BDA-Project: Ebola epidemic

2022-11-10

Introduction

For our Bayesian Data Analysis Project, we aim to model monthly number of infectious cases of the Ebola epidemic in various countries. As a simple model, the spread of the Ebola can be simulated by the SIS model shown by below equations:

$$\frac{dS}{dt} = -\beta SI + \gamma I \frac{dI}{dt} = \beta SI - \gamma I$$

Where, S is size of the susceptible population, I is size of the infected population. Also parameter β is the transmission rate of the disease and the parameter γ is the recovery rate. Considering that the size of the population is always constant and equal to 1, $1 = S(t) + I(t)$, we can solve the system above:

$$I(t) = \frac{I_\infty}{1 + Ve^{-\chi t}}$$

Where, $V = \frac{I_\infty}{I_0} - 1$, $\chi = \beta - \gamma$, and $I_\infty = (1 - \frac{\gamma}{\beta})$. Using the Taylor series, this can be expanded as below:

$$I(t) = I_0 + \frac{VI_0\chi}{V+1}\ t\ + \frac{I_0(V-1)V\chi^2}{2(V+1)^2}\ t^2\ + \dots$$

Taking only until the first order into account, we can write:

$$I(t) = I_0 + \frac{VI_0\chi}{V+1}\ t$$

But transmission rate depends on the number of susceptibles that an infected is on average in contact with. We assume this value has a linear relation with population density. We also assume that the initial number of infected people, I_0, is also depended on the population density. Then we can write our model as below:

$$I(t) = \mu + \beta_{pop\_den} \times \text{population density} + (\beta_{month\_diff} + \beta_{pop\_den,month\_diff} \times \text{population density}) \times t$$

As a result, we use the population density of the country, and number of months passed from the start of the pandemic as our predictors to model the number of Ebola cases. In the following section, more information on the Data can be found with explanatory illustrative figures.

Data

Data preprocessing

We collected the total number of Ebola cases from the Ebola dataset provided by WHO and shared by humandata.org (we use ebola_data_db_format.csv from WHO (2019)). The ebola cases we use in our project are an aggregate of suspected, confirmed and probable ebola cases, which are based on official information reported by ministries of health. The original dataset contains figures for ten countries:
Guinea, Italy, Liberia, Mali, Nigeria, Senegal, Sierra Leone, Spain, United Kingdom, United States. Each country spans it own date range of varying length between the end 2014 to the beginning of 2016, therefore we have different number of observations for each country. The case count is usually provided every 1 to 3 days (although the periodicity appear to vary by country), that we aggregate into monthly figures. We

did so by assigning the month value of 0 to the first date that appears in the dataset, and calculate the the other month values as the month difference from the first date. The first date in our original dataset was 2014-08-29 for Guinea, so all observations occuring 30 days after that have month difference value 0. The observations with date 2015-12-29 for the United States, on the other hand, has month_diff value 15 because the difference in time between this date and the first date in our dataset is 15 months. We opted for this approach, because we wanted the month_diff to represent the month count from the start of the outbreak. We decided to drop Italy and the United Kingdom from our analysis due to the fact that they had too many missing values. We also dropped observations having month = 0, since most of the remaining countries had observations starting from month 1. We aggregated the data for the remaining countries with the population density for each country at the given date. We use the population density from the WorldBank database. The population density can only access as a yearly figure. Therefore, using the previous example of the United states with month_diff value 15 calculated from date 2015-12-29, the pop_dens value will represent the population density of the United States in 2015.

Processed dataset

The final dataset contains 136 entries and four columns : Country, Cases, months_diff, and Pop_den. The variable months_diff is the number of months from the earliest data we could access cases information on, that is 2014-08-29 for Guinea. Country is a country among the 8 we selected. Cases is the aggregate of suspected, confirmed and probable ebola cases for the given country during the given months_diff, and pop_den is yearly population density that the country had during the given months_diff.

```
data <- read.csv(file = 'ebola_population_aggregate_same_span.csv')
head(data)
```

```
##   Country months_diff Cases  Pop_den
## 1  Guinea           1 11088 45.38080
## 2  Guinea           2 17013 45.38080
## 3  Guinea           3 21816 45.38080
## 4  Guinea           4 61876 46.36888
## 5  Guinea           5 61188 46.52489
## 6  Guinea           6 56884 46.52489
```

```
tail(data)
```

```
##           Country months_diff Cases  Pop_den
## 131 United States          12    84 35.06333
## 132 United States          13    68 35.06333
## 133 United States          14    52 35.06333
## 134 United States          15    44 35.06333
## 135 United States          16     4 35.06333
## 136 United States          18     4 35.31835
```

```
str(data)
```

```
## 'data.frame':    136 obs. of  4 variables:
##  $ Country    : chr  "Guinea" "Guinea" "Guinea" "Guinea" ...
##  $ months_diff: int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Cases      : num  11088 17013 21816 61876 61188 ...
##  $ Pop_den    : num  45.4 45.4 45.4 46.4 46.5 ...
```

Our final dataset contains 17 observations per country, one for each months_diff value (1,2,3,4,5,6,7,8,9, 10, 11, 12, 13, 14, 15,16 and 18 as month 17 was missing for every country in the original dataset, but the other observations are otherwise linear in time). The counts my country and month_diff are shown below.

```
## [1] "Observations by Country:"
```

