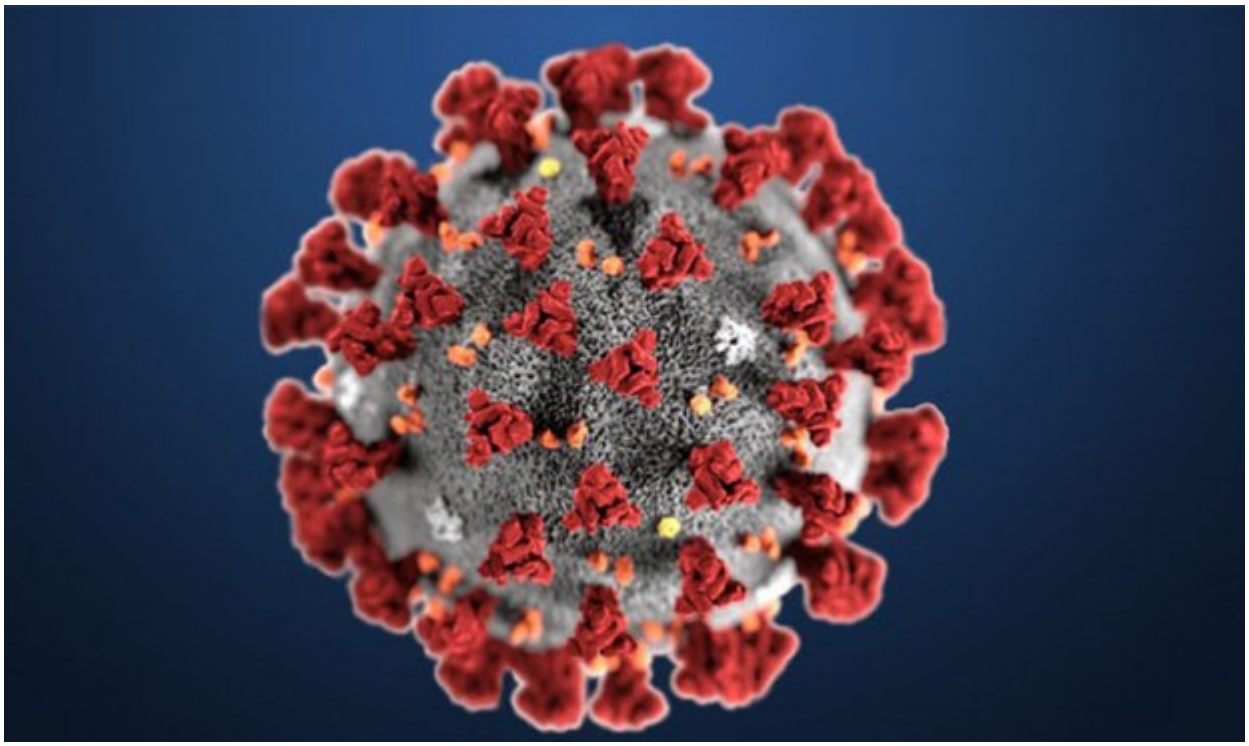
~ Modeling the Covid 19 pandemic in Norway and Spain ~

Alicia Lionneton and Silje Marie Anfindsen

06/06/2021



Facultat de Matemàtiques i Estadística
UNIVERSITAT POLITÈCNICA DE CATALUNYA

Abstract

The following report presents a predictive model for the total number of Covid-19 cases per million per day in Norway and Spain. It also tries to highlight the specific aspects of each country. The total number of Covid-19 cases per million is a convenient indicator because the new number of Covid-19 cases per million per day can be deduced from it and it is easy to predict it in the short time. We conducted a Bayesian analysis in order to make a prediction for the total number of Covid-19 cases per million for seven days. We stated that the total number of cases per million in one day was following a Normal distribution with variance sigma and centered in a value that was depending on the total number of cases per million of the day before times a particular parameter beta plus the total number of cases per million of seven days before times a particular parameter gamma plus a constant alpha. For each parameter, we obtained the posterior distribution through Markov Chain Monte Carlo algorithm. The prediction obtained for seven days concurs with the real data.

Contents

Introduction

Since 2019, people from all over the world have been infected by Covid-19. In function of the number of cases in their country, the governments implemented some measures in order to regulate the outbreak: curfew, lockdown, barrier gestures...

A predictive model of the number of cases could help them to anticipate what is going to happen in the next few days and help them to implement the accurate measures. In this report, a short-term predictive model is presented. It aims to predict how many total Covid-19 cases per million there will be in the next seven days in two chosen countries. After modeling the total Covid-19 cases per million for some days, several analyses can be done. It is possible to examine if the prediction is correct, to see if evolution of the total cases per million is the same in the two countries chosen and study for which country the modeling is the best. It is also possible to predict the number of new cases per million by each day.

Two countries are studied : Spain and Norway. Spain and Norway are different. Spain is in the south of Europe whereas Norway is in the north of Europe. There is 46.8 million people in Spain and 5.4 million people in Norway. A lot of measures were implemented in Spain while the restrictions were not numerous in Norway. The dataset used in this study was collected from the website ourworldindata [Max Roser, Hannah Ritchie, Esteban Ortiz-Ospina and Joe Hasell]. It contains a lot of variables. Only some of them were studied in this research : the name of the country, the day and the total number of cases per million.

In the case of this study, it was more appropriate to analyze the total cases per million because Spain and Norway have a different demography. It will have been difficult to compare only the number of total cases. The study is focused only on the period of May and June because it aims to do short term prediction.

The model uses Bayesian methods in order to achieve its goals. It was implemented in the 4.0.4 version of R software. Several packages have been used.

