Bayesian Data Analysis

Subhasish Basak (MDS201803)

Bayesian Analysis of COVID-19 dataset

May 4, 2020

1 Introduction

In the context of current situation of the pandemic due to **COVID-19** (corona-virus), in this project we run Bayesian Analysis to check whether the lock-down is effective or not, based on the daily incidence report for different countries. For our purpose we have worked with data sets for 3 countries viz. India, Spain and Italy. Followed by a simple linear regression model fitting we move on to Bayesian Regression modelling of the data set to visualize the spread of the virus.

- **Dataset** : We have used the data repository for the 2019 Novel Coronavirus by the Johns Hopkins University. The database is available here: https://github.com/CSSEGISandData/COVID-19
- **Methodologies used** : OLS analysis of linear regression, Bayesian Regression with Flat Prior, Bayesian Regression with Conjugate Priors.

2 Data Preprocessing

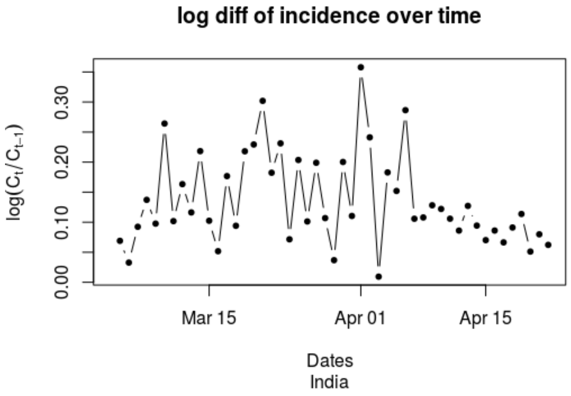
The original dataset contains data for several countries. We need to extract data only for the countries we are interested in.

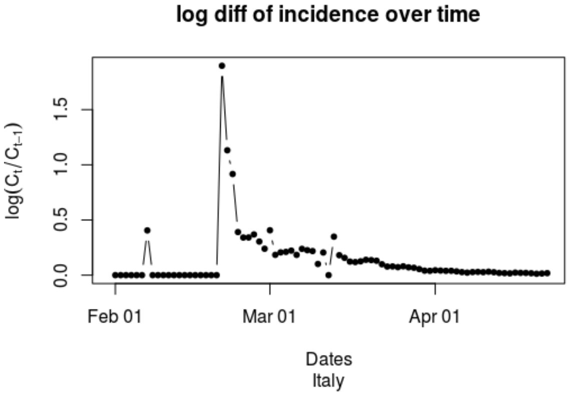
- **Splitting the data** : We are only interested in the date of incidence and the number of daily incidence reported. We extract those columns from the dataframe.
- **Computing log diff** : Here we work with the logarithm of the difference of the daily incidences reported instead of the raw figures. It is given by the equation,

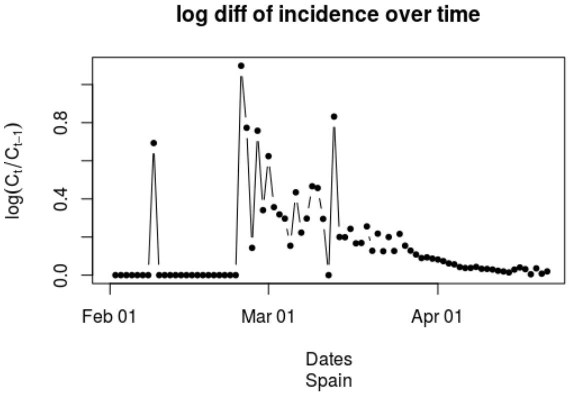
$$ln\_t = log(incidence_t) - log(incidence_{t-1}) \tag{1}$$

3 Visualization

We present below the plots of the log difference of the incidence numbers.







4 The Regression Model

We propose the following regression model, for modelling the log difference of the reported incidence rate, i.e. our response variable is *ln_t* and as the predictor variables we take *Time*, *Time*^{2} and *Lockdown period*. The variable *Time* just maintains a count of the days from the starting date. The variable *Lockdown* is a binary variable which is 1 when the lockdown is in effect.

$$ln\_t = \beta_0 + \beta_1 Time + \beta_2 Time^2 + \beta_3 Lockdown + \epsilon \tag{2}$$

Where ϵ is the i.i.d Gaussian noise.