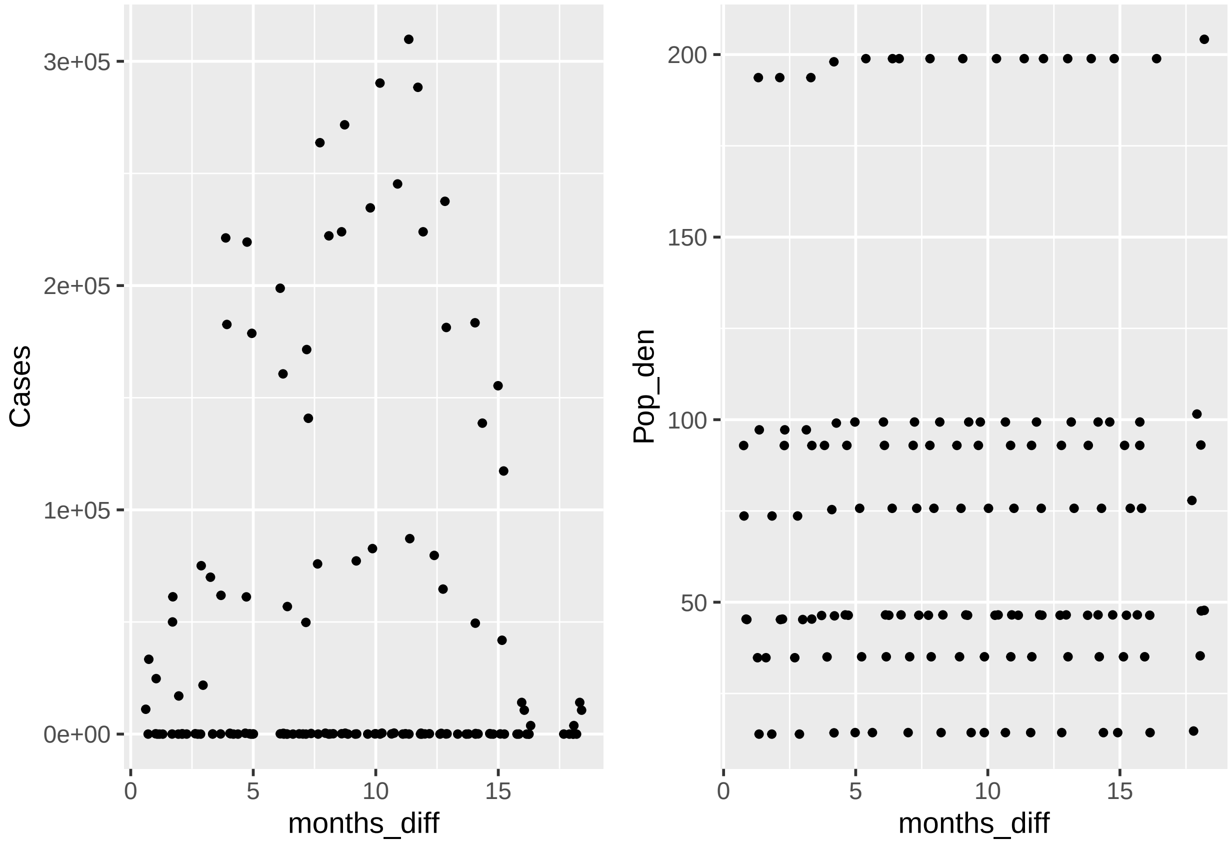
```
##
```

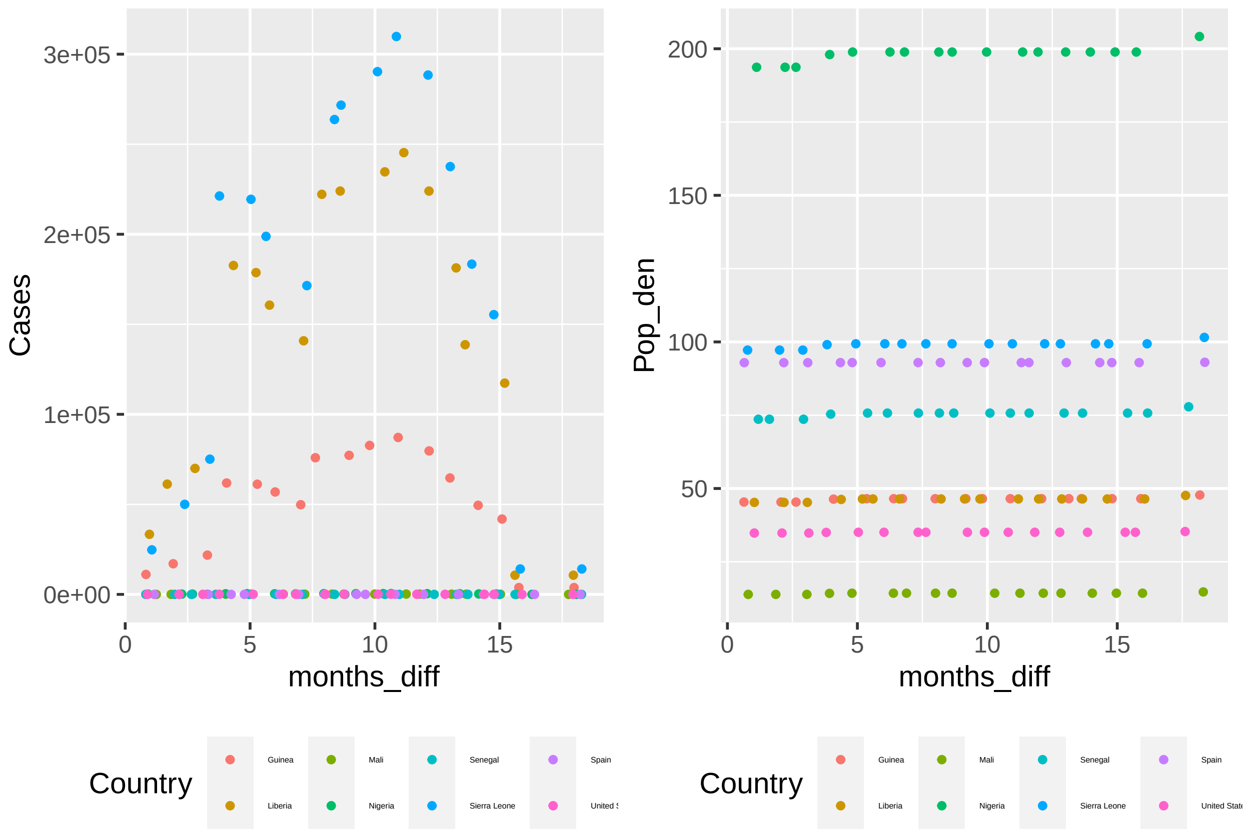
```
##        Guinea       Liberia          Mali       Nigeria       Senegal
##            17            17            17            17            17
##  Sierra Leone         Spain United States
##            17            17            17
```

```
## [1] "Observations by month_diff:"
```

```
##
##  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 18
##  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8
```

A more illustrative representation of the dataset can be seen from the scatter plots of below. In the top panels, we do not distinguish between countries. In the bottom panels, we differentiate between the observations of different countries by color. A small jitter is added to all graph to highlight overlapped points.





As can be observed from the panels above, there is obvious relationship between the country and the cases count by month. This will come to show that the pooled model will not accurately fit our data, because of this dependency of the country and the cases. On the other hand, a hierarchical model that allows for a degree of differentiation between countries, will result in a better fit, although a separate model might perform even better.

Models

The models we settled on are inspired by the study "A First Look at Multilevel Regression; or {Everybody's} Got Something to Hide Except Me and My Macaque" (Simpson (2022)). Similarly to their analysis on monkeys, our grouping variable is the country, which indicates to what group the observation belongs. We have one group-level covariate, which is the population density of the country. This does not vary at every observation, but changes by minimal amounts every year. Since the changes in density are very small for each country, we assume that population density is almost constant for each country and identify it as a group-level covariate. Lastly, our individual-level covariate is the month_diff variable, as it varies with the individual observations. We will run a pooled analysis and a hierarchical analysis on this dataset to evaluate which approach returns the best posterior predictive distributions. We run the models using cmdstan. For the MCMC attributes, we decided to use using the cmdstan defaults: that is, we use 4 chains, each having chain length 2000 and a burn-in of 1000. We agreed that four chains and chain length of 1000 after burn-in should be sufficient for the dataset we are using. The convergence analysis in the "Convergence analysis" section discusses the convergence of the parameters in the models more in detail, where we observed that both the pooled and the hierarchical models reach convergence with the given cmdstan settings.

Pooled

In the pooled approach, we assume that there is no difference between countries. We therefore consider the dataset of observations as a whole, and assume that in a linear regression approach the coefficient and parameters are shared among the countries. We argue that Cases observations j can be modeled as normal distribution whose mean depends on the month_diff and the population density. This is shown as reported below:

$$Cases_j \sim N(\mu + \beta_{pop\_den} pop\_den_j + \beta_{month\_diff} month\_diff_j + \beta_{pop\_den,month\_diff} pop\_den_j \times month\_diff_j, \sigma^2)$$

Following the pooled model approach, the parameter μ, β_{pop_den}, β_{month_diff}, $\beta_{pop\_den,month\_diff}$ and σ defining the normal distribution are drawn from the same prior distributions for every observation $Cases_j$.