I. Data description

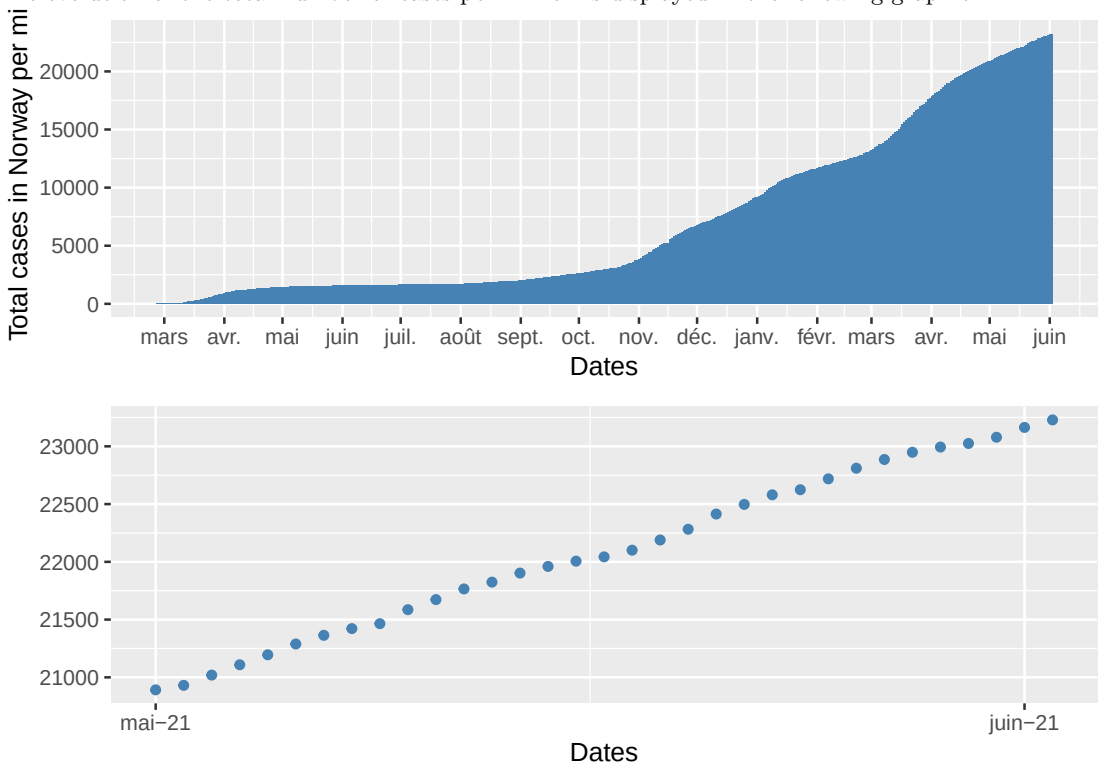
Before implementing any model, it is interesting to study how the data behaves. It helps to make correct assumptions and a posteriori to judge if the results obtained are likely to be true or not.

A. Norway

For Norway, data is available from the 26th of February 2020 to the 2nd of June 2021. In the following table, it is possible to observe the first values of the total cases per million and the new cases per million.

location	date	total_cases_per_million	new_cases_per_million
Norway	2020-02-23	NA	NA
Norway	2020-02-26	0.184	0.184
Norway	2020-02-27	0.184	0.000
Norway	2020-02-28	1.107	0.922
Norway	2020-02-29	2.767	1.660
Norway	2020-03-01	3.505	0.738

The evolution of the total number of cases per million is displayed in the following graphic.



The evolution of the total number of cases in Norway is not linear. From April 2020 to November 2020, the evolution of the total number of Covid-19 cases is very slow. It means that there were a few of new cases each day. From December 2020 to April 2021, each day the total number of cases enhances. The number of new cases by day was consequently increasing.

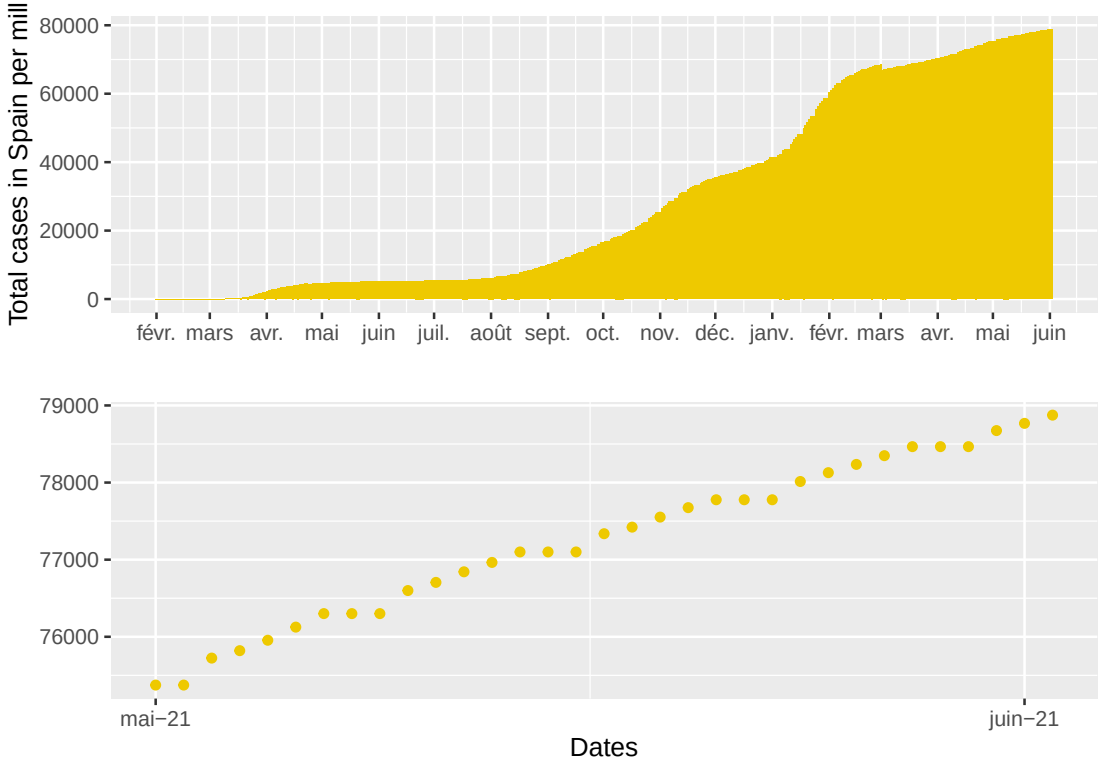
As we only want to make some predictions for the last day of May, it could be appropriate to have an idea of the total number of cases per million during the month of May. It can be seen that the total number of cases increases at a constant pace. It seems that it could be possible to multiply the value of one day by a constant in order to obtain the total number of cases of the day that follows.

