ind_species[, .N, by = type]
setkey(n_sp, type)

temporary = lm(fake_dt[, fake_vol] ~ 0 + fake_dt[, fake_mu])
```

```
sig_est = round(summary(temporary)$sigma, 3)
```

Here is a summary of the parameters value $\beta_0$ and $\beta_2$ (simulated), and $b_0$ and $b_2$, which are supposed to be the mean of $\beta_0$s and $\beta_2$s:

Table 1: Parameters value for both functional groups

| Parameter | Conifer | Broadleaf |
|---:|---|---|
| $b_0$ | 0.2 | 5.3 |
| $\beta_0$ (data) | 0.27 | 6.014 |
| $\sigma_0$ | 4.2 | 4.2 |
| $\sigma_0$ (data) | 5.186 | 5.186 |
| $b_1$ | 0.27 | 1.4 |
| $b_2$ | 4 | 0.21 |
| $\beta_2$ (data) | 4.896 | -0.16 |
| $\sigma_2$ | 6.98 | 6.98 |
| $\sigma_2$ (data) | 7.974 | 7.974 |
| $\sigma$ | 1.2 | 1.2 |
| $\sigma$ (lm) | 1.199 | 1.199 |

As can be seen, the data do not necessarily represent the true parameters value very well.

## Parameter recovery

> Caching execution
>
> To store some results that could be slow to obtain, or to avoid the recompilation of stan code, use the option `cache = TRUE`.

### Simplest model

I first define the simplest possible model, where there is no hierarchy, and I only try to estimate $b_0$, $b_1$, $b_2$, and the residual variance $\sigma$:

Prepare the data for Stan,

```
#### Stan data
## Data list
stanData = list(
    N = fake_dt[, .N],
    S = S,
    n_sp_conif = n_sp["conif", N],
    n_sp_broad = n_sp["broad", N],
    ind_start_conif = ind_species[type == "conif", start],
    ind_start_broad = ind_species[type == "broad", start],
    ind_end_conif = ind_species[type == "conif", end],
    ind_end_broad = ind_species[type == "broad", end],
    lim_broadleaf = lim_broadleaf,
    fake_hdn = fake_dt[, fake_hdn],
    fake_slenderness = fake_dt[, fake_slenderness],
    volume_m3 = fake_dt[, fake_vol_no_randeff]
)

## Common variables
n_chains = 4
iter_warmup = 500
iter_sampling = 1500
```

and run the simplest model (see Listing 1):

It gives the following results (see Table 1 for real values):

```
info_dt = pretty_summary(fit = fit, params = fit$metadata()$model_params[-1]) # -1 to remove
info_dt |>
    gt() |>
    cols_label(
        params_name = "Parameter",
        mean_params = "Mean",
        sd_params = "Std. dev",
        r_hat_params = "r hat"
    ) |>
    fmt_number(
        n_sigfig = 2
    ) |>
    tab_style(
        style = cell_borders(sides = "all", style = NULL),
        locations = cells_body()
    ) |>
```

| Parameter | Mean | Std. dev | r hat |
|---|---|---|---|
| b0[1] | 0.25 | 0.047 | 1.0 |
| b0[2] | 5.3 | 0.046 | 1.0 |
| b1[1] | 0.27 | 0.0018 | 1.0 |
| b1[2] | 1.4 | 0.0017 | 1.0 |
| b2[1] | 4.0 | 0.0016 | 1.0 |
| b2[2] | 0.21 | 0.0015 | 1.0 |
| sigma | 1.2 | 0.015 | 1.0 |

```
    tab_style(
        style = cell_text(weight = "bold"),
        locations = list(cells_column_labels(), cells_column_spanners())
    ) |>
    tab_style(
        style = cell_text(align = "right"),
        locations = cells_body(columns = params_name)
    )
```

### Hierarchical model

> ! Quick and dirty way (which does not always work)
>
> Simply adding two priors on $\sigma_0$ and $\sigma_2$ like bellow does not necessarily work (it does in my case because the data are informative enough):
>
> ```
> target += inv_gamma_lpdf(sigma_beta0 | 1, 1);
> target += inv_gamma_lpdf(sigma_beta2 | 1, 1);
> ```
>
> When there is not enough groups to describe the population parameters $b_0$, $\sigma_0$ and $b_2$, $\sigma_2$ the model can turn unidentifiable. In this case, a solution can be a non-centred parametrisation. Another example is available here. Similarily, when the local data (*i.e.,* within each group) is not 'informative enough', then the individual likelihood functions can degenerate (paragraph 3 of the non-centred parametrisation link).

A slightly more complex model, this time with a group effect (*i.e.,* random effect $\beta_0$s and $\beta_2$s to estimate), and the associated two variances $\sigma_0$ and $\sigma_2$:

```
data {
    // Dimensions and indices
```

```
    int N; // Number of individuals
    int S; // Number of species
    int<lower = 0, upper = S> n_sp_conif; // number of conifer species
    int<lower = S - n_sp_conif, upper = S - n_sp_conif> n_sp_broad; // number of broadleaf sp
    array[n_sp_conif] int ind_start_conif; // Conifer species index start
    array[n_sp_broad] int ind_start_broad; // Broadleaf species index start
    array[n_sp_conif] int ind_end_conif; // Conifer species index end
    array[n_sp_broad] int ind_end_broad; // Broadleaf species index end

    // Predictors
    vector [N] fake_hdn;
    vector [N] fake_slenderness;

    // Response variable
    vector[N] volume_m3;
}

parameters {
    // Fixed effects (population parameters)
    vector[2] b0;
    vector[2] b1;
    vector[2] b2;

    // Random effects (group parameters)
    vector[S] beta0;
    vector[S] beta2;

    // Variances
    real<lower = 0> sigma; // sd residuals
    real<lower = 0> sigma_beta0; // sd random effect beta0
    real<lower = 0> sigma_beta2; // sd random effect beta2
}

model {
    // Priors
    // --- Population parameters
    target += normal_lpdf(b0 | 0, 10);
    target += normal_lpdf(b1 | 0, 10);
    target += normal_lpdf(b2 | 0, 10);

    // --- Residual variance and population variance
    target += inv_gamma_lpdf(sigma | 1, 1); // Uses shape and scale
```

```
    target += inv_gamma_lpdf(sigma_beta0 | 1, 1);
    target += inv_gamma_lpdf(sigma_beta2 | 1, 1);

    // Hierarchy
    target += normal_lpdf(beta0[1:n_sp_conif] | b0[1], sigma_beta0);
    target += normal_lpdf(beta2[1:n_sp_conif] | b2[1], sigma_beta2);
    target += normal_lpdf(beta0[(n_sp_conif + 1):S] | b0[2], sigma_beta0);
    target += normal_lpdf(beta2[(n_sp_conif + 1):S] | b2[2], sigma_beta2);

    for (i in 1:n_sp_conif)
    {
        // Likelihood conifers
        target += normal_lpdf(volume_m3[ind_start_conif[i]:ind_end_conif[i]] | beta0[i] +
            b1[1]*fake_hdn[ind_start_conif[i]:ind_end_conif[i]] +
            beta2[i]*fake_slenderness[ind_start_conif[i]:ind_end_conif[i]], sigma);
    }

    for (i in 1:n_sp_broad)
    {
        // Likelihood broadleaves
        target += normal_lpdf(volume_m3[ind_start_broad[i]:ind_end_broad[i]] | beta0[n_sp_co
            b1[2]*fake_hdn[ind_start_broad[i]:ind_end_broad[i]] +
            beta2[n_sp_conif + i]*fake_slenderness[ind_start_broad[i]:ind_end_broad[i]], sig
    }
}
```

This time, the model needs the data generated with the random effects:

```
stanData[["volume_m3"]] = fake_dt[, fake_vol]
```

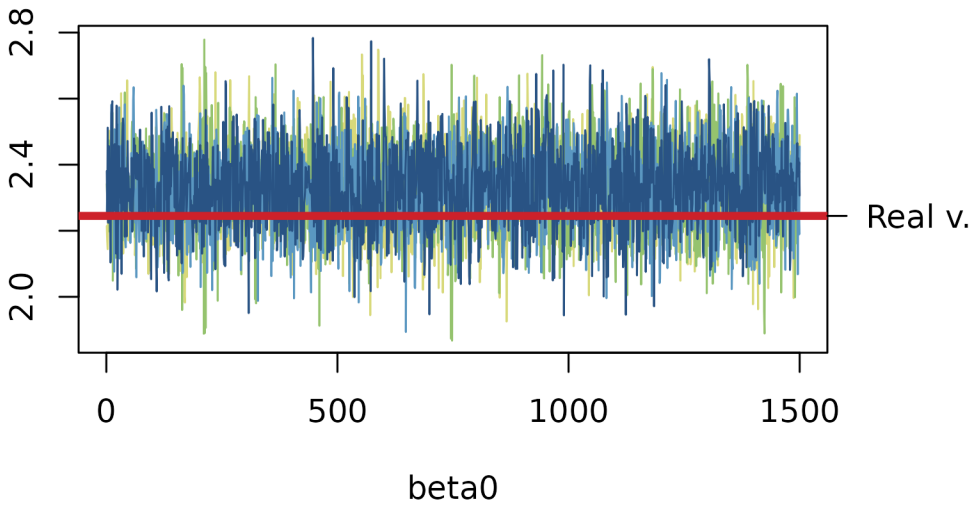
and now run the hierarchical model:

```
fit = hierarchical_model$sample(data = stanData, chains = n_chains, parallel_chains = ifelse
    refresh = 200, iter_warmup = iter_warmup, iter_sampling = iter_sampling)
```

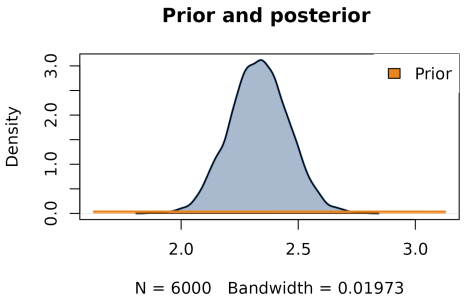
This gives the following results (see Table 1 for real values):

### Data with correlations

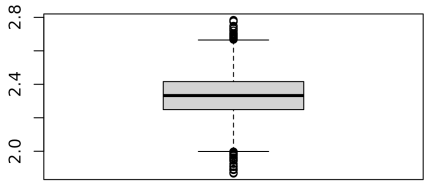
Now, we consider the same model, but with a correlation between slenderness and hardiness, and we run the same hierarchical model.




(a) Trace plot

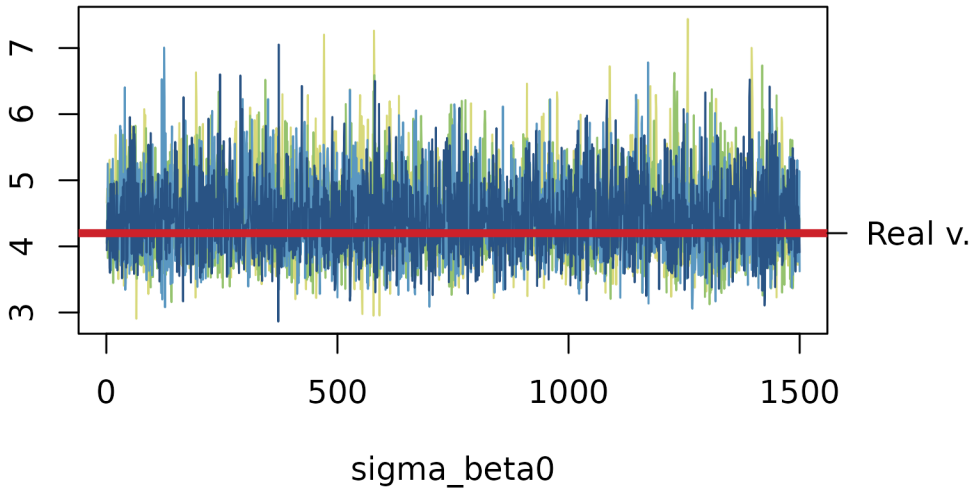



(b) Posterior

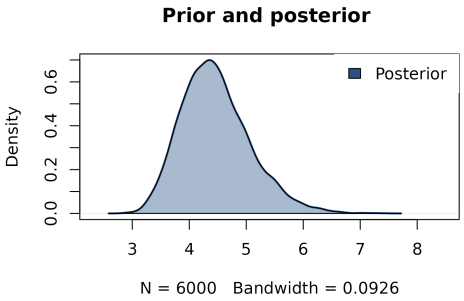



(c) Boxplot

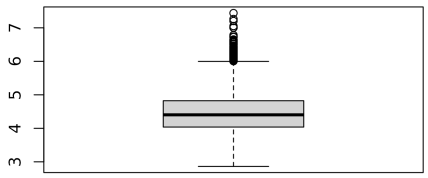
Figure 1: Trace plot and posterior of $\beta_0$ for the fifth species




(a) Trace plot




(b) Posterior




(c) Boxplot

Figure 2: Trace plot and posterior of $\sigma_0$

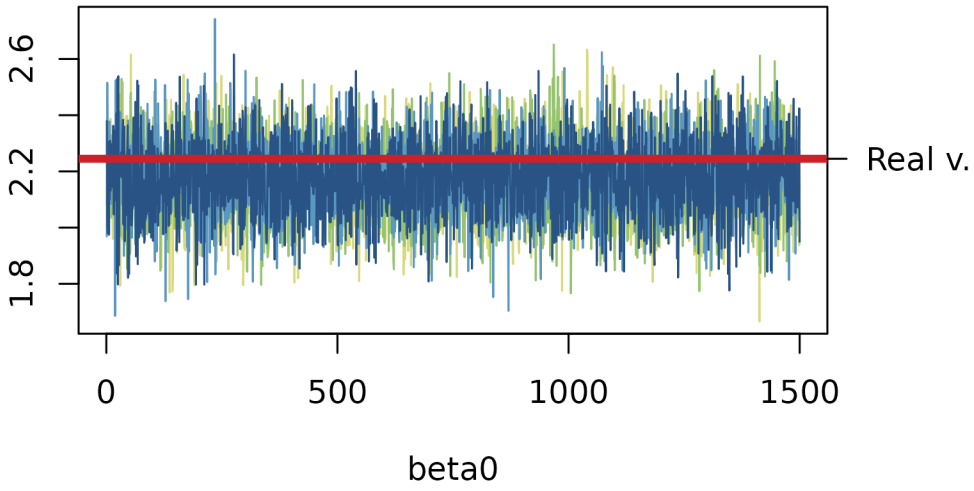