We want weakly information or uninformative priors for this model, as long as they are proper priors. More information on prior choices will be given in the "Prior choices" section. We define these priors as:

$$\begin{aligned} \mu &\sim N(0, 100000) \\ \beta_{pop\_den} &\sim N(60, 10000) \\ \beta_{month\_diff} &\sim N(0, 10000) \\ \beta_{pop\_den,month\_diff} &\sim N(0, 10000) \\ \sigma &\sim gamma(100, 0.001) \end{aligned}$$

Where the second gamma parameter refers to the rate of the gamma distribution. The stan model is reported below:

```r
writeLines(readLines('pooled.stan'))
```

```
## data {
##   int<lower=0> N; // number of observations
##   vector[N] month; //vector of months_diff
##   vector[N] pop_den; //population density
##   vector[N] cases;
## }
##
## parameters {
##   real mu;
##   real beta_month;
##   real beta_pop;
##   real beta_month_pop;
##   real<lower=0> sigma;
## }
##
## model {
##
##   // Priors
##   mu ~ normal(0, 100000);
##   beta_month ~ normal(0, 10000);
##   beta_pop ~ normal(60, 10000);
##   beta_month_pop ~ normal(0,10000);
##   sigma ~ gamma(100, 0.001);
##
##
##   for (i in 1:N){
##     cases[i] ~ normal(mu + beta_month*month[i] + beta_pop*pop_den[i] + beta_month_pop*month[i]*pop_de
##   }
## }
##
## generated quantities {
```

```
##   #real ypred;
##   vector[N] ypred;
##   vector[N] log_lik;
##
##   #ypred = normal_rng(mu, sigma); // pooled predictive distribution
##
##   for(i in 1:N)
##     {
##       ypred[i] = normal_rng(mu + beta_month*month[i] + beta_pop*pop_den[i] + beta_month_pop*month[i]*
##       log_lik[i] = normal_lpdf(cases[i] | mu + beta_month*month[i] + beta_pop*pop_den[i] + beta_mont
##    }
##
## }
```

```r
stan_data_pooled <- list(
  cases = data$Cases,
  N = nrow(data),
  month=data$months_diff,
  pop_den=data$Pop_den
)


pooled <- cmdstan_model(stan_file = "pooled.stan")
model_pooled <- pooled$sample(data = stan_data_pooled, refresh=1000)
```

Hierarchical

In the hierarchical approach, we do not assume shared priors as in the pooled approach, neither completely distinguished group level priors as in the separate approach, but we differentiate between countries by assigning each country its own prior that's sampled from the same hyper-prior distribution for each country. This should allow some differentiation between countries while maintaining a level of dependency among all observations by sampling from the same hyper-priors. We can then say that we can model the number of cases for country i in month_diff j as reported below:

$$Cases_{ij}|\mu_j, \beta, \sigma \sim N(\mu_j + \beta_{pop\_den_j} pop\_den_{ij} + \beta_{month_j} month_{ij} + \beta_{month,pop\_den_j} month_{ij} \times pop\_den_{ij}, \sigma^2)$$

Priors:

$$\begin{aligned}
\sigma &\sim gamma(100, 0.001) \\
\mu_j &\sim N(\mu_h, \tau) \\
\beta_{pop\_den_j} &\sim N(\beta_{hpop\_den}, \tau_{hpop\_den}) \\
\beta_{month,pop\_den_j} &\sim N(\beta_{hmonth,pop\_den}, \tau_{hmonth,pop}) \\
\beta_{month_j} &\sim N(\beta_{hmonth}, \tau_{month})
\end{aligned}$$

Hyper-priors:

$$
\begin{aligned}
\mu_h &\sim normal(1000, 10000)\\
\tau &\sim gamma(10000, 1)\\
\beta_{hpop\_den} &\sim N(60, 10000)\\
\beta_{hmonth,pop\_den} &\sim N(0, 10000)\\
\beta_{hmonth} &\sim N(0, 10000)\\
\tau_{hpop\_den} &\sim gamma(10000, 1)\\
\tau_{hmonth,pop\_den} &\sim gamma(10000, 1)\\
\tau_{month} &\sim gamma(10000, 1)
\end{aligned}
$$

The stan code for the hierarchical model can be observed below.

```
writeLines(readLines("hier.stan"))
```

```
##
## data {
##   int<lower=0> M; //number of countries
##   int<lower=0> N;//max number of obs across all the countries
##
##   vector[M] month[N];//vector of months
##   vector[M] pop_den[N];//population density
##   vector[M] cases[N];
##
## }
##
## parameters {
##   vector[M] mu;
##   vector[M] beta_month;
##   vector[M] beta_pop;
##   vector[M] beta_month_pop;
##   real<lower=0>sigma;
##   real<lower=0> tau;
##   real hyper_mu;
##   real hyper_beta_month;
##   real hyper_beta_pop;
##   real hyper_beta_month_pop;
##   real <lower=0>hyper_tau_month;
##   real <lower=0>hyper_tau_pop;
##   real <lower=0>hyper_tau_month_pop;
## }
##
## model {
##
##   tau ~ gamma(10000,1);
##
##   sigma ~ gamma(100,0.001);
##
##   hyper_mu ~ normal(1000, 10000);
##   hyper_beta_pop ~ normal(60,10000);
##   hyper_beta_month ~ normal(0,10000);
##   hyper_beta_month_pop ~ normal(0,10000);
##   hyper_tau_pop ~ gamma(10000,1);
##   hyper_tau_month ~ gamma(1000,1);
##   hyper_tau_month_pop ~ gamma(1000,1);
##
##
##
##
##   for (j in 1:M){
##     mu[j] ~ normal(hyper_mu, tau);
##     beta_month[j] ~ normal(hyper_beta_month, hyper_tau_month);
##     beta_pop[j] ~ normal(hyper_beta_pop, hyper_tau_pop);
##     beta_month_pop[j] ~ normal(hyper_beta_month_pop, hyper_tau_month_pop);
##
##     for (n in 1:N){
##       cases[n,j] ~ normal(mu[j] + beta_month[j]*month[n,j] + beta_pop[j]*pop_den[n,j] + beta_month_po
##     }
##   }
## }
##
## generated quantities {
##   vector[M] ypred[N];
##   vector[M] log_lik[N];
##
##   for (j in 1:M){
##     for(i in 1:N){
##       log_lik[i,j] = normal_lpdf(cases[i,j] | mu[j] + beta_month[j]*month[i,j] + beta_pop[j]*pop_den
##       ypred[i,j] = normal_rng(mu[j] + beta_month[j]*month[i,j] + beta_pop[j]*pop_den[i,j] +
## beta_month_pop[j]*month[i,j]*pop_den[i,j], sigma);
##     }
##   }
## }
```

We feed the model the data in the form of 17x8 matrices, where the columns represent the countries and the rows the observations.

```r
months<- matrix(nrow = 8, ncol = 17)
pop_den <- matrix(nrow = 8, ncol = 17)
cases <- matrix(nrow = 8, ncol = 17)


row_count <- 1
for (i in 1:8){
  for (j in 1:17){
    months[i,j] = data$months_diff[row_count]
    pop_den[i,j] = data$Pop_den[row_count]
    cases[i,j] = data$Cases[row_count]
    row_count <- row_count + 1
  }
}
```

```r
stan_data_hier <- list(
  cases = t(cases),
  N = 17,
  M = 8,
  month=t(months),
  pop_den=t(pop_den)
)
```

```
hier <- cmdstan_model(stan_file = "hier.stan")
model_hier <- hier$sample(data = stan_data_hier, refresh=1000)
```

Prior choices

Pooled model

For the pooled model, we define the priors for μ, σ, β_{pop_den}, β_{month}, β_{month,pop_den}.

For μ we decided on $\mu \sim N(0, 100000)$, reasoning that the intercept of cases should be 0 and the standard deviation needs to be sufficiently large to be uninformative. The choice of 0 as the mean is grounded by the fact that the new ebola outbreak started in 2014, so we assume that any months following the ones in our dataset had no ebola cases. We assign a large standard deviation to account for the possibility of still having cases before the start of our timeline.

We chose $\sigma \sim gamma(100, 0.001)$, which yields a mean of 10^6 and a variance of 1^8. We reason that sigma should have a large mean and and a large variance, as well as being positive. Since the number of cases is in the order of 10^5, we require sigma to be significantly larger than this to be uninformative.