B. Spain

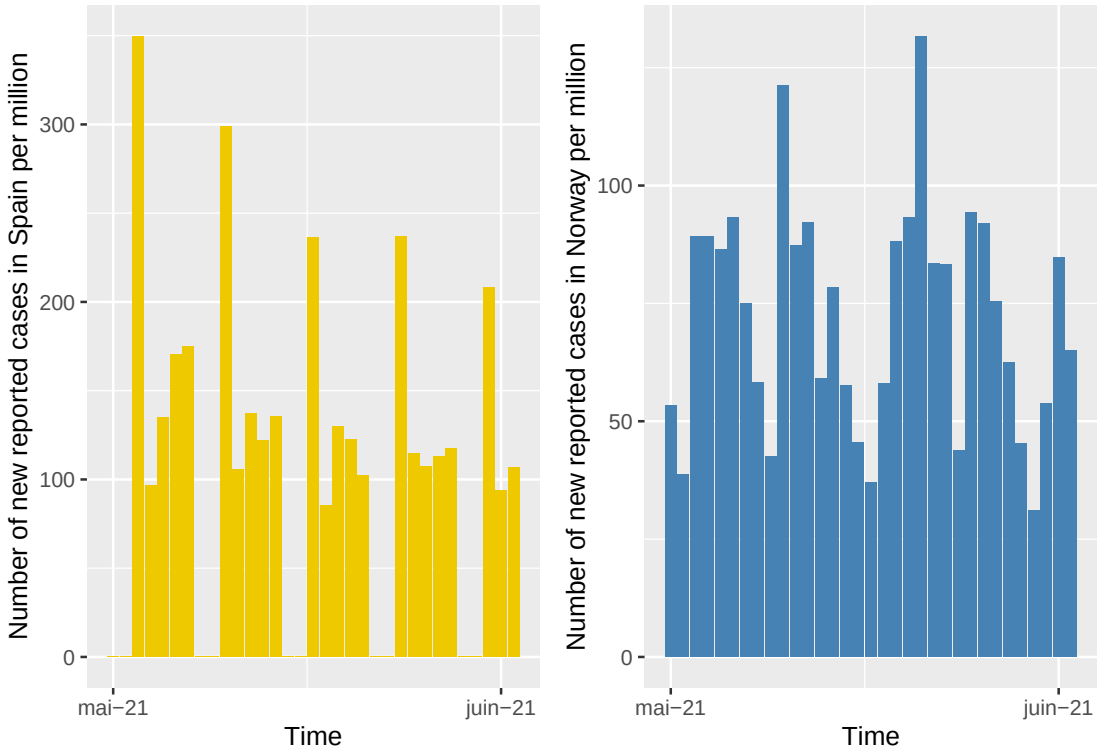
For Spain, data is available from the 2nd of February 2020 to the 2nd of June 2021. In the following table, it is possible to observe the first values of the total cases per million and the new cases per million.

location	date	total_cases_per_million	new_cases_per_million
Spain	2020-02-01	0.021	0.021
Spain	2020-02-02	0.021	0.000
Spain	2020-02-03	0.021	0.000
Spain	2020-02-04	0.021	0.000
Spain	2020-02-05	0.021	0.000
Spain	2020-02-06	0.021	0.000

The evolution of the total number of cases per million is shown in the following graphic.



The evolution of the total number of cases per million of Spain has a pattern similar to the one of Norway. However, it is not in the same scale because the number of cases in Spain is greater than in Norway independent of the population size. The graphic below is a zoom on the month of May 2021. The total number of cases is increasing slowly. The value of one day depends on the value of the day before.



The graphic above shows the number of new cases per million in both countries. We notice that there is a seasonality present in the two datasets. A week seems to have the same trend as the week before meaning we have a weekly seasonality in the data set. This is especially visible for the weekends in Spain where you see that there are no reported cases and than there is a "jump" on each Monday. In Norway it seems that new cases are also reported in the weekends, but maybe less people test themselves or the test capacity is less. We can for both countries conclude that there is a weekly seasonality in the data set.

This seasonality for the number of new cases per million is likely to have an impact on the total cases per million. From the graphic below we expect that in order to make a prediction for one day, it could be necessary to take into account the value of the total number of cases per million of the same day one week before.

II. The Bayesian model

To predict the value of the total number of cases per million for day $n+1$, it is necesary to take into account the value of day n and day $n-6$ because there is a time dependency. Thus, the predicted value was defined as a function of this two values. The predicted value will follow a normal distribution centered in the effect of day n and day $n-6$ as it can be seen in the following equation.

$$y_{n+1} \sim \mathcal{N}(\alpha + \beta * y_n + \gamma * y_{n-6}\ ; \sigma_y)$$

$\alpha + \beta * y_n + \gamma * y_{n-6}$ is an autoregression function that represents the dependency on the previous values.

All the constants introduced in the model will follow a particular distribution. It was consequently mandatory to introduce a prior distribution for α,β,γ and σ.

For Spain and Norway, the same prior distributions were chosen because there were not a lot of information about how much the value of days n and $n-6$ could have an impact on the day n. A constant α is introduced because it could help the model to be more accurate.