4.1 Results & conclusions of Liner Regression

Given the summary results of the linear regression model we conclude the following :

- For India the coefficient corresponding to *Time* is statistically significant at 1% level, which is too weak to conclude the Effect of *Time* on the log differenced incidence rate. On the other hand the *Lockdown* variable turns out to be significant.
- For Italy **none** of the variables turn out to be significant even in 1% level.
- For Spain also the coefficients corresponding *Time* and *Time*^{2} are not significant and the coefficient of lockdown is significant at 5% level.

We present below the summary results of fitting regression with OLS method for all the countries,

R code 4.1: Output for India

```
OLS results for India

Call:
lm(formula = ln_t ~ Time + I(Time^2) + lock_down, data = country_data2)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.144447 -0.035494 -0.006156  0.031356  0.201765 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)   
(Intercept)  6.698e-02  3.637e-02   1.842  0.07167 . 
Time         9.571e-03  3.582e-03   2.672  0.01028 * 
I(Time^2)   -1.780e-04  5.233e-05  -3.401  0.00136 **
lock_down   -3.882e-02  4.024e-02  -0.965  0.33956   
---
Signif. codes:  0   ***   0.001   **   0.01   *   0.05   .   0.1   
       1

Residual standard error: 0.0672 on 48 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.2534,    Adjusted R-squared:  0.2068 
F-statistic: 5.432 on 3 and 48 DF,  p-value: 0.002682
```

R code 4.2: Output for Italy

```
OLS results for Italy

Call:
lm(formula = ln_t ~ Time + I(Time^2) + lock_down, data = country_data2)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.25208 -0.09384 -0.03464  0.02490  1.63468 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)   
(Intercept) -4.418e-02  9.378e-02  -0.471  0.63888   
Time         1.768e-02  5.421e-03   3.262  0.00162 **
I(Time^2)   -1.703e-04  5.088e-05  -3.347  0.00124 **
lock_down   -2.949e-01  1.078e-01  -2.735  0.00766 **
---
Signif. codes:  0   ***   0.001   **   0.01   *   0.05   .   0.1   
       1

Residual standard error: 0.2424 on 81 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.1965,    Adjusted R-squared:  0.1667 
F-statistic: 6.603 on 3 and 81 DF,  p-value: 0.0004775
```

R code 4.3: Output for Spain

```
OLS results for Spain

Call:
lm(formula = ln_t ~ Time + I(Time^2) + lock_down, data = country_data2)

Residuals:
     Min       1Q   Median       3Q      Max
-0.34021 -0.07738 -0.03602  0.03375  0.84684

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -8.562e-02  7.281e-02  -1.176   0.2431
Time         1.835e-02  4.046e-03   4.536 1.99e-05 ***
I(Time^2)   -1.944e-04  4.022e-05  -4.833 6.40e-06 ***
lock_down   -1.606e-01  8.440e-02  -1.903   0.0606 .
---
Signif. codes:  0   ***   0.001   **   0.01   *   0.05   .   0.1
      1

Residual standard error: 0.1927 on 80 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.2591,   Adjusted R-squared:  0.2313
F-statistic: 9.326 on 3 and 80 DF,  p-value: 2.339e-05
```

5 Testing Assumption

After fitting the linear model using Ordinary Least Square (OLS) method we now test our assumption of Normality on the residuals. In equation (2) we have assumed the ϵs are independent and identically distributed Gaussian r.v. To check this assumption we use the **Shapiro-Wilks test** of Normality. The test has the null hypothesis that the population is *normally distributed*. Here are the results of the test:

- **India**
  - **p-value** : 0.02
  - **conclusion** : at 5% level of significance we fail to accept the null hypothesis and conclude the residuals are not Gaussian.
- **Italy**
  - **p-value** : $2.485e - 15$
  - **conclusion** : at 5% level of significance we fail to accept the null hypothesis and conclude the residuals are not Gaussian.
- **Spain**
  - **p-value** : $4.978e - 11$
  - **conclusion** : at 5% level of significance we fail to accept the null hypothesis and conclude the residuals are not Gaussian.