We chose $\beta_{pop\_den} \sim N(60, 10000)$. We center it at 60 since this the average population density in the entire world, and give it a standard deviation of 10000 to make it sufficiently uninformative. The choice of 60 lies in the idea of disease spreading in networks, where the upper bound of new diseases in a kilometer square is equivalent to the the population density. We chose $\beta_{month} \sim N(0, 10000)$ as a sufficiently uninformative prior, since it detains no assumptions about the sign or magnitude of the parameter. We also chose $\beta_{month,pop\_den} \sim N(0, 10000)$ arguing that we cannot make any assumption about the sign and the magnitude of this parameter.

Hierarchical model

For the hierarchical model, we chose our hyper priors to be uninformative and proper. For μ_h we decided on $\mu_h \sim N(100, 10000)$, giving us a sufficiently wide and uninformative distribution. We keep the shared sigma the same as in the pooled model, $\sigma \sim gamma(100, 1/1000)$. For the τ parameters we argue that it should be positive with a large variance and large enough mean. We assign it as $\tau \sim gamma(10000, 1)$, which yields a mean of 10000 and a variance of 10000. We apply the same reasoning made for our beta priors in the pooled model to our beta hyper priors: we chose $\beta_{hpop\_den} \sim N(60, 10000)$, $\beta_{hmonth} \sim N(0, 10000)$ and $\beta_{hmonth,pop\_den} \sim N(0, 10000)$. We also identify all hyper τs for the β parameters as $gamma(10000, 1)$ distributions.

Convergence analysis

In this section, we evaluate that our models have reached convergence by observing the value of $\hat{R}$. If this parameter is close to 1, we can safely assume that our MCMC has reached convergence. We also analyze the model performances by their ESS values. Normally, we would observe the Rhat values and ESS values for each parameter, and possibly evaluate chain plots to visually observe whether our chains have converged. Since our models are heavily parametrized, doing so for every individual parameter would be inefficient. Instead, we run the command cmdstan_diagnose() on our pooled and hierarchical models. This will tell us whether there were divergent transitions in the MCMCs, whether the $\hat{R}$ for each parameter is satisfactory (sufficiently close to 1) and whether the effective samples size ESS for each parameter is also satisfactory.

Pooled

```
model_pooled$cmdstan_diagnose()
```

```
## Processing csv files: /tmp/Rtmprwglth/pooled-202212021951-1-213ded.csv, /tmp/Rtmprwglth/pooled-20221
##
## Checking sampler transitions treedepth.
## Treedepth satisfactory for all transitions.
##
## Checking sampler transitions for divergences.
## No divergent transitions found.
##
## Checking E-BFMI - sampler transitions HMC potential energy.
## E-BFMI satisfactory.
##
## Effective sample size satisfactory.
##
## Split R-hat values satisfactory all parameters.
##
## Processing complete, no problems detected.
```

We observe that there were no divergent transitions found in the MCMC computation for the pooled model, and that both the ESS and $\hat{R}$ values are satisfactory. We can assume that our MCMC for the pooled model have reached convergence and that the model parametrization is adequate.

Hierarchical

```
model_hier$cmdstan_diagnose()
```

```
## Processing csv files: /tmp/Rtmprwglth/hier-202212021951-1-3e35c5.csv, /tmp/Rtmprwglth/hier-2022120219
##
## Checking sampler transitions treedepth.
## Treedepth satisfactory for all transitions.
##
## Checking sampler transitions for divergences.
## No divergent transitions found.
##
## Checking E-BFMI - sampler transitions HMC potential energy.
## E-BFMI satisfactory.
##
## Effective sample size satisfactory.
##
## Split R-hat values satisfactory all parameters.
##
## Processing complete, no problems detected.
```

As can be observed, with the current parametrization, the MCMCs for our hierarchical models had no divergent transitions, and the $\hat{R}$ and SSE metrics for our parameter are satisfactory. We may therefore assume that our chains have converged and the model is satisfactory. Initially we had a hyperprior $\mu_h \sim gamma(1, 0.01)$, which resulted into almost all transition diverging and three fourth of our parameters to fail convergence according to the $\hat{R}$ diagnostic. The reason for this was probably related to the fact that this prior was too narrow and informative. After changing it to a more wide prior $\mu_h \sim N(1000, 10000)$ the MCMCs performed significantly better.

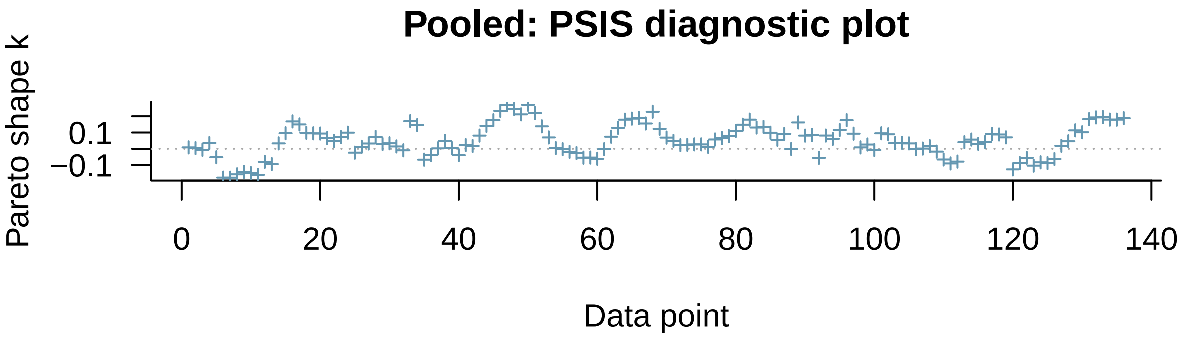
Model comaparison

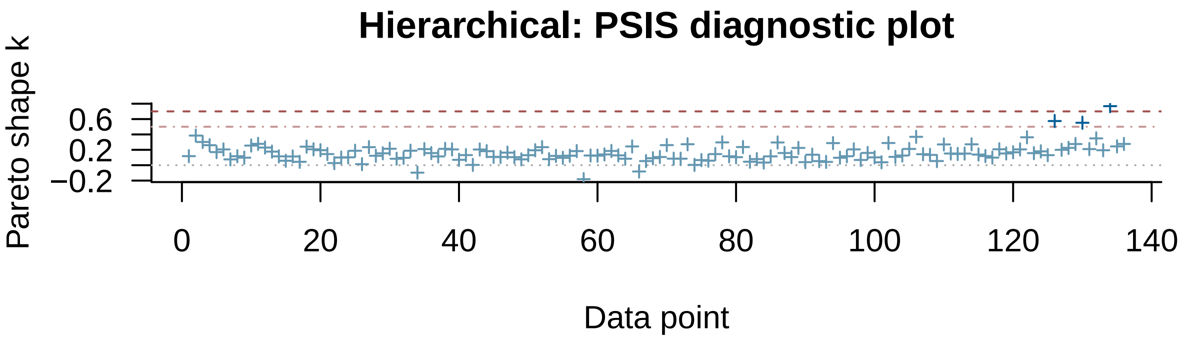
We use the loo function to observe the k-values of our models and compare them based on their elpd difference. We select the best model based on the elpd comparison results and analyze the reliability of each model by observing the proportion of k-values < 0.7.

```
## This is loo version 2.5.1
```

```
## - Online documentation and vignettes at mc-stan.org/loo
## - As of v2.0.0 loo defaults to 1 core but we recommend using as many as possible. Use the 'cores' ar

## Warning: Dropping 'draws_df' class as required metadata was removed.

## Warning: Relative effective sample sizes ('r_eff' argument) not specified.
## For models fit with MCMC, the reported PSIS effective sample sizes and
## MCSE estimates will be over-optimistic.

## Warning: Some Pareto k diagnostic values are too high. See help('pareto-k-diagnostic') for details.

## Warning: Dropping 'draws_df' class as required metadata was removed.

## Warning: Relative effective sample sizes ('r_eff' argument) not specified.
## For models fit with MCMC, the reported PSIS effective sample sizes and
## MCSE estimates will be over-optimistic.
```

We first observe the k-diagnostics for each model. These are presented in the two below and summarized in the following tables.