$$\alpha \sim \mathcal{N}(0,\ 10)$$

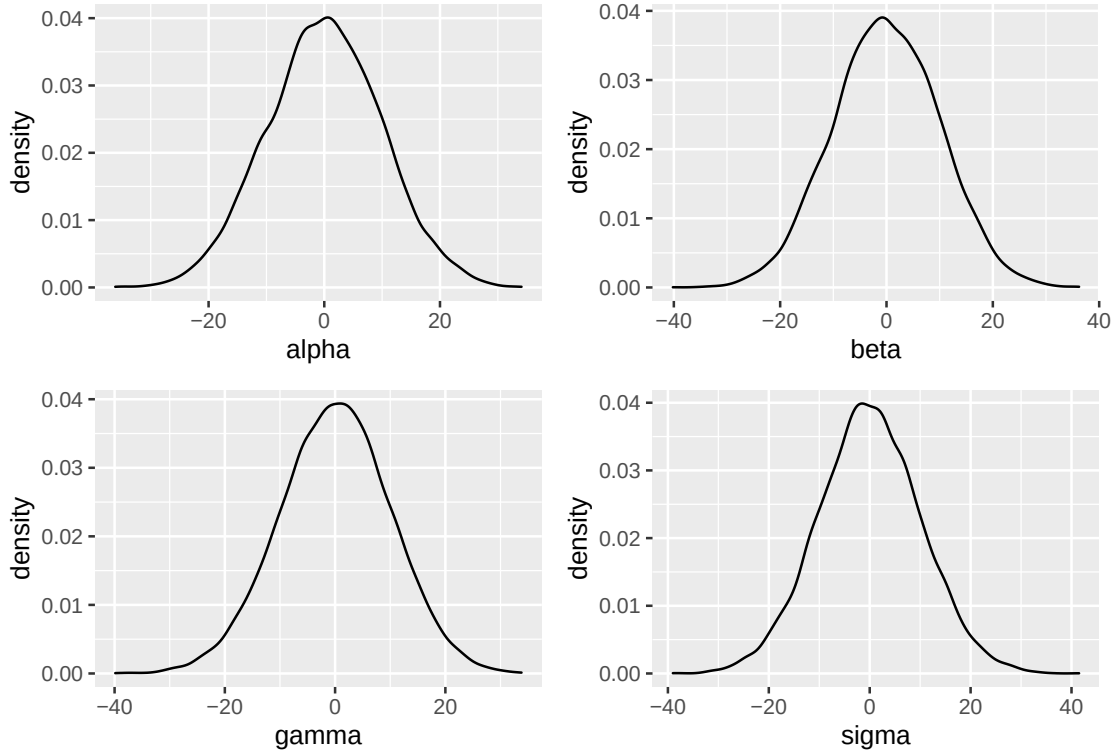
$$\beta \sim \mathcal{N}(0,\ 10)$$

$$\gamma \sim \mathcal{N}(0,\ 10)$$

$$\sigma \sim \mathcal{N}(0,\ 10)$$

A normal distribution allows us to choose around which value the parameter is. The priors are centered in 0 and they have a high variability because we do not have much information about them a priori.

They are displayed in the following graphics.



III. The results

A. Convergence

We used the data from the 10 of May to the 26 of May. We performed a Markov Chain Monte Carlo algorithm in order to find the posterior distribution of our parameter. We checked the convergence of the model of Norway and the model of Spain by examining the R-hat values for all parameters and plotting the chains traces.

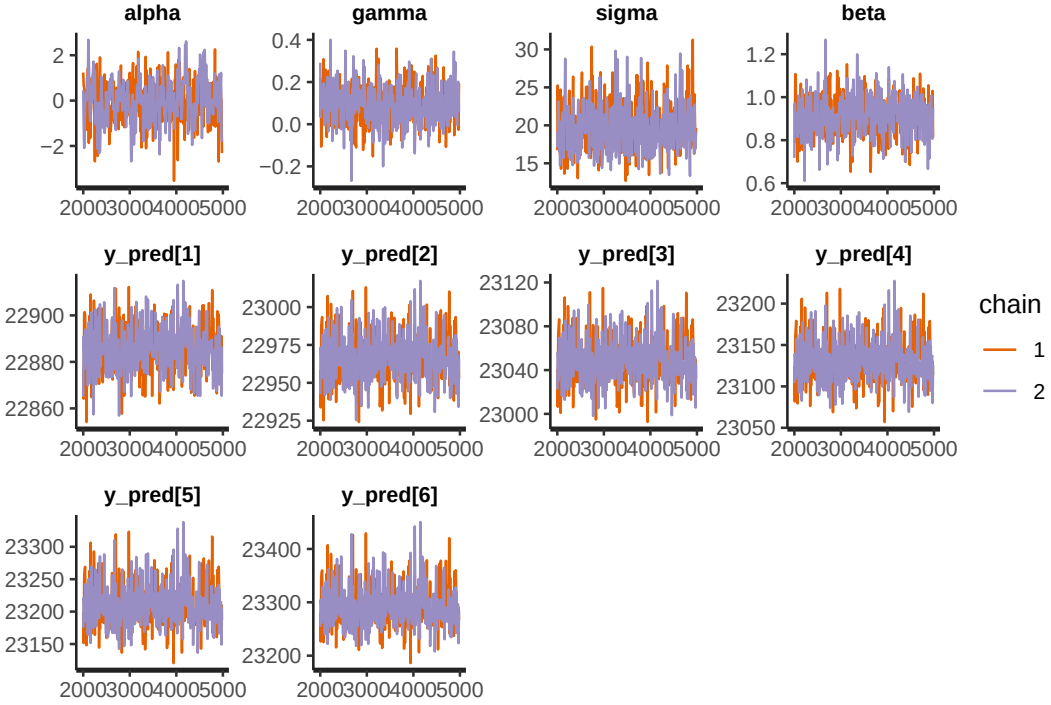
1. Norway

The table below shows the mean value, the effective sample size and the R-hat value of each parameter. For the model for Norway, all the R-hat values are equal to one which indicates convergence.

	mean	n_eff	Rhat
alpha	-0.05	313.72	1
beta	0.91	489.42	1
gamma	0.09	489.02	1
sigma	19.75	526.02	1

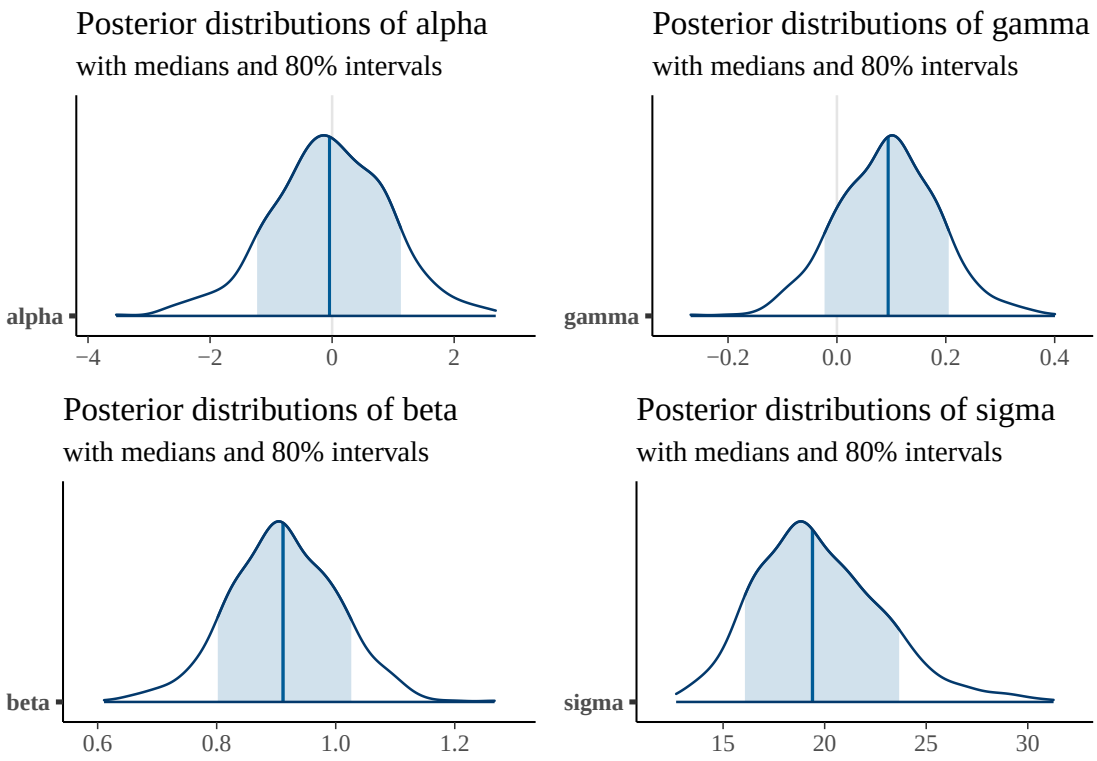
Below is the trace plot for each chain in the fitting of the parameters of the model. We can confirm that there is convergence as the chains mix well.

```
## 'pars' not specified. Showing first 10 parameters by default.
```



We examined the mean of the parameters and noticed that beta was very close to one, while gamma was almost zero. This indicates that the model is very influenced by the number of cases per million of the day before and less affected by the number of cases of one week in advance. The alpha coefficient is also close to zero and thus has a low impact on the model.

