Hierarchical Poisson Models for Earthquake Inference

Ryuta Yoshimatsu

```
library(stats)
library(dplyr)
library(ggplot2)
library(devtools)
library(gridExtra)
library(grid)
library(tidyverse)
library(knitr)
library(lubridate)
library(reshape2)
library(vcd)
library(maps)
library(resample)
library(rjags)
library(kableExtra)
```

1 Executive Summary

We use historical data of earthquakes and estimate probabilities of large earthquakes occurring in eight different countries in year 2022. The Bayesian framework for inference is used and Poisson processes are assumed for temporal occurrences of earthquakes. Posterior distributions for the expected mean of rate for the different countries are obtained, however, shortcomings of the current approach are also identified. These include: 1. the data do not strictly follow Poisson process and 2. there is a presence of possible correlation between the observations.

2 Introduction

In some parts of the world, earthquakes pose great natural threats to humans which cause immense damage and affect many people's lives. Despite its relevance, predicting occurrences of major earthquakes remains a big challenge. In this report, we present an attempt to estimate the probability of large earthquakes occurring in a given country within the year 2022 using the Bayesian framework of inference.

3 Data

Our goal is to make statistical inference for the probabilities of earthquake occurring in the future. To this end, we use the historical data of earthquakes, which we collect from a worldwide earthquake catalog from the United States Geological Survey database (https://earthquake.usgs.gov/earthquakes/search/). We analyze the earthquakes that occurred after 1970 with the magnitude greater than 7.0. We restricted our data set to these conditions because the records prior to 1970 and below the magnitude 6.0 could be prone to errors and incompleteness (see the website for details).

```
earthquakes <- read.csv('earthquakes_processed.csv', header = TRUE)
earthquakes$time <- as.Date(substr(earthquakes$time,1,10), format('%Y-%m-%d'))
```

```
earthquakes$year <- year(earthquakes$time)
print(head(earthquakes), row.names=FALSE)
```

```
##        time latitude longitude mag     country year
##  2021-11-28  -4.4528  -76.8109 7.5        Peru 2021
##  2021-10-02 -21.1265  174.8958 7.3     Vanuatu 2021
##  2021-09-08  16.9465  -99.7530 7.0      Mexico 2021
##  2021-08-14  18.4335  -73.4822 7.2       Haiti 2021
##  2021-08-11   6.4748  126.7151 7.1 Philippines 2021
##  2021-07-29  55.3635 -157.8876 8.2         USA 2021
```

```
print(tail(earthquakes), row.names=FALSE)
```

```
##        time latitude longitude mag     country year
##  1970-05-27   27.236   140.230 7.1       Japan 1970
##  1970-04-29   14.520   -92.653 7.3      Mexico 1970
##  1970-04-07   15.791   121.630 7.4 Philippines 1970
##  1970-02-28   52.487  -174.915 7.1         USA 1970
##  1970-01-10    6.785   126.682 7.2 Philippines 1970
##  1970-01-04   24.185   102.543 7.1       China 1970
```

The earthquakes in the data set are visualized on the world map below (Figure 1). Note that earthquakes that occurred far out into the sea are omitted from the data set.

```
world_map <- map_data("world")
p <- ggplot() + coord_fixed() + xlab("") + ylab("")
base_world_messy <- p + geom_polygon(data=world_map, aes(x=long, y=lat, group=group), colour="light green"
cleanup <- theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
        panel.background = element_rect(fill = 'white', colour = 'white'),
        axis.line = element_line(colour = "white"), legend.position="none",
        axis.ticks=element_blank(), axis.text.x=element_blank(),
        axis.text.y=element_blank())
base_world <- base_world_messy + cleanup
map_data <- base_world + geom_point(data=earthquakes, aes(x=longitude, y=latitude, size=mag), colour="Deep
print(map_data)
```

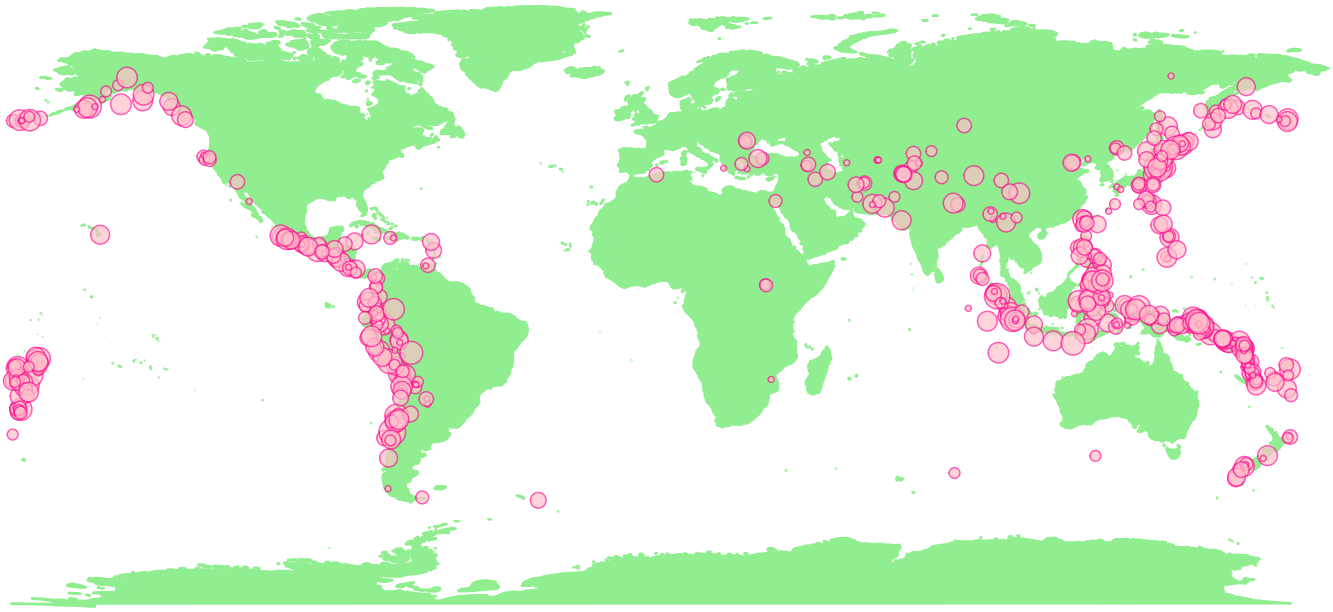


Figure 1: World map with major earthquakes since 1970.