Pareto-k diagnostics for the pooled model:

```
pareto_k_table(loo_pol)
```

```
##
## All Pareto k estimates are good (k < 0.5).
```

Pareto-k diagnostics for the hierarchical model:

```
pareto_k_table(loo_hier)
```

```
## Pareto k diagnostic values:
##                          Count Pct.    Min. n_eff
## (-Inf, 0.5]   (good)     133   97.8%   721
##  (0.5, 0.7]   (ok)         2    1.5%   254
##    (0.7, 1]   (bad)        1    0.7%   61
##    (1, Inf)   (very bad)   0    0.0%   <NA>
```

The pareto k diagnostistics for both the pooled and the hierarchical model are very good. The pooled model appears to perform slightly better on the k-diagnostics, but more accurate model comparison is achieved by comparing the elpds of the models. We use loo_compare for this purpose.

```
com <- loo_compare(loo_pol, loo_hier)
com
```

```
##        elpd_diff se_diff
## model2   0.0       0.0
## model1 -58.6      15.3
```

Model 2, which is the hierarchical model, appears to perform better on the prediction accuracy than the pooled model. We therefore conclude that with our current parametrizations for the pooled and the hierarchical model, the hierarchical model appears to be the most reliable model.

Posterior predictive checks

After fitting our model, we perform posterior predictive checks by understanding how the actual number of ebola cases compares with simulated draws from the model. If the model is a good fit, we would expect the observed data to look similar to the predicted draws.

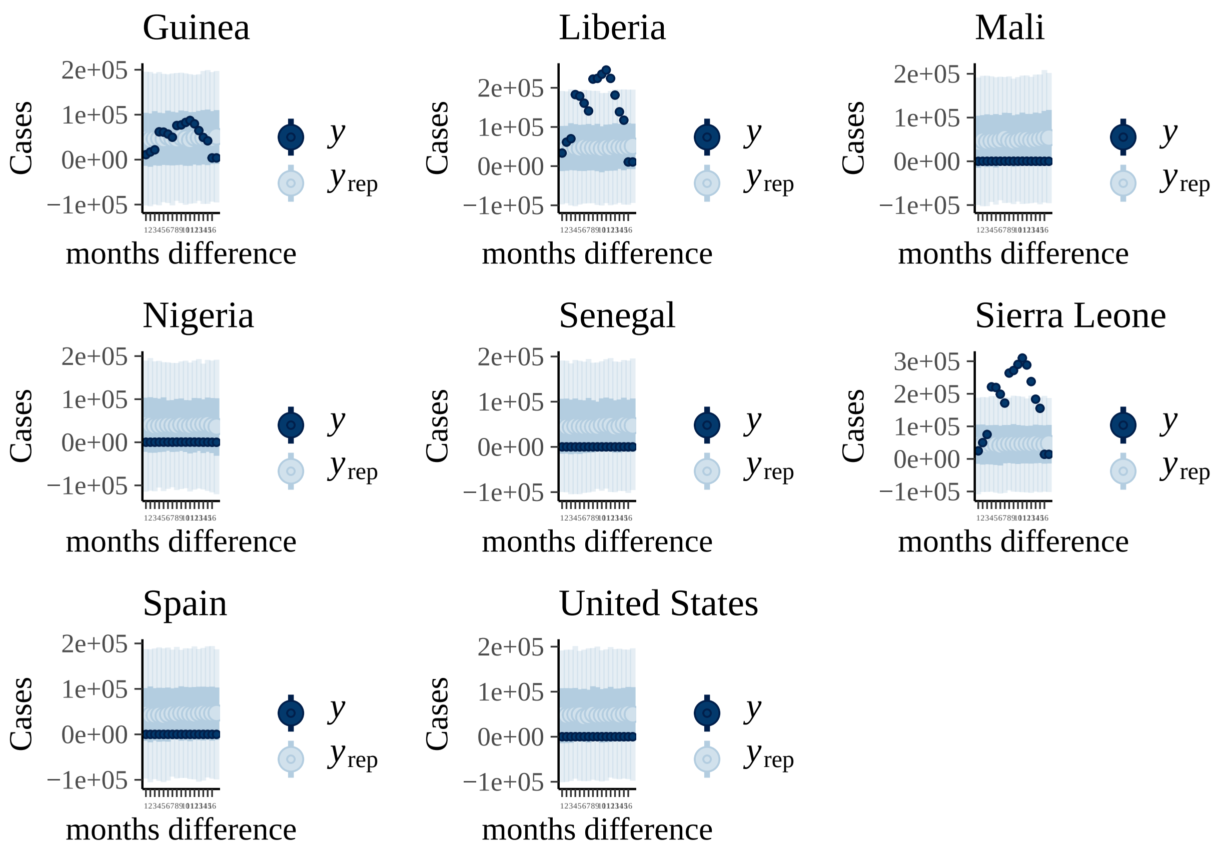
We will make this comparison by using the same values of the predictors – months difference and population density – which we denote by X, to obtain the simulated draws. In this case, the predictive distribution is given by

$$p(\hat{y}|y, X) \sim \int p(\hat{y}|\theta, X)p(\theta|y, X)d\theta.$$

Graphically, this can be done using the R package Bayesplot.

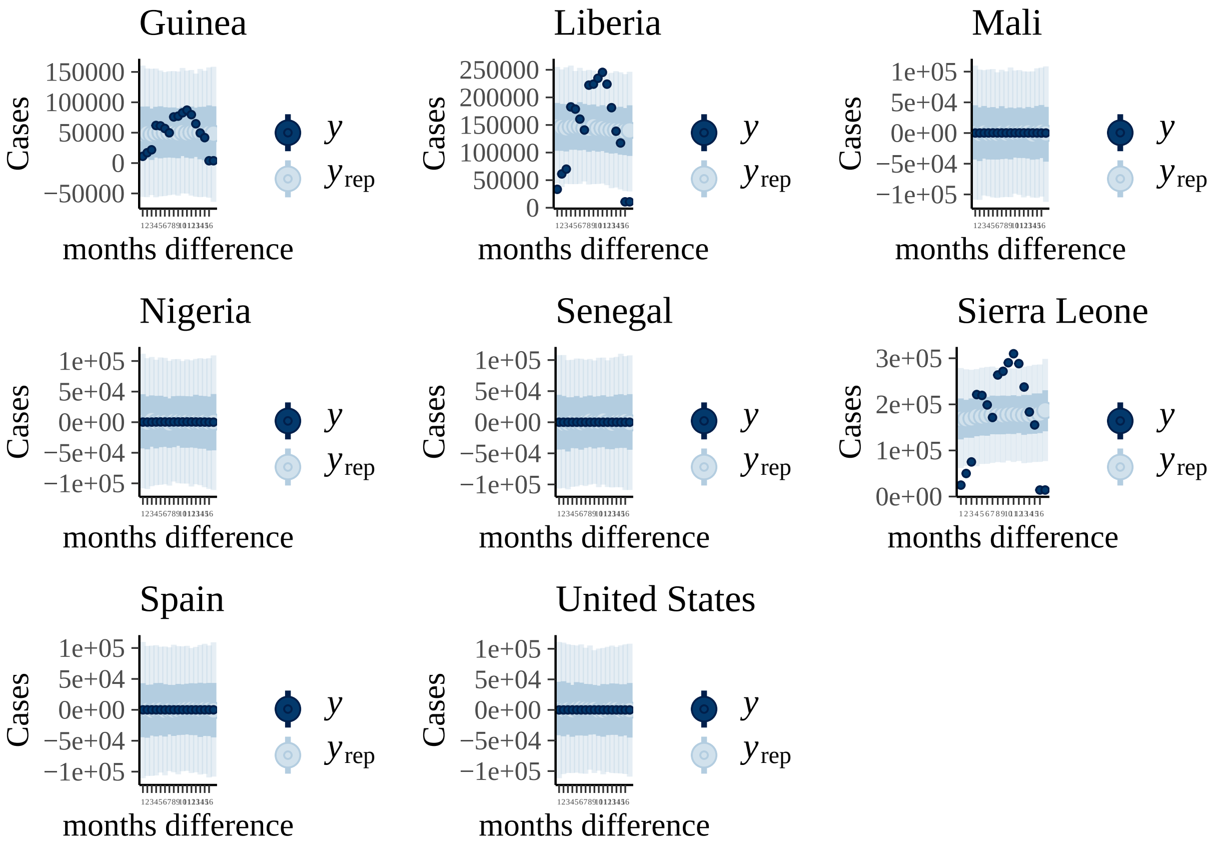
Pooled:

```
## This is bayesplot version 1.9.0

## - Online documentation and vignettes at mc-stan.org/bayesplot

## - bayesplot theme set to bayesplot::theme_default()

##    * Does _not_ affect other ggplot2 plots

##    * See ?bayesplot_theme_set for details on theme setting
```