The following graphics display the posterior distribution of the parameters of interest. The means are not anymore equal to zero for all the coefficients.



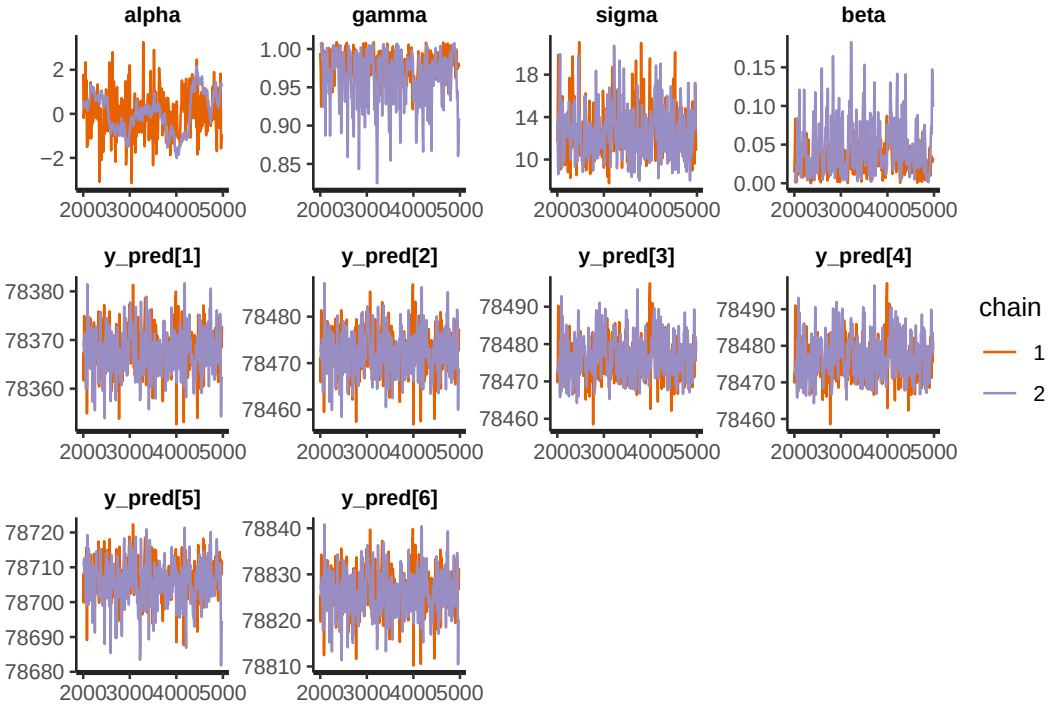
2. Spain

The table below shows the mean value of each parameter, the effective sample size and the R-hat value for the model for Spain. All the R-hat values are near of one, indicating convergence.

	mean	n_eff	Rhat
alpha	0.02	40.84	1.00
beta	0.04	99.55	1.04
gamma	0.97	99.63	1.04
sigma	12.62	427.20	1.00

Below is the trace plot for each chain in the fitting of the parameters of the model. We can also for this model confirm that there is convergence as the chains mix well.

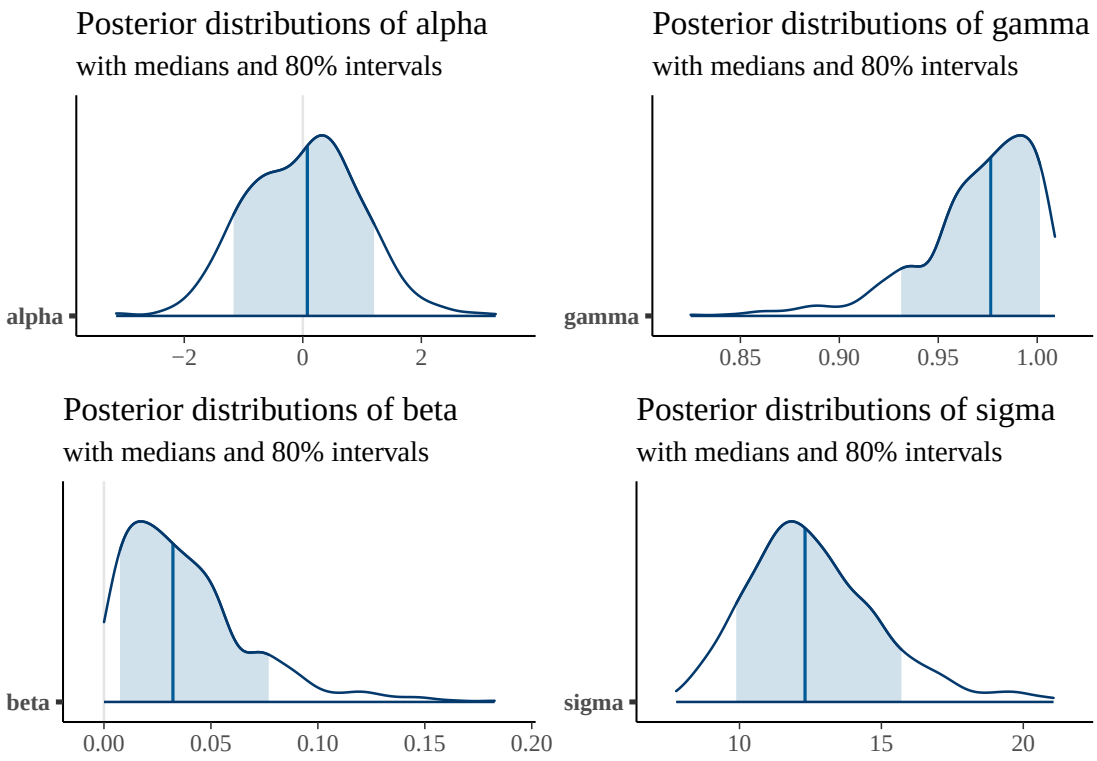
```
## 'pars' not specified. Showing first 10 parameters by default.
```



From the mean of the parameters, we noticed that gamma was greater than the other coefficients, indicating that for this model, the value of seven days before dominates the model a lot compared to the one of one day before, which is linked with the beta coefficient. As with the Norwegian data, the alpha coefficient has a low impact on the model as it is very small.