Thus we propose to use Laplace distribution instead of Gaussian. Here are the QQ plots of the corresponding countries:

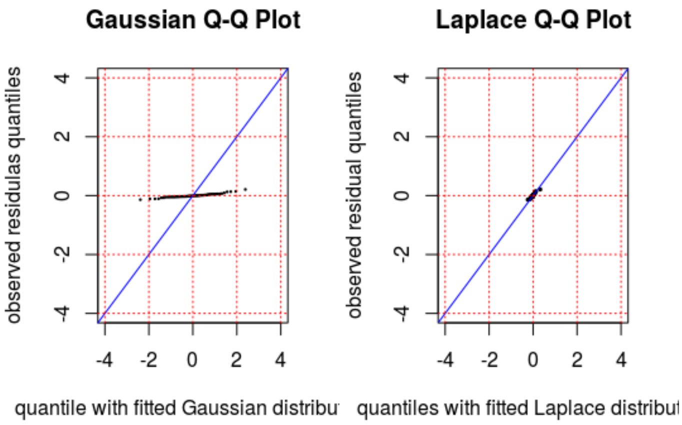


Figure 1: India

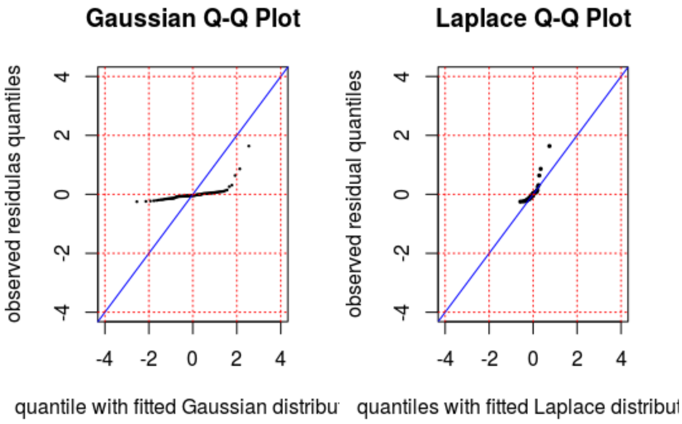


Figure 2: Italy

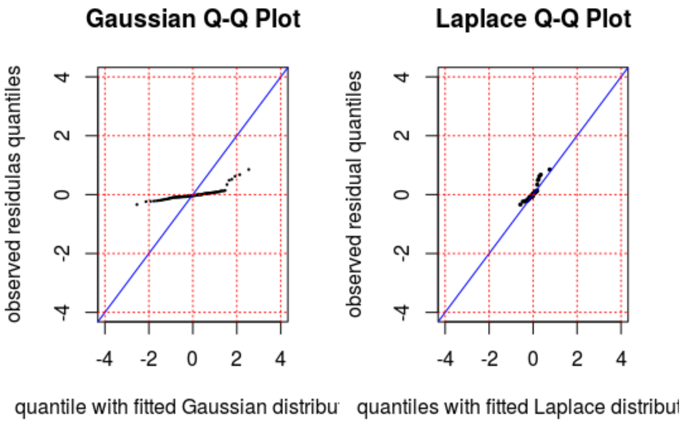


Figure 3: Spain

6 Bayesian regression

We consider the same equation (2) but here we take Laplace distribution of the errors terms.

$$\epsilon \sim Laplace(0, \lambda) \tag{3}$$

Thus the log diff of incidence follows Laplace distribution. Also we consider the following priors on the parameters:
$(\beta_0, \beta_1, \beta_2, \beta_3)$ independently follows Cauchy$(0, 1)$ distribution and λ follows Gamma$(1, 1)$.
Since we do not know the closed form of the posterior distribution, next we use the **Metropolis-Hastings** algorithm to simulate from the posterior distribution of the parameters and compute the summary statistics.

6.1 Results & Conclusions of Bayesian Analysis

We first present the results of the implementation.

R code 6.1: Output for India

```
Bayesian Regression (using MH algo with Laplace prior) model fitting summary
    for India
       beta_0 beta_1 beta_2  beta_3 lambda
median 0.0670 0.0089 -1e-04 -0.0792 0.0438
mean   0.0683 0.0081 -1e-04 -0.0769 0.0447
sd     0.0227 0.0023  0e+00  0.0347 0.0063
2.5%   0.0330 0.0025 -2e-04 -0.1318 0.0355
97.5%  0.1176 0.0119 -1e-04  0.0089 0.0589
```

R code 6.2: Output for Italy

```
Bayesian Regression (using MH algo with Laplace prior) model fitting summary
    for Italy
        beta_0 beta_1 beta_2  beta_3 lambda
median -0.0463 0.0105 -1e-04 -0.1817 0.0966
mean   -0.0460 0.0105 -1e-04 -0.1702 0.0961
sd      0.0228 0.0013  0e+00  0.0389 0.0078
2.5%   -0.0870 0.0079 -1e-04 -0.2414 0.0829
97.5%  -0.0087 0.0130 -1e-04 -0.0849 0.1131
```