From the plots for Mali, Nigeria, Senegal, Spain and the United States, we see that the actual observations look compatible with the simulated draws. This is also arguably true for Guinea, since its observations fall within the bodies of the boxplots. However, the model does not seem to fit well with observations from Liberia and Sierra Leone. Hence, the pooled model may not be the most suitable when we take all the countries data into account.

Hierarchical:



From the plots for Mali, Nigeria, Senegal, Spain and the United States, we see that the actual observations look compatible with the simulated draws. By inspection, the fit looks better for the hierarchical model as compared to the pooled model when considering the aforementioned countries. However, as with the pooled model, the model does not seem to fit well with observations from Liberia and Sierra Leone. This might suggest that the model specification is not precise enough.

Prior sensitivity analysis

The hierarchical model:

The hierarchical model is a better model for analyzing our data (Model Selection section). To check the sensitivity of this model to the choice of priors, we try different priors for this model. Because we have so many hyperparameters, we evaluate the effect of the prior distribution only on the posterior distributions of the directly affected parameters. The first section of priors is as below, keeping all else equal (colored red in the graphs):

$$\begin{aligned} hyper\_tau\_pop &\sim gamma(10000, 10) \\ hyper\_tau\_month &\sim gamma(1000, 10) \\ hyper\_tau\_month\_pop &\sim gamma(1000, 10) \end{aligned}$$

Which we also change to the following set (colored blue in the graph), keeping all else equal:

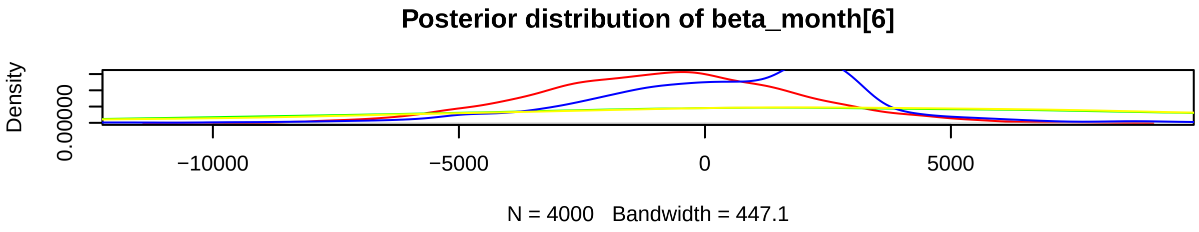
$$\begin{aligned} hyper\_tau\_pop &\sim N(10000, 10000) \\ hyper\_tau\_month &\sim N(10000, 10000) \\ hyper\_tau\_month\_pop &\sim N(10000, 10000) \end{aligned}$$

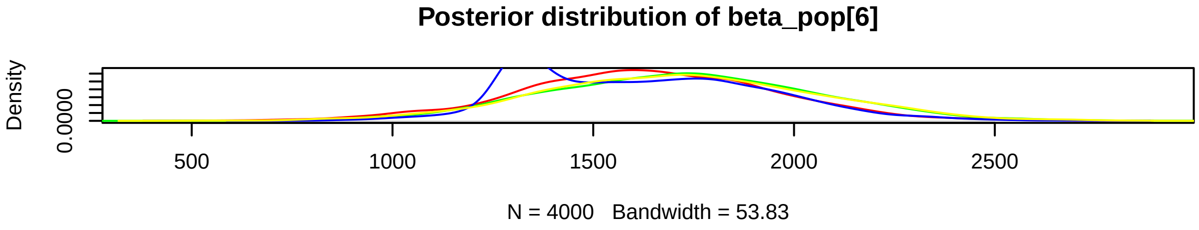
And lastly we change the hyper betas instead of the taus (colored yellow in the graph),keeping all else equal, as follows:

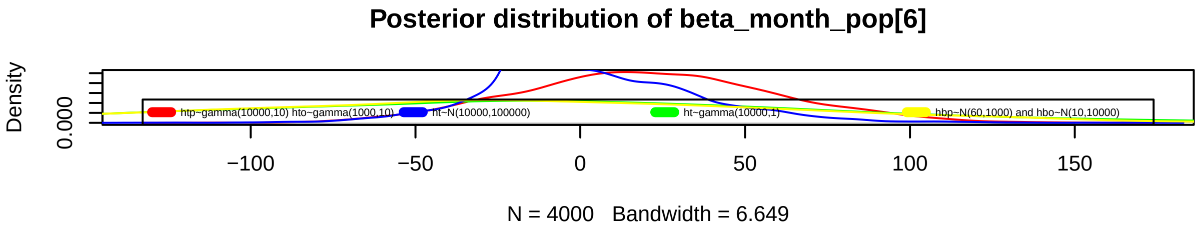
$$hyper\_beta\_pop \sim N(60, 1000)$$
$$hyper\_beta\_month \sim N(10, 10000)$$
$$hyper\_beta\_month\_pop \sim N(10, 10000)$$

We compare these two set of changes with the beta posteriors achieved whit the hyperprior parameters used in the model section (colored green in the graph). We select country 6 to evaluate the difference on the posterior distributions of our βs between these new hyperpriors and the hyperpriors chosen in the model section. Running the model with these parameters:

```
stan_data_hier <- list(
  cases = t(cases),
  N = 17,
  M = 8,
  month=t(months),
  pop_den=t(pop_den)
)
hier_1 <- cmdstan_model(stan_file = "hier_sen1.stan")
model_hier_1 <- hier_1$sample(data = stan_data_hier, refresh=1000)
hier_3 <- cmdstan_model(stan_file = "hier_sen3.stan")
model_hier_3 <- hier_3$sample(data = stan_data_hier, refresh=1000)
hier_4 <- cmdstan_model(stan_file = "hier_sen4.stan")
model_hier_4 <- hier_4$sample(data = stan_data_hier, refresh=1000)
```