Table 1: The earthquake dataset.

year	Japan	Philippines	Solomon.Islands	USA	Guinea	Indonesia	Russia	Vanuatu
2019	0	0	0	0	2	2	0	0
2020	0	0	0	2	1	0	2	0
2021	2	2	0	1	0	0	0	1

```
agg <- earthquakes %>%
  group_by(year, country) %>%
  summarise(count=n(), .groups = 'drop') %>%
  pivot_wider(names_from = country, values_from = count, values_fill = 0)
names(agg) <- make.names(names(agg), unique=TRUE)
```

We limit the number of countries for the analysis to eight. The choice of this number is completely arbitrary and this is solely for brevity's sake. We take the top countries with the largest number of earthquakes into our analysis. The sample data shown below holds the number of earthquakes by year per country (see Table 1). The same data are shown in a time series format in Figure 2.

```
# Select countries with total earthquakes >= 30 (this filter is arbitrary)
cols <- which(colSums(agg) >= 30)
agg <- agg %>% select(cols)
kable(tail(agg,3), caption="The earthquake dataset.") %>% kable_styling(html_font=8)
```

```
print(summary(agg[,-1]))
```

```
##      Japan         Philippines     Solomon.Islands       USA
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :1.0000   Median :1.0000   Median :0.0000   Median :1.0000
##  Mean   :0.9231   Mean   :0.7692   Mean   :0.5962   Mean   :0.7115
##  3rd Qu.:2.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :6.0000   Max.   :3.0000   Max.   :4.0000   Max.   :2.0000
##      Guinea        Indonesia         Russia          Vanuatu
##  Min.   :0.000   Min.   :0.000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.000   1st Qu.:0.000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :1.000   Median :1.000   Median :0.0000   Median :0.0000
##  Mean   :1.077   Mean   :1.423   Mean   :0.6731   Mean   :0.8077
##  3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :5.000   Max.   :5.000   Max.   :4.0000   Max.   :3.0000
```

```
agg_melt <- melt(agg, id.vars = 'year', variable.name = 'country')
print(ggplot(agg_melt, aes(year, value)) + geom_line(aes(colour = country)))
```

We plot the auto-correlations of the time series (see Figure 3) to evaluate the that the independence of the earthquakes. We acknowledge the level of auto-correlation in the plot for Guinean and other countries. These are interesting observations and deserve further investigations, but it is for now out of scope for this assignment and, therefore, we proceed with our analysis.

```
p <- c()
i <- 1
for(column in colnames(agg))
{
  if (column != "year")
  {
    bacf <- acf(agg[[column]], plot=FALSE)
    bacfdf <- with(bacf, data.frame(lag, acf))
    p[[i]] <- ggplot(data=bacfdf, mapping=aes(x=lag, y=acf)) +
```

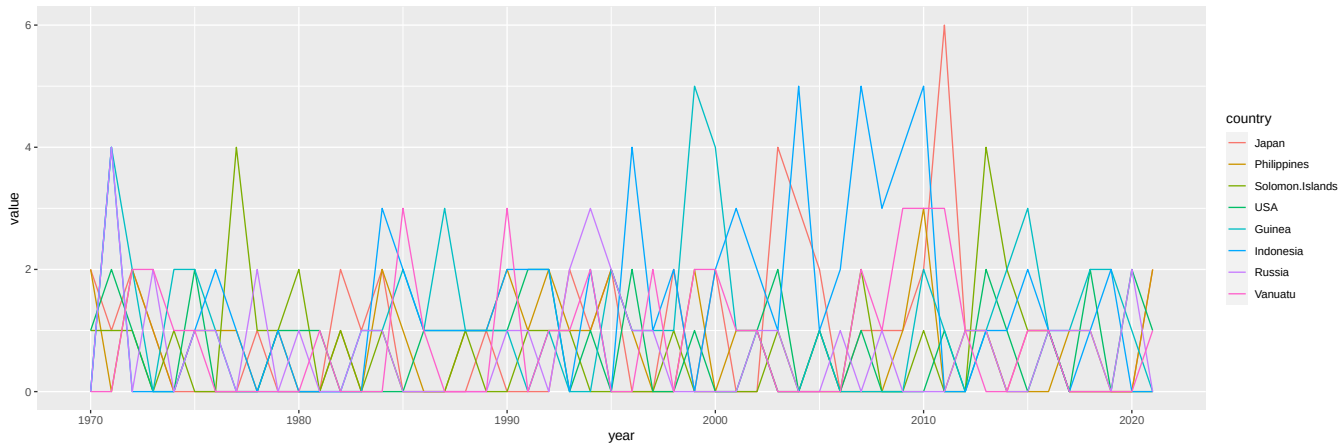


Figure 2: Time series of earthquake occurances.

```
      geom_segment(mapping=aes(xend=lag, yend=0)) +
      geom_hline(aes(yintercept = 0.2), linetype = 3, color = 'darkblue') +
      geom_hline(aes(yintercept = -0.2), linetype = 3, color = 'darkblue') +
      ggtitle(column) +
      theme(plot.title = element_text(hjust = 0.5))
    i = i + 1
  }
}

grid.arrange(
  p[[1]], p[[2]], p[[3]], p[[4]], p[[5]], p[[6]], p[[7]], p[[8]],
  nrow=2,
  bottom = textGrob(
    "",
    gp = gpar(fontface=3, fontsize=9),
    hjust=1,
    x=1
  )
)
```

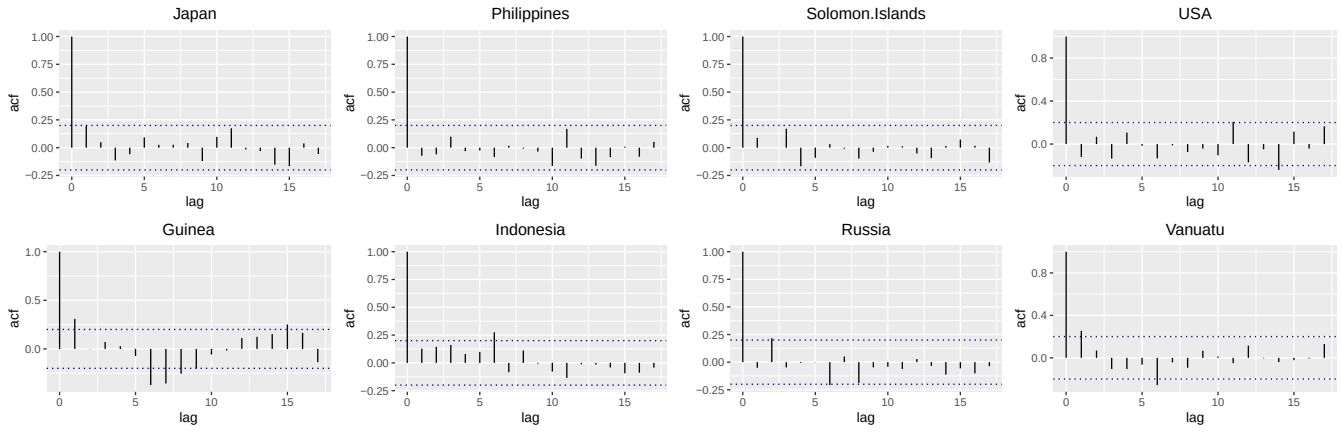


Figure 3: Autocorrelation of the timeseries.