R code 6.3: Output for Spain

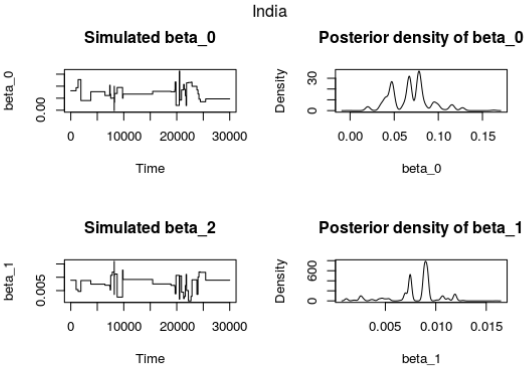
```
Bayesian Regression (using MH algo with Laplace prior) model fitting summary
    for Spain
        beta_0 beta_1 beta_2  beta_3 lambda
median -0.0658 0.0123 -1e-04 -0.1311 0.1069
mean   -0.0737 0.0122 -1e-04 -0.1312 0.1079
sd      0.0393 0.0028  0e+00  0.0723 0.0116
2.5%   -0.1602 0.0068 -2e-04 -0.2696 0.0866
97.5%  -0.0091 0.0175 -1e-04  0.0216 0.1365
```

Given the summary results of the posterior mean and standard deviation of the parameters we decide upon the *statistical significance* of a coefficient, whether it includes 0 in its symmetric posterior quantile-based credible intervals or not.

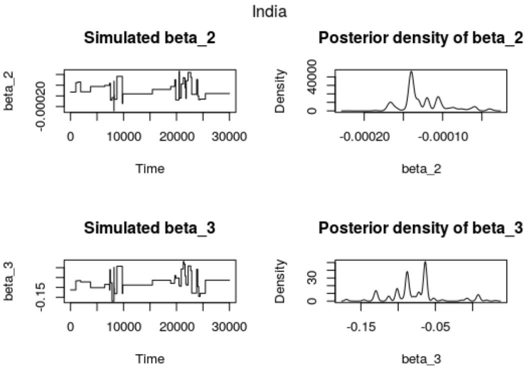
- For India the all the coefficients except the one corresponding to *Lockdown* are significant for 0.005 and 0.995 -th qunatile based CI, since they do not include 0.
- For Italy all the coefficient are significant for 0.05 and 0.95 quantile based CI.
- For Spain also the all the coefficients except the one corresponding to *Lockdown* are significant for 0.005 and 0.995 -th qunatile based CI, since they do not include 0.

6.2 Plots & Visualization of the posterior densities

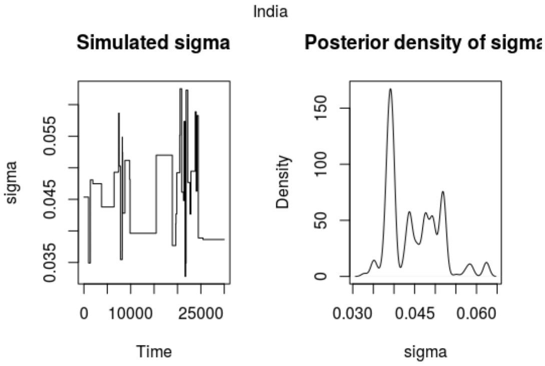
In this section we present the plots of the Monte-Carlo simulations obtained from the Metropolis-Hastings algorithm (after discarding the burning samples). For different countries we have used different *Number of simulations* and *Number of Burn-in samples*, depending on the iterations after which the Markov Chain stabilizes.



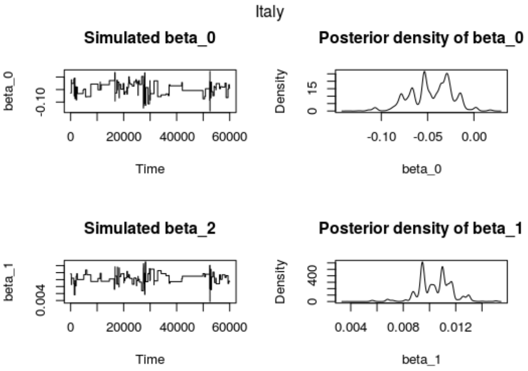
(a)