From what we discovered in the data description part, it makes sense that the model for Spain has a greater gamma and then lower beta than the model for Norway, which is almost the opposite. As we noticed Spain does not report any cases at all in the weekends, which Norway has done. Thus we could expect the y_{n+1} for Spain to have a greater dependency on the coefficient gamma linked with y_{n-6} than in the model for Norway.

The posterior distribution of the parameters are presented in the fourth following graphics.

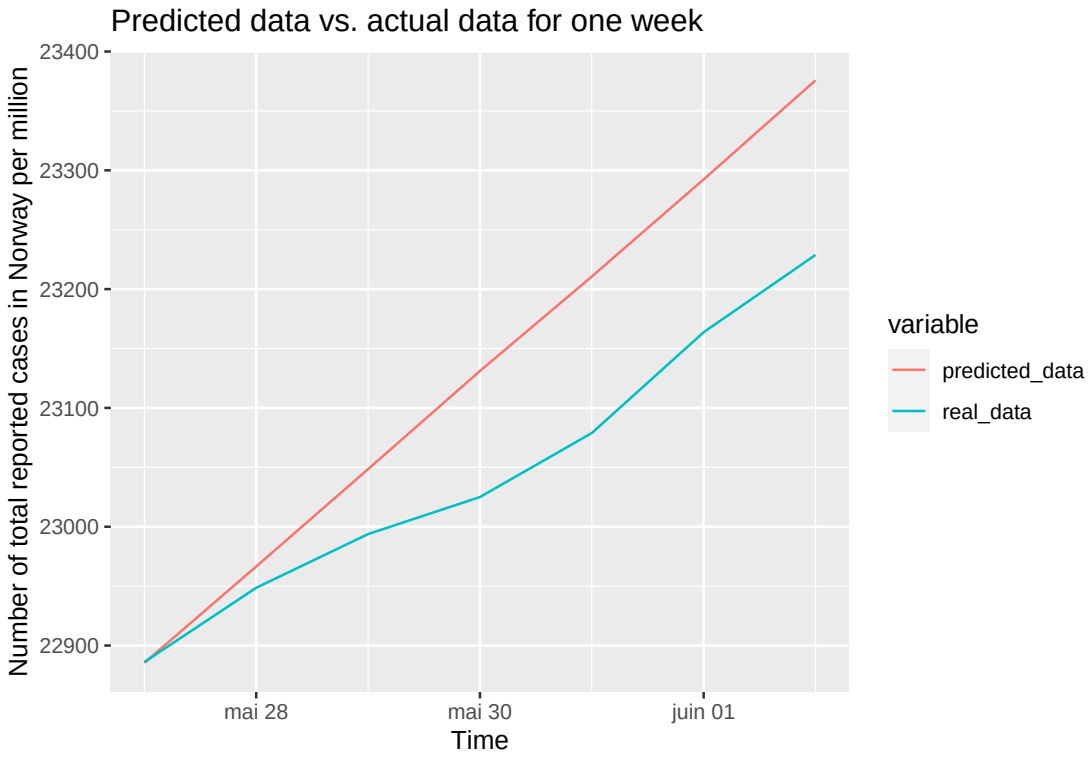


C. Prediction

We did a prediction of the total number of cases per million from the 27th of May to the 2nd of June. To know if the predictions we obtained were right, we compared them with the real total number of cases per million of this week. We did the comparison for Norway and Spain.

1. Norway

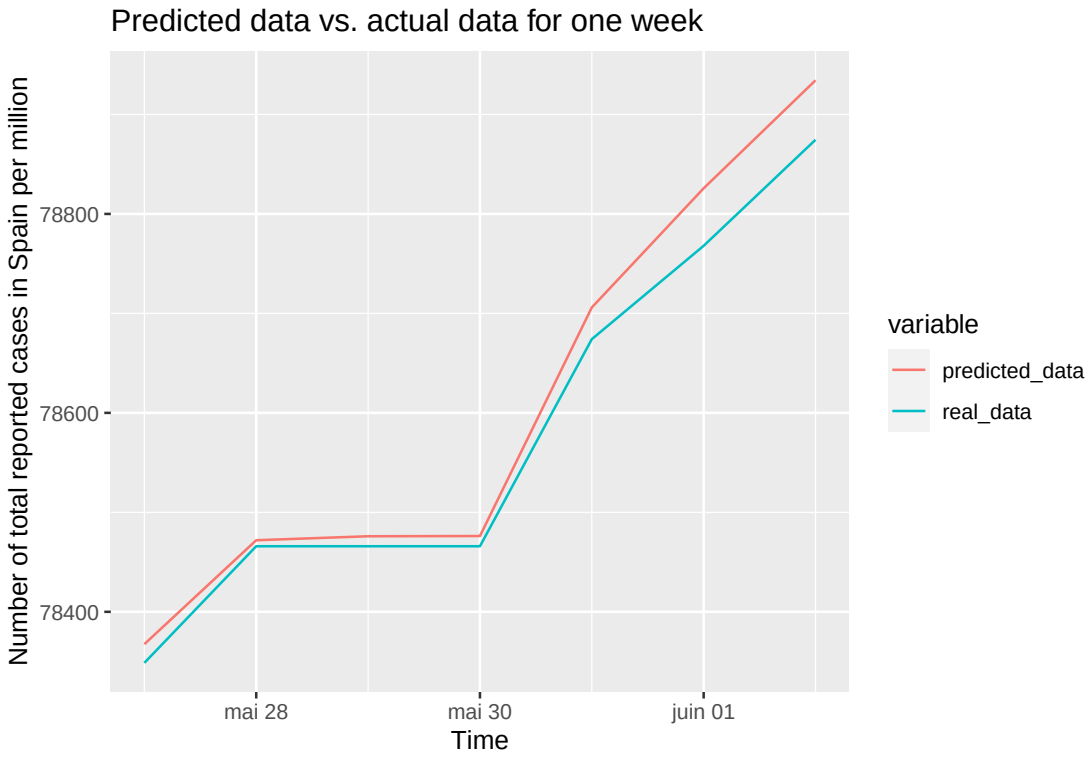
The prediction for Norway is presented in the graphic below.