We plot the distribution of the earthquakes by country in Figure 4. All are positively skewed with median lying on either 0 or 1 and mean taking value between 0.60 and 1.07.

```
h <- c()
i <- 1
for(column in colnames(agg))
{
  if (column != "year")
  {
    h[[i]] <- ggplot(agg, aes_string(column)) + geom_histogram(binwidth=1, fill='blue', alpha=0.5)
    i = i + 1
  }
}

grid.arrange(
  h[[1]], h[[2]], h[[3]], h[[4]], h[[5]], h[[6]], h[[7]], h[[8]],
  nrow=2,
  bottom = textGrob(
    "",
    gp = gpar(fontface=3, fontsize=9),
    hjust=1,
    x=1
  )
)
```

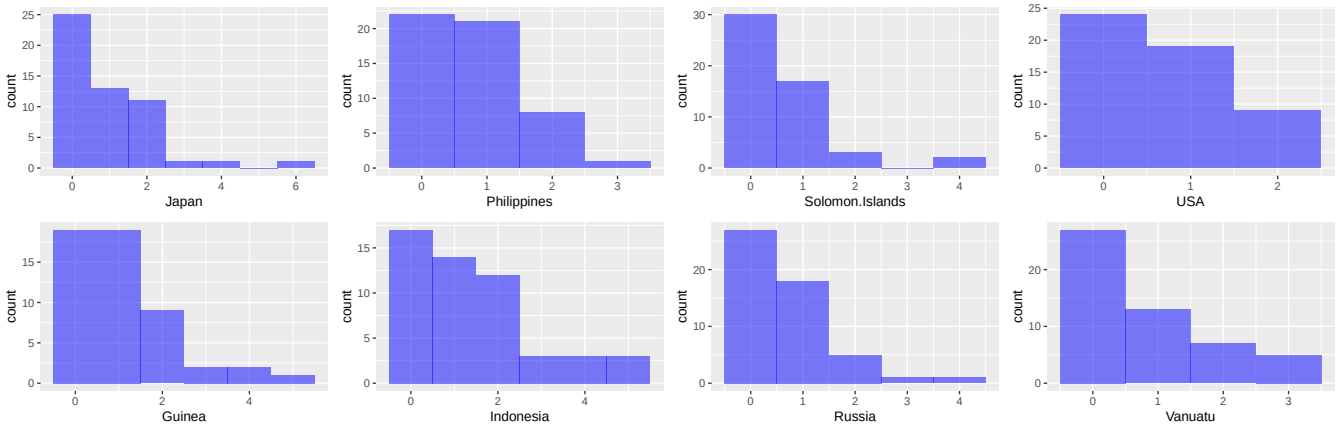


Figure 4: Histogram: number of earthquakes in a year

Finally, we run a goodness of fit test for each distribution using the Chi-squared statistics to check whether the data are not obviously inconsistent with the underlying distribution being Poisson (see Table 2). If the p value of the test is larger than 0.05, we can support the null hypothesis which states that the process is a Poisson process. Again, we acknowledge that for some countries, we fail to support the null hypothesis. There are other distributions that are known to fit the earthquakes temporal distribution better [Min-Hao Wua: Earthquake, Poisson and Weibull distributions], however, we carry on with the inference task assuming a Poisson process for each distribution.

```
pvalues <- c()
i=1
for(column in colnames(agg))
{
  if (column != "year")
  {
    gf = goodfit(agg[[column]], type="poisson", method="ML")
    gf.summary = capture.output(summary(gf))[[5]]
    pvalue = unlist(strsplit(gf.summary, split = " "))
    pvalue = as.numeric(pvalue[length(pvalue)])
    pvalues[[i]] <- c(column, pvalue)
```

Table 2: pvalues of Chi-square tests

country	p-value
Japan	0.02159118
Philippines	0.3905217
Solomon.Islands	0.007532658
USA	0.02987068
Guinea	0.4667118
Indonesia	0.04640559
Russia	0.6198651
Vanuatu	0.03627001

```
    i = i + 1
  }
}
pvalues <- as.data.frame(do.call(rbind, pvalues))
colnames(pvalues) <- c("country", "p-value")
kable(pvalues, caption='pvalues of Chi-square tests') %>% kable_styling(html_font=8)
```

4 Model

We assume that the temporal occurrence of earthquakes with magnitude greater than 7.0 is a Poisson process: i.e. independent, stationary and do not occur simultaneously. Therefore, we employ Poisson likelihood with a hierarchical structure, where the hierarchical grouping is done by country. Each country sits on a different junction tectonic plates and hence the earthquakes occurring in the same country should share the same distribution parameter (lambda), but across different countries, it is natural to assume that these lambdas are different. We specify Gamma prior on the intra-country lambdas ($\lambda[1],\ldots,\lambda[8]$), whose mean ($\mu$) comes from yet another Gamma distribution with the mean at the empirical mean of the observation and standard deviation from an exponential distribution. Note that μ is our prior on the inter-country mean of the rate as well.