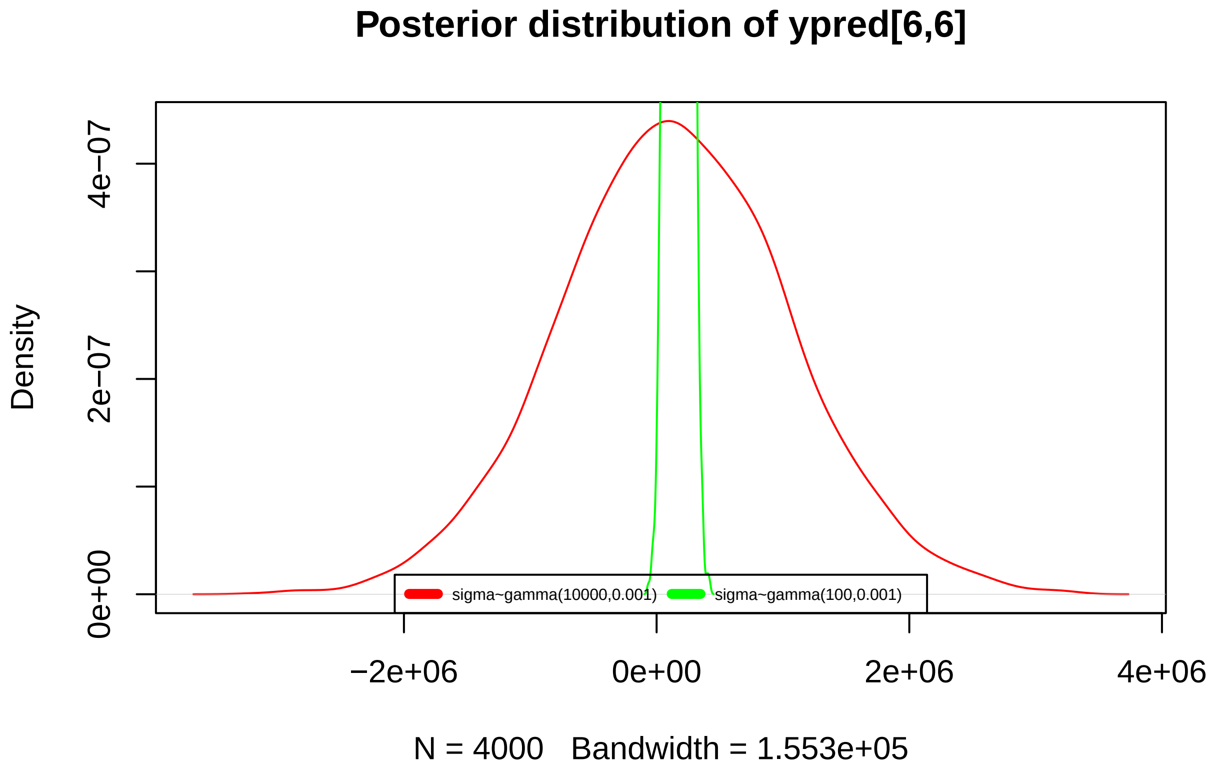
In the graph, htp stands for hyper tau pop, and hto stands for hyper tau others (when the hyper distributions are not all the same). Observing the graphs, we note that the posteriors drawn from the hyper beta taus $\tau \sim (10000, 1)$ we use in our model (shown in yellow, where we change only the hyper-betas, and the green model which is the original model) are narrower compared to the posteriors drawn from chaining the hyper tau to $\tau \sim (10000, 1)$. Changing the hyper betas (model shown in yellow) doesn't seem to affect the outcome when we keep our original hyper tau. We conclude that our choice of hyper tau $\tau \sim (10000, 1)$ results in a much more uninformative posterior for the beta_month and the beta_month_pop, possibly reflecting that 1) our choice of hyper taus are centered at a too high mean 2) we do not have sufficient observations for the likelihood to balance the effect of the prior 3) we cannot infer any information about the sign and magnitude of the effect of the month and the interaction of the month and the population density on the number of cases.

In the second test, we change below priors:

$$\sigma \sim gamma(1000, 0.001)$$

We evaluate the change on the ypred distribution for 6th month observation of the 6th country, since the predictive distribution is directly affected by the sigma parameter. Running the model with these parameters:

```r
hier_2 <- cmdstan_model(stan_file = "hier_sen2.stan")
model_hier_2 <- hier_2$sample(data = stan_data_hier, refresh=1000)
```



Based on the posterior plots, we observe that our results are highly sensitive to the specified standard deviations parameterising the prior distributions. As illustrated above, our chosen sigma induces a much narrower posterior distribution over this specific ypred as compared to the new sigma prior. This indicates that our results are heavily dependent on our choice of priors, making them non-robust. This sensitivity analysis reveals that our current modelling set-up may not be suitable for adequately describing our observations.

The pooled model:

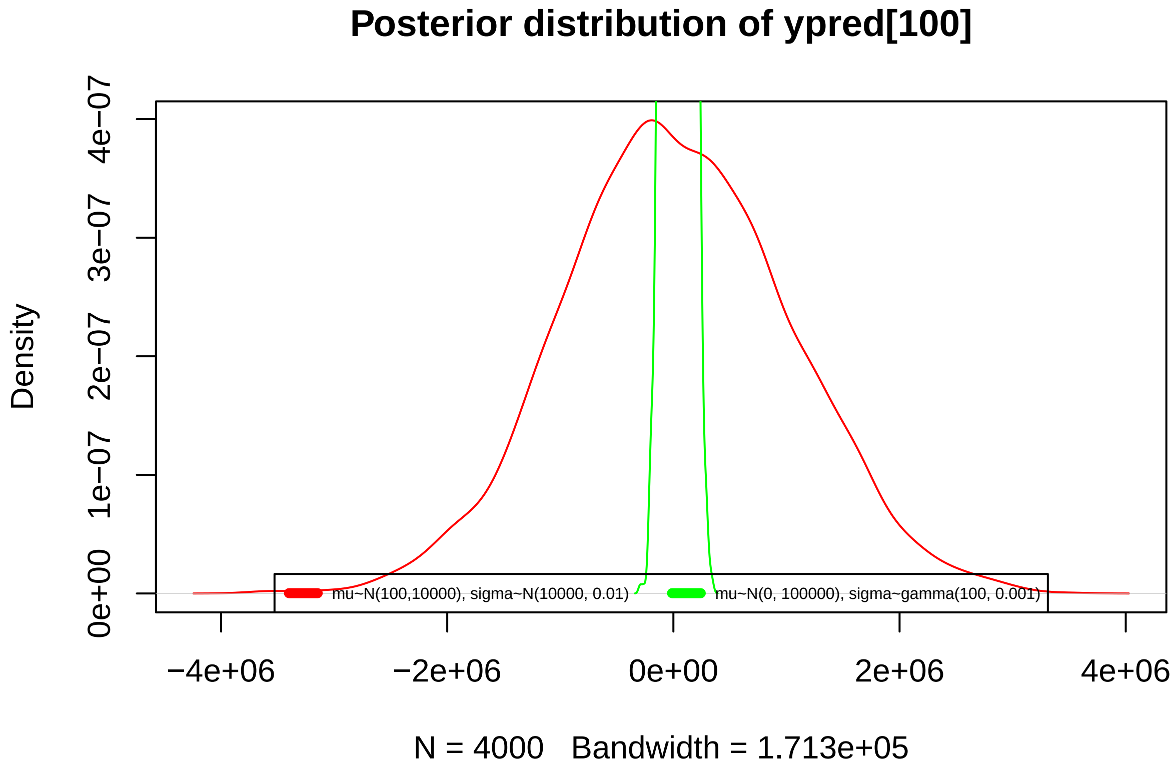
We can also test the prior sensitivity for the pooled model. As the first test, we change below priors (red) and evaluate the change on the posteriors from the priors used in our model (green). Again, we observe the differences in ypred[100], that is the predictive distribution for the datapoint at index 100.