The first and the second prediction are very near of the real value. The third prediction stays accurate but it is less precise. The model is predicting that the total number of cases per million will be around 23 050 people in May 29th and the real total number of cases per million was 23 000 people. For the fourth last days, the prediction is less precise. There is almost a difference of 150 people between the total number of cases per million and the real total number of cases per million.

2. Spain

The prediction for Spain is exhibited in the graphic below.



The prediction for Spain is exact. The model is able to grasp the different evolutions for each day. That is because of the gamma coefficient that we introduced in the model. The value predicted for one day depends a lot on the value of seven days before. Only the last prediction is displaying a difference of 50 people between the total number of cases per million and the real total number of cases per million.

Conclusion

Our results show that with our bayesian model with some time dependencies, we can predict the total number of cases of the covid-19 outbreak per million in Spain and Norway for one week in the future.

The model for the two countries use the same prior information and the same statistical model. As we noticed from the data description part that Spain had a more dominating weekly seasonality than Norway, perhaps we could have chosen a gamma distribution with greater mean compared to the mean of the beta distribution for Spain. The total number of Covid 19 cases in Norway depends more on the value of the day before than the value of seven days before. Consequently, the mean of the beta coefficient could have been slightly enhanced by 0.5.

In the model, we only used data from May 2021 and not from before. This is in order to make the model more precise. If we had used more data, the coefficients of our model would have been bigger or smaller because the trend is changing along the time. We wanted to only capture the trend of the last month.

In that way it made more sense to look at a shorter time interval where the total number of cases per million had an almost linear increasing trend. Another reason for doing this is because of the many restrictions that vary both in time and country. As the restrictions are different, the trend can change. Consequently, it was appropriate to select only the total number of Covid-19 cases per million of May to do the prediction into the week that begins the 27 of May and ends the 2 of June.

Our analysis could have been enriched by the Gompertz model [ÁBerihuete, A.;Sánchez-Sánchez, M.; Suárez-Llorens,A.] if we had more time and more knowledge. This model is able to take into account more information and can precisely predict the number of new cases. It adequately takes into account some time dependencies.

References

1. Max Roser, Hannah Ritchie, Esteban Ortiz-Ospina and Joe Hasell (2020) - "Coronavirus Pandemic (COVID-19)". Published online at OurWorldInData.org. Retrieved from: 'https://ourworldindata.org/coronavirus' [Online Resource]

2. ÁBerihuete, A.;Sánchez-Sánchez, M.; Suárez-Llorens,A. A Bayesian Model of COVID-19 Cases Based on the Gompertz Curve. Mathematics 2021, 9, 228. https://doi.org/10.3390/math9030228