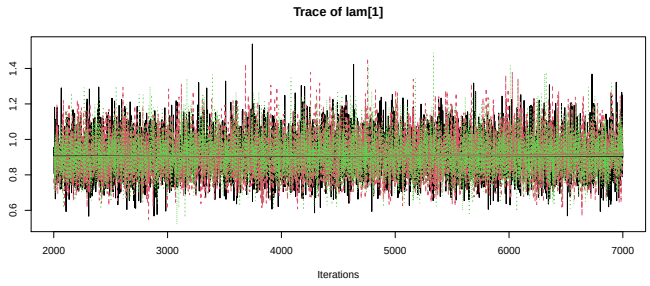
```
stats <- merge(x=as.data.frame(colMeans(agg)), y=as.data.frame(colVars(agg)), by=0) %>% filter(Row.names !=
colnames(stats) <- c("country","mean", "variance")
print(paste0("Inter-country empirical mean: ", mean(stats$mean)))
print(paste0("Inter-country empirical variance: ", var(stats$mean)))
```

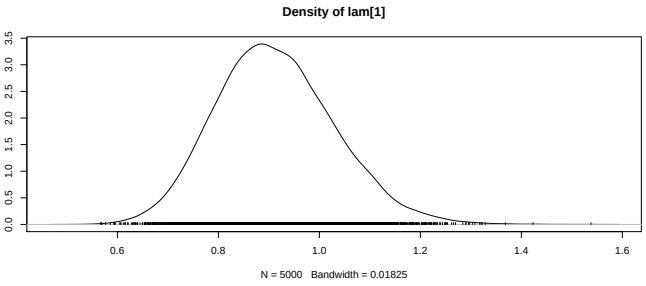
We fit the model using JAGS and R and generate three chains of simulation, but throw away the first 1000 steps as burn-in. We then produce 5000 more steps for each chain.

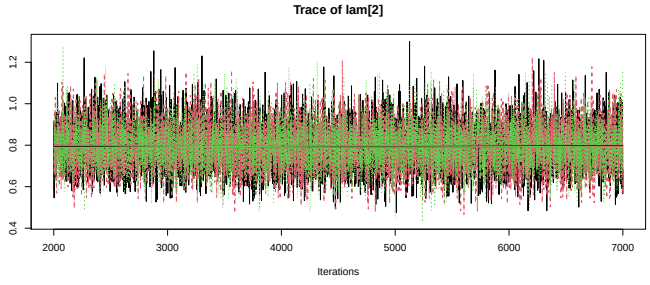
```
## Compiling model graph
##    Resolving undeclared variables
##    Allocating nodes
## Graph information:
##    Observed stochastic nodes: 416
##    Unobserved stochastic nodes: 10
##    Total graph size: 848
##
## Initializing model
```

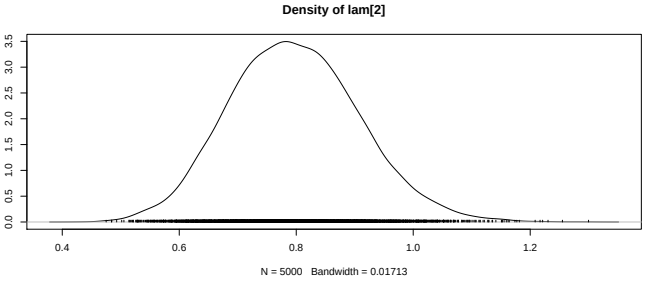
The trace plots of the parameters and Gelman-Rubin diagnostics indicate the convergence of the simulations (i.e. potential scale reduction factors all close to 1), and the effective sample size of all parameters are on the order of thousands, which guarantees us a reliable estimation of credible intervals.

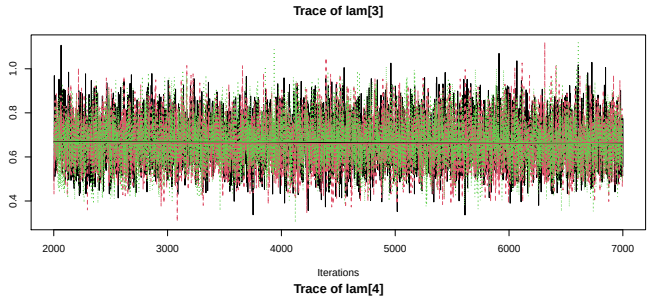
```
# convergence diagnostics
plot(mod_sim)
```