| Parameter | Mean | Std. dev | r hat |
|---|---|---|---|
| b0[1] | 0.26 | 1.1 | 1.0 |
| b0[2] | 5.9 | 1.3 | 1.0 |
| b1[1] | 0.27 | 0.0018 | 1.0 |
| b1[2] | 1.4 | 0.0017 | 1.0 |
| b2[1] | 4.8 | 1.7 | 1.0 |
| b2[2] | -0.15 | 2.3 | 1.0 |
| sigma | 1.2 | 0.016 | 1.0 |
| sigma_beta0 | 4.5 | 0.60 | 1.0 |
| sigma_beta2 | 7.8 | 1.1 | 1.0 |

| Parameter | Mean | Std. dev | r hat |
|---|---|---|---|
| b0[1] | 0.26 | 1.1 | 1.0 |
| b0[2] | 6.0 | 1.4 | 1.0 |
| b1[1] | 0.27 | 0.0021 | 1.0 |
| b1[2] | 1.4 | 0.0020 | 1.0 |
| b2[1] | 4.7 | 1.8 | 1.0 |
| b2[2] | -0.18 | 2.3 | 1.0 |
| sigma | 1.2 | 0.015 | 1.0 |
| sigma_beta0 | 4.5 | 0.63 | 1.0 |
| sigma_beta2 | 7.8 | 1.1 | 1.0 |

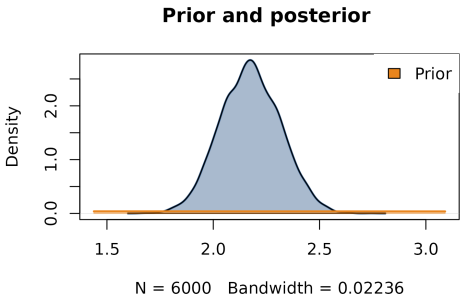
```
set.seed(1969 - 08 - 18) # Woodstock seed
fake_dt[, fake_slenderness_cor := rnorm(n = .N, mean = 1.2 - 0.5*fake_hdn, sd = 15)]