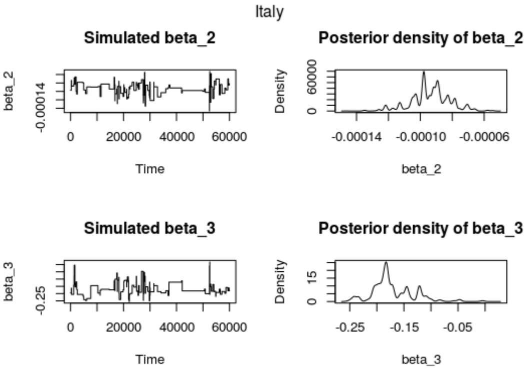
(b)



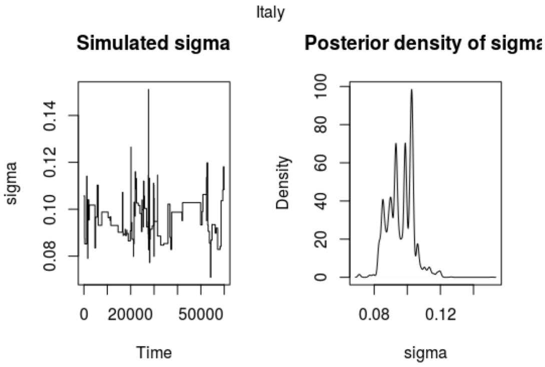
(c)

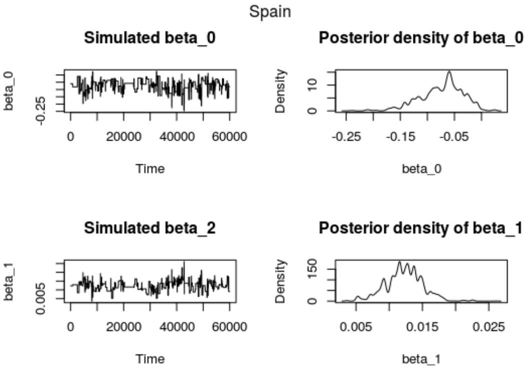


(a)

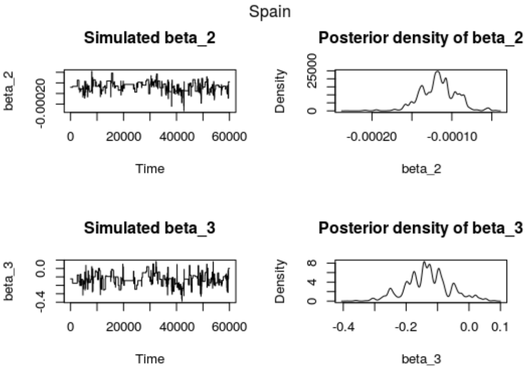


(b)

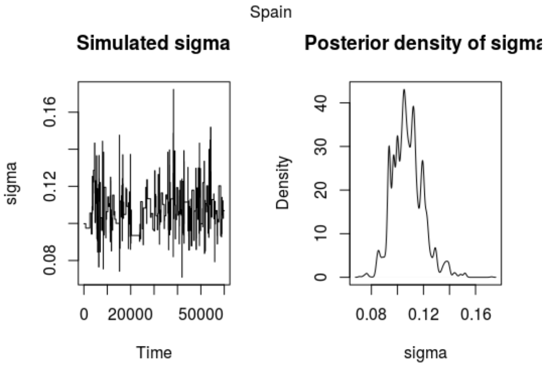

Italy
Simulated sigma
Posterior density of sigma
sigma
0.08
0.10
0.12
0.14
0
20000
50000
Time
Density
0
20
40
60
80
100
0.08
0.12
sigma


Spain
Simulated beta_0
Posterior density of beta_0
beta_0
-0.25
0
20000
40000
60000
Time
Density
0
10
-0.25
-0.15
-0.05
beta_0
Simulated beta_2
Posterior density of beta_1
beta_1
0.005
0
20000
40000
60000
Time
Density
0
150
0.005
0.015
0.025
beta_1

(a)


Spain
Simulated beta_2
Posterior density of beta_2
beta_2
-0.00020
0
20000
40000
60000
Time
Density
0
25000
-0.00020
-0.00010
beta_2
Simulated beta_3
Posterior density of beta_3
beta_3
-0.4
0.0
0
20000
40000
60000
Time
Density
0
4
8
-0.4
-0.2
0.0
0.1
beta_3

(b)


Spain
Simulated sigma
Posterior density of sigma
sigma
0.08
0.12
0.16
0
20000
50000
Time
Density
0
10
20
30
40
0.08
0.12
0.16
sigma

(c)