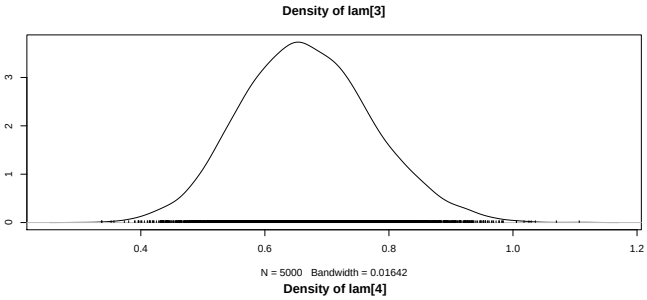

Trace of lam[1]
Iterations

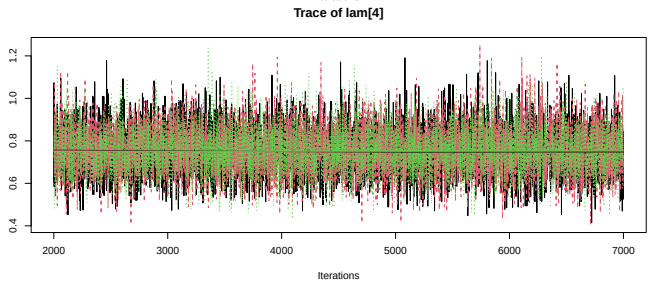

Density of lam[1]
N = 5000 Bandwidth = 0.01825

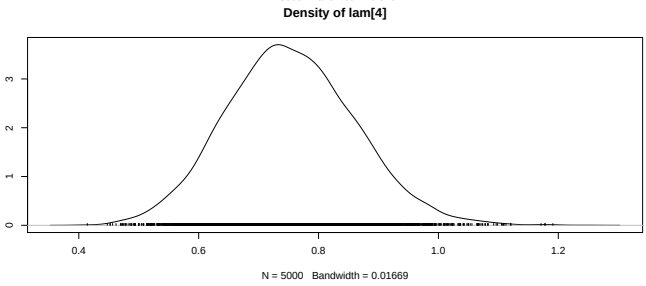

Trace of lam[2]
Iterations

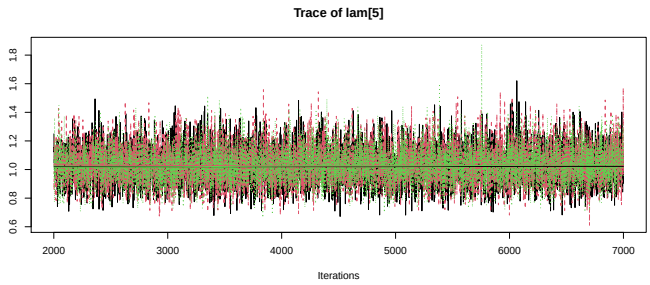

Density of lam[2]
N = 5000 Bandwidth = 0.01713

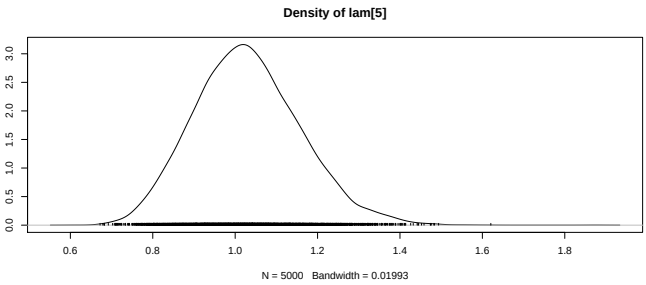

Trace of lam[3]
Iterations

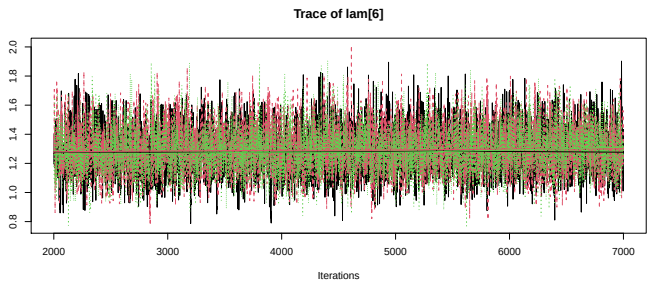

Density of lam[3]
N = 5000 Bandwidth = 0.01642

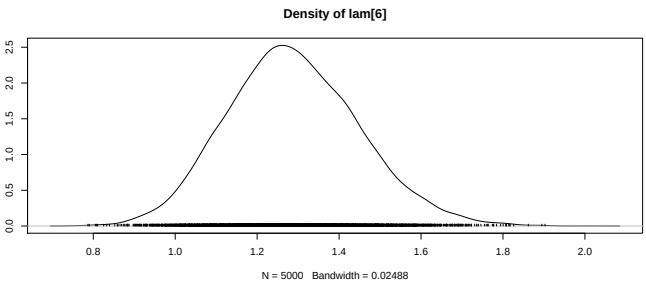

Trace of lam[4]
Iterations

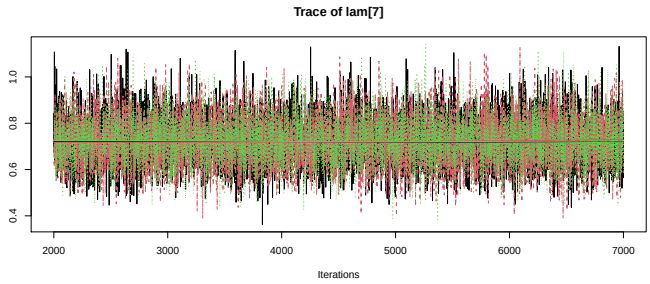

Density of lam[4]
N = 5000 Bandwidth = 0.01669

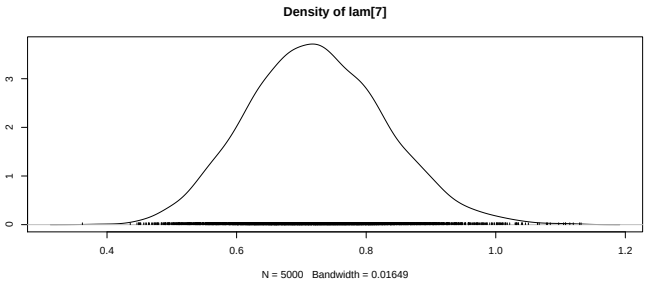

Trace of lam[5]
Iterations

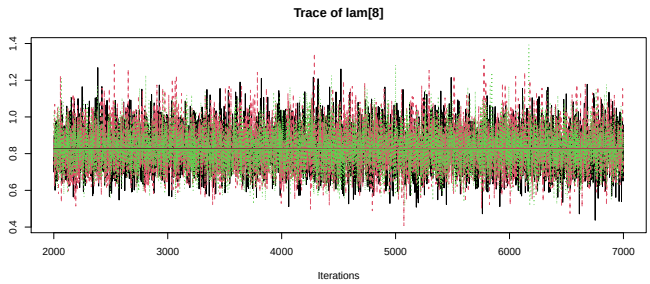

Density of lam[5]
N = 5000 Bandwidth = 0.01993

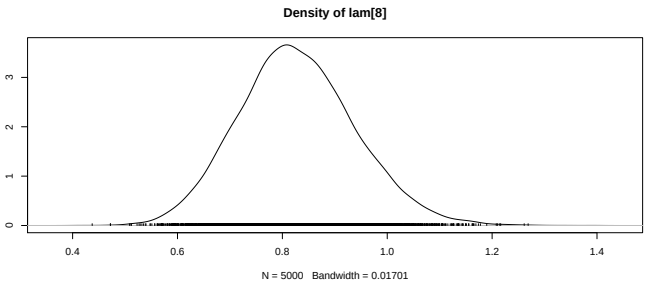

Trace of lam[6]
Iterations

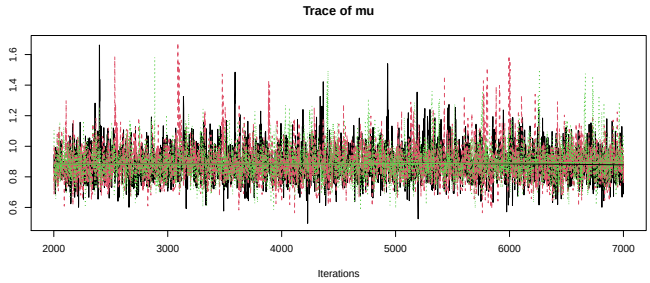

Density of lam[6]
N = 5000 Bandwidth = 0.02488

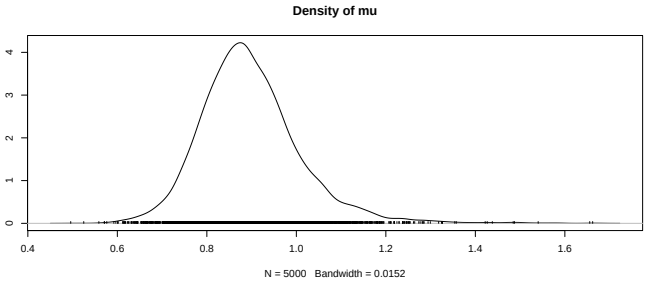


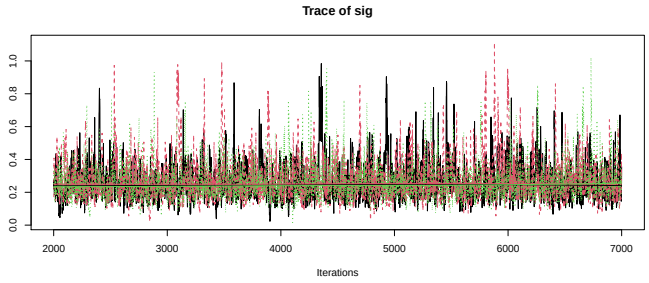