$$\mu \sim N(100, 10000)$$
$$\sigma \sim N(10000, 0.01)$$

```r
pooled_1 <- cmdstan_model(stan_file = "pooled_sen1.stan")
```

```
## Warning in readLines(stan_file): incomplete final line found on
## 'pooled_sen1.stan'
```

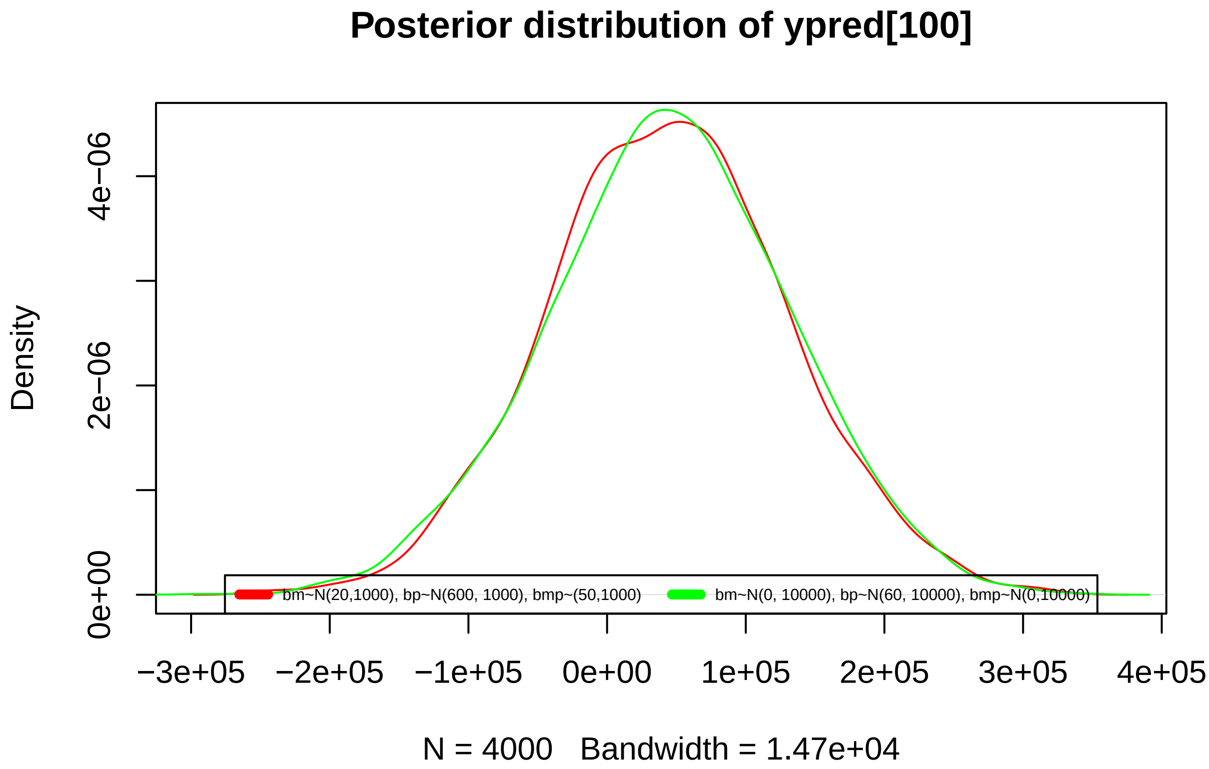
```r
model_pooled_1 <- pooled_1$sample(data = stan_data_pooled, refresh=1000)
model_pol_1_draws <- model_pooled_1$draws(format='df')
```



And as a second test, we change the βs priors as shown below and evaluate their effect on the posterior distributions of ypred[100].

$$beta\_month \sim N(20, 1000)$$
$$beta\_pop \sim N(600, 1000)$$
$$beta\_month\_pop \sim N(50, 1000)$$

```r
pooled_2 <- cmdstan_model(stan_file = "pooled_sen2.stan")
model_pooled_2 <- pooled_2$sample(data = stan_data_pooled, refresh=1000)
model_pol_2_draws <- model_pooled_2$draws(format='df')
```



We draw the same conclusions made for the hierarchical model. The changes in the βs do not appear to affect our chosen predictive posterior distribution significantly, meanwhile the change in the sigma parameter is very apparent. Our chosen sigma induces a much narrower posterior distribution over this specific ypred as compared to the new sigma prior.This sensitivity analysis reveals that our current modelling set-up may not be suitable for adequately describing our observations.

Discussion

In this project, we have assumed a linear relationship between the number of ebola cases and the predictors considered. However, following our sensitivity analyses, it is observed that our results are highly sensitive to our prior specifications, suggesting that the results are not robust to the choice of priors used. As we saw in the introduction, the size of infected population has a non linear relationship with time, which we approximated to be linear. This possibly indicates that this assumption is too strong, and that a linear model may not be the most appropriate for our modelling set-up. Some possible follow up steps include: considering higher order terms to better approximate the the number of Ebola cases, and having the relationship modeled in a non-linear fashion, such as via kernel functions.

The size of our dataset must also be considered. After preprocessing, we are left only with 8 countries with 17 observations each. A follow up step could be collecting more observations and possibly considering more predictors, as it is possible that our posterior distributions result uninformative due to the low number of observations.

We assume there is no correlation between parameters, which is potentially a strong assumption to make. In the future, we could consider encoding this information into the model, for instance in the form of a covariance matrix between the parameters.

Lastly, in our model comparison, the hierarchical model resulted to be more reliable than the pooled model. Nevertheless, it is possible that a separate model might have also been a reliable model if the dataset included more observations per country. We should consider it for further analysis.

Conclusion

In summary, we tried to model the number of Ebola cases per month for different countries. To do so, we assumed a linear relationship between number of the cases in each country and the time passed since the start of the pandemic. We also assumed that the transmission per time rate of the disease has a linear relation with the population density in the country. We used two different models: a "Hierarchical model" and a "Pooled model". The dataset that we used for our models consists of 8 countries and 17 observations in each country, where each observation correspond to a different month from the beginning of the epidemic. After training our models, we checked their convergence, and evaluated that both the hierarchical and the pooled models converge. However, after comparing the elpd values and the pareto-k optimality parameters of the two models, we found the hierarchical model a better choice for our dataset. In our posterior predictive analysis, we compared the actual number of Ebola cases with the simulated draws from our models. Both the hierarchical and pooled models return appropriate simulations when compared with the empirical data for most of the countries. As the final step, we checked our model's sensitivity to the choice of priors. Although our model is simple, it gives us reasonable predicts. However as a future work, we can modify it by for instance adding higher order relations so that it give us a more accurate prediction.

Self-reflection

While doing this project, we had the opportunity to delve deeper into understanding Bayesian Data Analysis applied to a real-world scenario. We learnt how to engage in the Bayesian Workflow: from fitting candidate models, performing posterior predictive checks, comparing model results, to doing prior sensitivity analyses. One of our biggest takeaways was learning the probabilistic programming language Stan, which we have never been exposed to before. We also learned our to conduct predictive posterior checkings more in detail than what was explored in the assignments.

References

Simpson, Dan. 2022. "A First Look at Multilevel Regression; or Everybody's Got Something to Hide Except Me and My Macaques." September 6, 2022. https://dansblog.netlify.app/2022-09-04-everybodys-got-something-to-hide-except-me-and-my-monkey.html.

WHO. 2019. "Number of Ebola Cases and Deaths in Affected Countries." November 10, 2019. https://data.humdata.org/dataset/ebola-cases-2014.