Appendix

```
library(knitr)
include_graphics("image1.jpg")

library(dplyr)
library(ggplot2)
library(gridExtra)
library(rstan)
library(bayesplot)
library(tidyverse)
library(reshape2)

library(dplyr)
library(readxl)
table <- read_excel("~/Ecole/Erasmus/analyse bayésienne/
                    projet/projet/covid.xlsx")
norway <- subset(table, location =="Norway")
spain <- subset(table, location =="Spain")

spain <- select(spain, location, date,
                total_cases_per_million, new_cases_per_million)
norway <- select(norway, location, date,
                 total_cases_per_million, new_cases_per_million )

write.csv(norway, "norway")
write.csv(spain, "spain")

norway <- read.csv("norway")
spain <- read.csv("spain")
spain <- select(spain,
                location, date, total_cases_per_million, new_cases_per_million)
norway <- select(norway,
                 location, date,total_cases_per_million, new_cases_per_million)

spain$date <- as.Date(spain$date)
norway$date <- as.Date(norway$date)


kable(head(norway))

gg_no1 <- ggplot(data=norway) +
  geom_bar(mapping = aes(x=date,y = total_cases_per_million),
           stat = "identity", fill="steelblue") +
  labs(x="Dates", y="Total cases in Norway per mill") +
  scale_x_date(date_breaks = "months" , date_labels = "%b")

sub_no <- subset(norway, date >= "2021-05-01")

gg_no2 <- ggplot(data=sub_no) +
  geom_point(mapping = aes(x=date,y = total_cases_per_million),
             stat = "identity", color="steelblue") +labs(x="Dates",y="") +
  scale_x_date(date_breaks = "months" , date_labels = "%b-%y")

grid.arrange(gg_no1, gg_no2, nrow=2)

kable(head(spain))
```

```r
gg_esp1 <- ggplot(data=spain) +
  geom_bar(mapping = aes(x=date,y = total_cases_per_million),
           stat = "identity", fill="gold2") +
  labs(x="",y="Total cases in Spain per mill") +
  scale_x_date(date_breaks = "months" , date_labels = "%b")

sub_es <- subset(spain, date >= "2021-05-01")

gg_esp2 <- ggplot(data=sub_es) +
  geom_point(mapping = aes(x=date,y = total_cases_per_million),
             stat = "identity", color="gold2") +labs(x="Dates",y="") +
  scale_x_date(date_breaks = "months" , date_labels = "%b-%y")

grid.arrange(gg_esp1,gg_esp2,nrow=2)


gg_es <- ggplot(data=sub_es) +
  geom_bar(mapping = aes(x=date,y = new_cases_per_million),
           stat = "identity", fill="gold2") +
  labs(x="Time",y="Number of new reported cases in Spain per million") +
  scale_x_date(date_breaks = "month" , date_labels = "%b-%y")

gg_no <- ggplot(data=sub_no) +
  geom_bar(mapping = aes(x=date,y = new_cases_per_million),
           stat = "identity",
           fill="steelblue") +
  labs(x="Time",y="Number of new reported cases in Norway per million") +
  scale_x_date(date_breaks = "month" , date_labels = "%b-%y")

grid.arrange(gg_es, gg_no,ncol = 2)


# Prior distribution of the parameters for Spain and Norway
alpha <- rnorm(10000, 0,10)
gg_a <- ggplot(tibble(alpha), aes(alpha)) +
  geom_density()

beta <- rnorm(10000, 0,10)
gg_beta <- ggplot(tibble(beta), aes(beta)) +
  geom_density()

gamma <- rnorm(10000, 0,10)
gg_gamma <- ggplot(tibble(gamma), aes(gamma)) +
  geom_density()

sigma <- rnorm(10000, 0,10)
gg_sigma <- ggplot(tibble(sigma), aes(sigma)) +
  geom_density()

grid.arrange(gg_a,gg_beta,gg_gamma, gg_sigma, nrow=2, ncol=2)

covid_norway_may <- norway[norway$date >= "2021-05-10" &
                             norway$date <= "2021-05-26" ,]

# Stan model for Norway
stan_norway <- list(
  n = nrow(covid_norway_may),
  mu_a = 0,
```

```
  mu_beta = 0,
  mu_gamma = 0,
  mu_sigma =0,
  y = covid_norway_may$total_cases_per_million
)

fit_norway <- stan("stan_number_infected.stan", data = stan_norway, chains = 2,
                   iter = 5000, warmup = 2000, thin = 10, refresh=0)

summary_nor <- summary(fit_norway,
                       pars = c("alpha", "beta", "gamma", "sigma"),
                       probs = c(0.1, 0.9))$summary
summary_nor <- summary_nor[,c("mean", "n_eff", "Rhat")]
summary_nor <- round(summary_nor, 2)
kable(summary_nor[,c("mean", "n_eff", "Rhat")])

traceplot(fit_norway)

posterior <- as.data.frame(fit_norway)

plot_title <- ggtitle("Posterior distributions of alpha",
                      "with medians and 80% intervals")
gg10 <- mcmc_areas(posterior,
           pars = c("alpha"),
           prob = 0.8) + plot_title

plot_title <- ggtitle("Posterior distributions of gamma",
                      "with medians and 80% intervals")
gg11 <- mcmc_areas(posterior,
           pars = c("gamma"),
           prob = 0.8) + plot_title

plot_title <- ggtitle("Posterior distributions of beta",
                      "with medians and 80% intervals")
gg12 <- mcmc_areas(posterior,
           pars = c("beta"),
           prob = 0.8) + plot_title

plot_title <- ggtitle("Posterior distributions of sigma",
                      "with medians and 80% intervals")
gg13 <- mcmc_areas(posterior,
           pars = c("sigma"),
           prob = 0.8) + plot_title

grid.arrange(gg10,gg11,gg12,gg13,nrow=2,ncol=2)


covid_spa_may <- spain[spain$date >= "2021-05-12" &
                         spain$date <= "2021-05-26" ,]

stan_spain <- list(
  n = nrow(covid_spa_may),
  mu_a = 0,
  mu_beta = 0,
  mu_gamma = 0,
  mu_sigma =0,
  y = covid_spa_may$total_cases_per_million
)
```

```r
fit_spain <- stan("stan_number_infected.stan", data = stan_spain, chains = 2,
                  iter = 5000, warmup = 2000, thin = 10, refresh=0)


summary_spain <- summary(fit_spain,
                         pars = c("alpha", "beta", "gamma", "sigma"),
                         probs = c(0.1, 0.9))$summary
summary_spain <- summary_spain[,c("mean", "n_eff", "Rhat")]
summary_spain <- round(summary_spain, 2)
kable(summary_spain[,c("mean", "n_eff", "Rhat")])

traceplot(fit_spain)

posterior_s <- as.data.frame(fit_spain)

plot_title <- ggtitle("Posterior distributions of alpha",
                      "with medians and 80% intervals")
gg1 <- mcmc_areas(posterior_s,
           pars = c("alpha"),
           prob = 0.8) + plot_title

plot_title <- ggtitle("Posterior distributions of gamma",
                      "with medians and 80% intervals")
gg2 <- mcmc_areas(posterior_s,
           pars = c("gamma"),
           prob = 0.8) + plot_title

plot_title <- ggtitle("Posterior distributions of beta",
                      "with medians and 80% intervals")
gg3 <- mcmc_areas(posterior_s,
           pars = c("beta"),
           prob = 0.8) + plot_title

plot_title <- ggtitle("Posterior distributions of sigma",
                      "with medians and 80% intervals")
gg4 <-mcmc_areas(posterior_s,
           pars = c("sigma"),
           prob = 0.8) + plot_title

grid.arrange(gg1,gg2,gg3,gg4, nrow=2,ncol=2)



#posterior predictive
posterior_nor <- as.data.frame(fit_norway)
y_pred_nor <- rep(0,7)
y_pred_nor[1] <- mean(posterior_nor$`y_pred[1]`)
y_pred_nor[2] <- mean(posterior_nor$`y_pred[2]`)
y_pred_nor[3] <- mean(posterior_nor$`y_pred[3]`)
y_pred_nor[4] <- mean(posterior_nor$`y_pred[4]`)
y_pred_nor[5] <- mean(posterior_nor$`y_pred[5]`)
y_pred_nor[6] <- mean(posterior_nor$`y_pred[6]`)
y_pred_nor[7] <- mean(posterior_nor$`y_pred[7]`)


#make plot for prediction vs. data
covid_nor_june <- norway[norway$date >= "2021-05-27" ,]
predicted_data <- y_pred_nor
```

```r
real_data <- covid_nor_june$total_cases_per_million

new_data_nor <- data.frame(covid_nor_june$date, predicted_data, real_data )

# melt the data to a long format
df3 <- melt(data = new_data_nor, id.vars = "covid_nor_june.date")

# plot, using the aesthetics argument 'colour'
ggplot(data = df3, aes(x = covid_nor_june.date, y = value, colour = variable))+
  geom_line() +
  ggtitle("Predicted data vs. actual data for one week") +
  labs(x="Time",y="Number of total reported cases in Norway per million")




#posterior predictive distributions
posterior <- as.data.frame(fit_spain)

y_pred_es <- rep(0,7)
y_pred_es[1] <- mean(posterior$`y_pred[1]`)
y_pred_es[2] <- mean(posterior$`y_pred[2]`)
y_pred_es[3] <- mean(posterior$`y_pred[3]`)
y_pred_es[4] <- mean(posterior$`y_pred[4]`)
y_pred_es[5] <- mean(posterior$`y_pred[5]`)
y_pred_es[6] <- mean(posterior$`y_pred[6]`)
y_pred_es[7] <- mean(posterior$`y_pred[7]`)


#make plot for prediction vs. data
covid_spa_june <- spain[spain$date >= "2021-05-27" ,]

predicted_data <- y_pred_es
real_data <- covid_spa_june$total_cases_per_million
new_data <- data.frame(covid_spa_june$date, predicted_data, real_data )


# melt the data to a long format
df2 <- melt(data = new_data, id.vars = "covid_spa_june.date")

# plot, using the aesthetics argument 'colour'
ggplot(data = df2, aes(x = covid_spa_june.date, y = value, colour = variable))+
  geom_line() +
  ggtitle("Predicted data vs. actual data for one week") +
  labs(x="Time",y="Number of total reported cases in Spain per million")
```