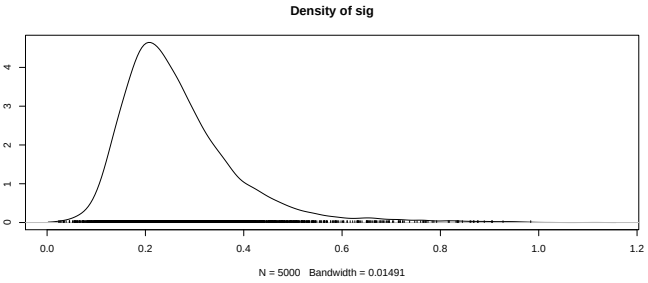












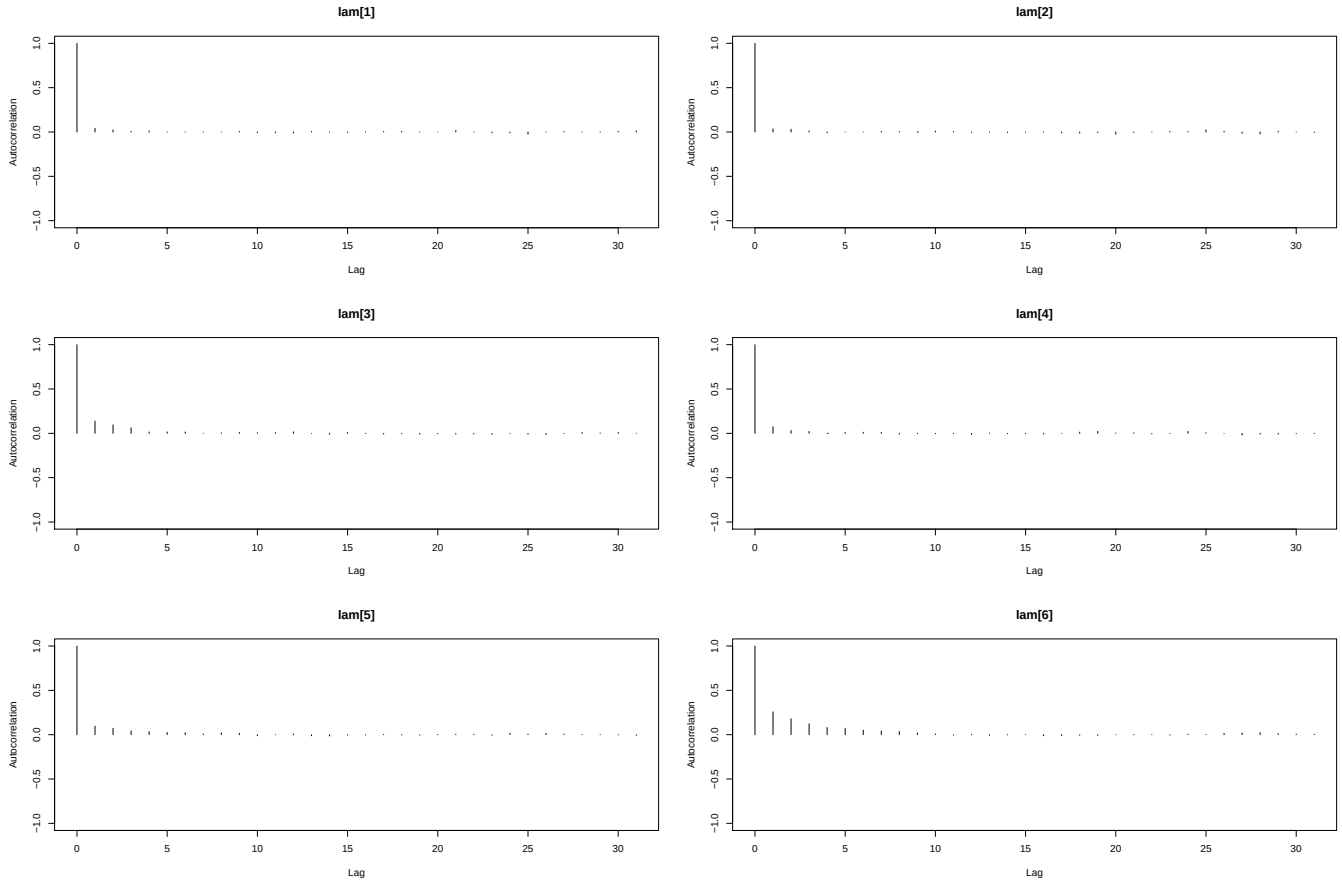

```
gelman.diag(mod_sim)
```

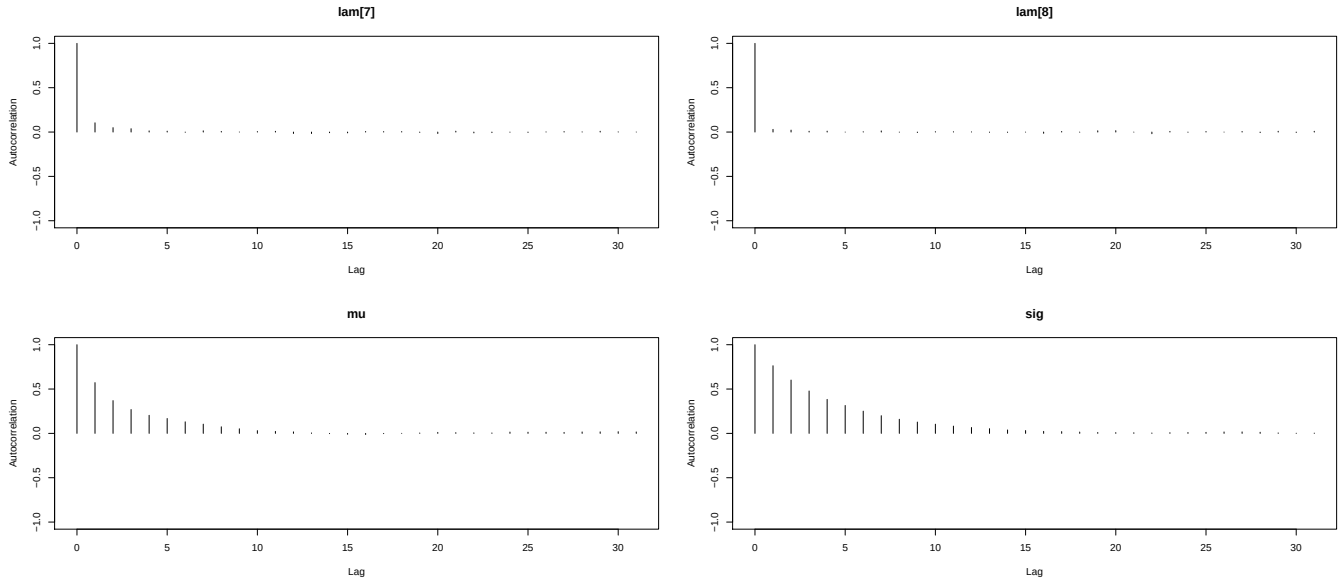
```
## Potential scale reduction factors:
##
##        Point est. Upper C.I.
## lam[1]          1          1
## lam[2]          1          1
## lam[3]          1          1
## lam[4]          1          1
## lam[5]          1          1
## lam[6]          1          1
## lam[7]          1          1
## lam[8]          1          1
## mu              1          1
## sig             1          1
##
## Multivariate psrf
##
## 1
```

```
autocorr.diag(mod_csim)
```

```
##              lam[1]       lam[2]       lam[3]       lam[4]       lam[5]
## Lag 0   1.000000000 1.0000000000  1.000000000  1.000000000  1.000000000
## Lag 1   0.042038775 0.0340269022  0.138754829  0.074462038  0.095471261
## Lag 5  -0.001351647 0.0007506054  0.016078093  0.011433759  0.025528427
## Lag 10 -0.009236064 0.0101377058  0.007347021 -0.004644956 -0.012275205
## Lag 50 -0.013071659 0.0018926528 -0.013390878  