fake_dt[, fake_mu_cor := beta0 + b1*fake_hdn + beta2*fake_slenderness_cor]
fake_dt[, fake_vol_cor := rnorm(.N, fake_mu_cor, sigma)]

stanData[["fake_slenderness"]] = fake_dt[, fake_slenderness_cor]
stanData[["volume_m3"]] = fake_dt[, fake_vol_cor]

fit = hierarchical_model$sample(data = stanData, chains = n_chains, parallel_chains = ifelse
    refresh = 200, iter_warmup = iter_warmup, iter_sampling = iter_sampling, adapt_delta = 0
```

With a correlation of -0.5 (-0.51 in the real data), we get the following results (see Table 1 for real values):

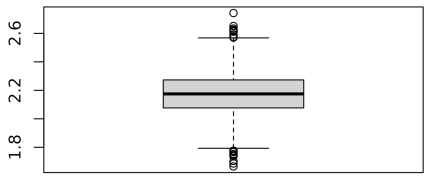
Which still works in our case…




(a) Trace plot

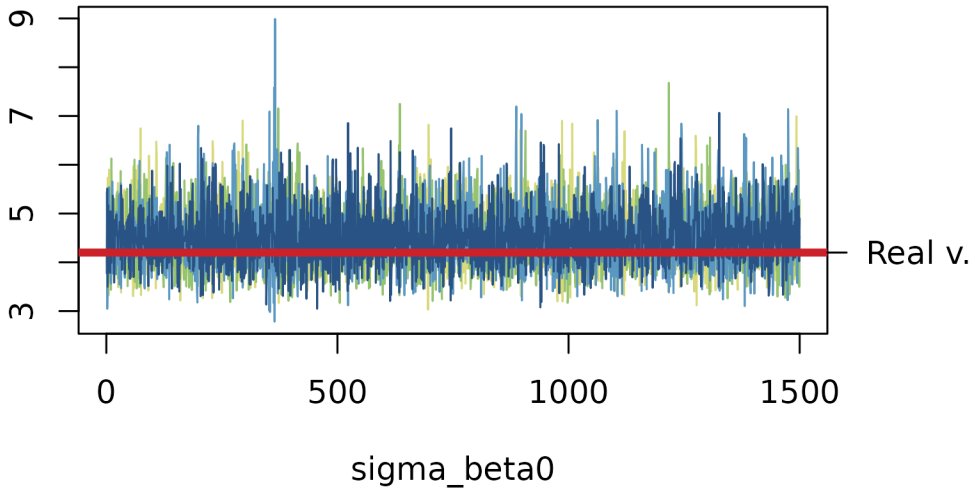



(b) Posterior

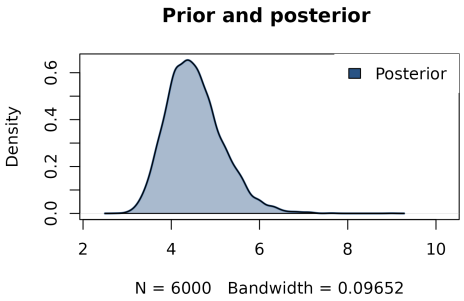



(c) Boxplot

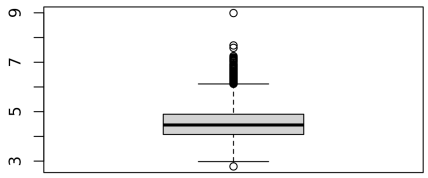
Figure 3: Trace plot and posterior of $\beta_0$ for the fifth species




(a) Trace plot




(b) Posterior




(c) Boxplot

Figure 4: Trace plot and posterior of $\sigma_0$

Note that I made the test with $\sigma = 35$ (instead of $\sigma = 1.2$). In other words, I put the data less informative. It was still working, although the parameters' variance increased a lot (makes sense!). I do not show these results.

Lindstrom, Mary J, and Douglas M Bates. 1990. "Nonlinear Mixed Effects Models for Repeated Measures Data." *Biometrics* 46 (3): 673. https://doi.org/10.2307/2532087.

**Listing 1**

```
data {
    // Dimensions and indices
    int N; // Number of individuals
    int lim_broadleaf; // Number of individuals

    // Predictors
    vector [N] fake_hdn;
    vector [N] fake_slenderness;

    // Response variable
    vector[N] volume_m3;
}

parameters {
    // Fixed effects
    vector[2] b0;
    vector[2] b1;
    vector[2] b2;

    // Variance
    real<lower = 0> sigma; // sd residuals
}

model {
    // Priors
    target += normal_lpdf(b0 | 0, 10);
    target += normal_lpdf(b1 | 0, 10);
    target += normal_lpdf(b2 | 0, 10);

    target += inv_gamma_lpdf(sigma | 1, 1);

    target += normal_lpdf(volume_m3[1:(lim_broadleaf - 1)] | b0[1] +
        b1[1]*fake_hdn[1:(lim_broadleaf - 1)] +
        b2[1]*fake_slenderness[1:(lim_broadleaf - 1)], sigma);

    target += normal_lpdf(volume_m3[lim_broadleaf:N] | b0[2] +
        b1[2]*fake_hdn[lim_broadleaf:N] +
        b2[2]*fake_slenderness[lim_broadleaf:N], sigma);
}
```