0.006460934 -0.008377382
##               lam[6]      lam[7]        lam[8]           mu         sig
## Lag 0   1.0000000000 1.000000000  1.0000000000  1.000000000 1.000000000
## Lag 1   0.2576272215 0.104382506  0.0304692719  0.572559276 0.762904985
## Lag 5   0.0725406502 0.010310221 -0.0003083674  0.168080509 0.314893247
## Lag 10  0.0078860786 0.005587956  0.0043513382  0.031079526 0.103949057
## Lag 50 -0.0004574178 0.004671540 -0.0011412207 -0.007573592 0.003417224
```

```
autocorr.plot(mod_csim)
```






```
effectiveSize(mod_csim)
```

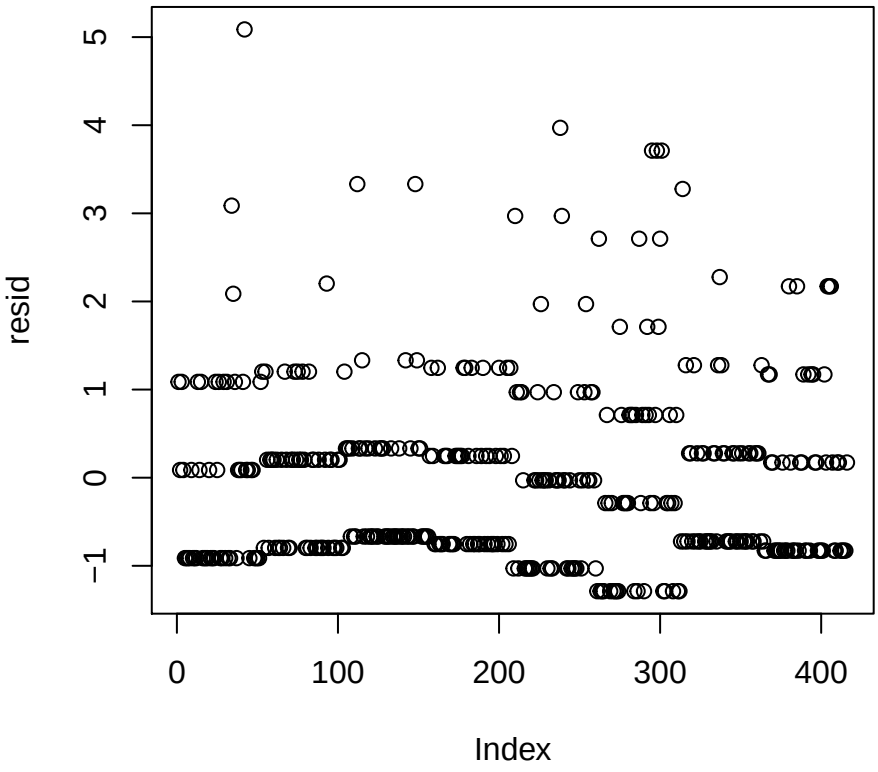
```
##    lam[1]    lam[2]    lam[3]    lam[4]    lam[5]    lam[6]    lam[7]    lam[8]
## 13188.607 13208.891  8978.222 11837.645  9047.838  5340.246 10604.753 13527.522
##        mu       sig
##  2882.869  1744.441
```

We check the fit via residuals. With a hierarchical model, there are two levels of residuals: the observation level and the country mean level. To simplify, we look at the residuals associated with the posterior means of the parameters.

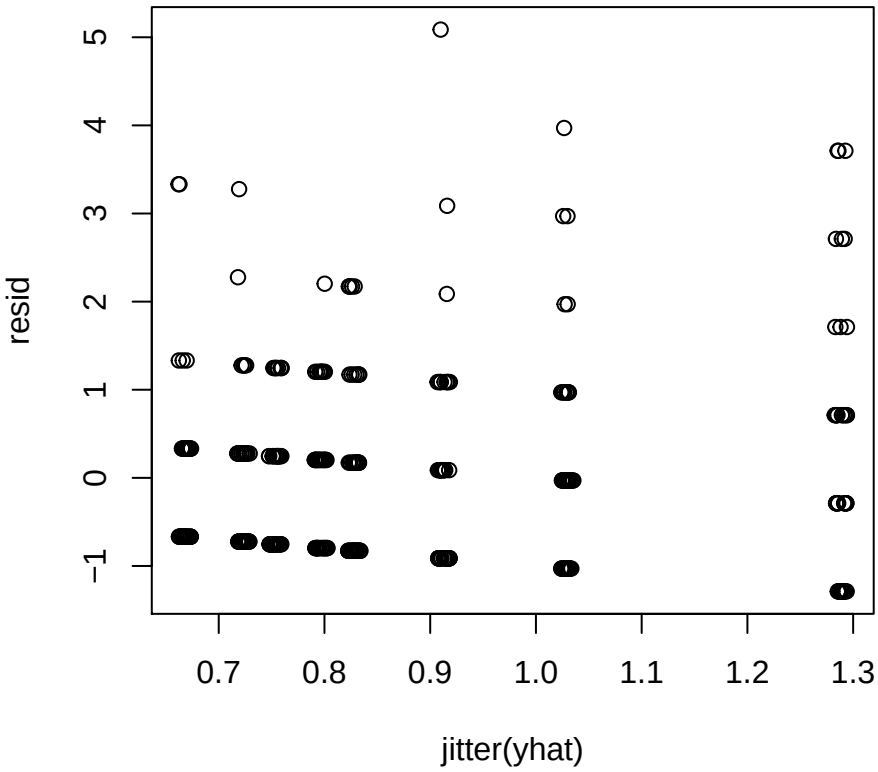
The observation residuals, based on the estimates of country means, seems to be right skewed indicating that the model struggles to fit to years when an unexpectedly large number of earthquakes occurred. For example, in 2011 in Japan, there were in total 6 earthquakes with magnitude greater 7. These earthquakes were thought to had been correlated (i.e. triggered by the first earthquake with magnitude 9 in the Tohoku area). Our model obviously fails to capture this accurately since the independence of events are assumed. The country mean level residual on the other hand look fine. Note that we omitted the plots for the limitation of pages but these were produced in the source code written in R.

```
pm_params = colMeans(mod_csim)
yhat = rep(pm_params[1:8], each=52)

# Observation level residuals
resid = agg_melt$value - yhat
plot(resid)
```




```
plot(jitter(yhat), resid)
```

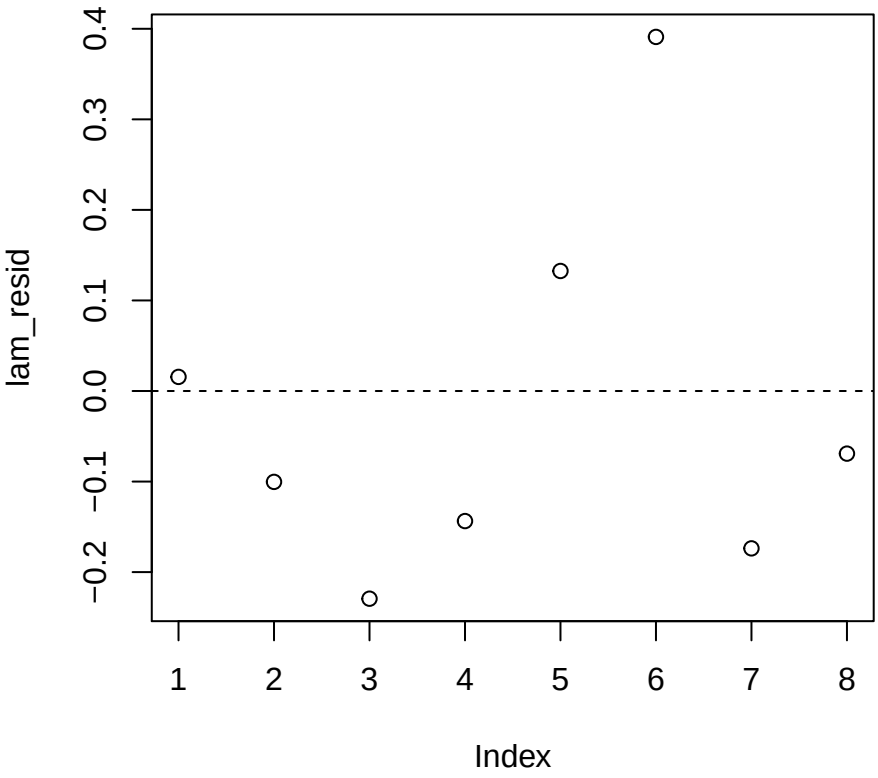



```
# Country level residuals
lam_resid = pm_params[1:8] - pm_params["mu"]
print(lam_resid)
```

```
##      lam[1]      lam[2]      lam[3]      lam[4]      lam[5]      lam[6]
##  0.01558189 -0.10036076 -0.22936049 -0.14362812  0.13254521  0.39102708
```

```
##      lam[7]      lam[8]
## -0.17379540 -0.06914549
```

```
plot(lam_resid)
abline(h=0, lty=2)
```



5 Results

We present the posterior summary in Table 3. The means of the parameter lambda in Poisson posterior distribution is the expected rate of occurrence. For example, $\lambda[1]$ is the expected mean rate for Japan and the model states that with probability 0.5981315 [`1-ppois(0,0.9116303)`], there will be at least one earthquake with a magnitude greater than 7.0 occurring in Japan in 2022. The same statement could be made for other countries by using the posterior distribution of their lambdas. We acknowledge that for some countries, the data are not statistically consistent with the Poisson process assumption and also from inspecting the residual plot of the fit, it is clear that the models are not always accurate.

```
kable(as.data.frame((summary(mod_sim)$statistics)), caption='Posterior distribution of the parameters.') %
```

6 Conclusions

In this report, we estimated the probability of large earthquakes occurring in a given country within the year 2022. We have assumed Poisson processes for earthquake temporal occurrence distributions and applied hierarchical models based on countries. We have obtained the posterior distributions for the expected mean of the rate for eight different countries. However, along the process, we have identified some shortcomings of the approach (i.e. the data not strictly following Poisson process, possible correlation between the observations).

Table 3: Posterior distribution of the parameters.

	Mean	SD	Naive SE	Time-series SE
lam[1]	0.9127631	0.1177954	0.0009618	0.0010312
lam[2]	0.7968205	0.1105618	0.0009027	0.0009563
lam[3]	0.6678207	0.1059896	0.0008654	0.0011002
lam[4]	0.7535531	0.1080043	0.0008819	0.0009801
lam[5]	1.0297264	0.1289274	0.0010527	0.0013112
lam[6]	1.2882083	0.1606158	0.0013114	0.0021373
lam[7]	0.7233858	0.1064469	0.0008691	0.0010477
lam[8]	0.8280357	0.1113417	0.0009091	0.0009692
mu	0.8971812	0.1133909	0.0009258	0.0021011
sig	0.2670889	0.1201078	0.0009807	0.0029137