Bayesian Data Analysis

Chennai Mathematical Institute

COVID 19 - Bayesian Analysis

Data

We have COVID-19 data and the data used is pulled from global COVID-19 data repository from the John Hopkins University from where we can get the data to do our analysis.

Link to the Data Repository - COVID-19 Data Repository

Task

The task is to run OLS analysis of linear regression and run a Bayesian analysis for COVID-19 disease progression for India. The task is also to run a Bayesian analysis to check if lockdown has worked for Italy and Spain.

Procedure

- First we read the data and preprocess for a particular country
- We then Visualize the response variable over time which gives us a idea about the data
- We then choose the predictors
- We run OLS analysis of linear regression and interpret
- We check for nomality of residuals - Shapiro-Wilk test and Q-Q plot
- If not normal we check whether residuals follow Laplace distribution - Q-Q plot
- Now we choose prior distribution of the parameters
- If not a closed form solution for posterior, we use Metropolis-Hastings to get the posterior
- We then check the trace plot and decide the burn-in period
- Considering Burn-in period we find the posterior summary, trace plot
- We interpret our result based on the posterior summary
- We interpret our result based on the posterior summary(Mean, S.D.,95% Credible Intervals)

Functions Written

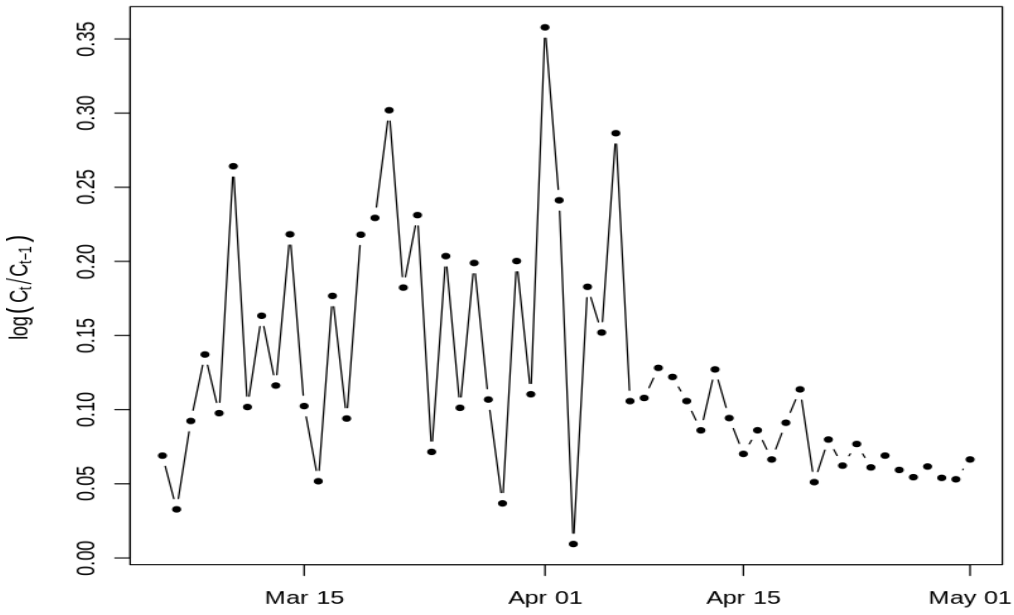
- For Pre-processing the data frame for a given country
- For Gaussian Q-Q plot
- For Laplace Q-Q plot - It again uses the following functions-
  - For density of Laplace distribution
  - For the negative log-likelihood function
  - For simulating from the fitted Laplace distribution
- For Posterior summary

- For trace plot
  - If argument "density_plot"==TRUE then it also plots the posterior distribution
- For model fitting using Metropolis-Hasting - It uses the following functions
  - For log-likelihood
  - For log prior
  - For log posterior

Analysis for India

Here the Data for India is considered from the date 4_{th} of March and till 1_{st} of May. The Lockdown started from 23_{rd} of March.

Visualization



From 7_{th} April onwards there is a change in the graph. It indicates that the increase in the cases is happening more or less in same rate.

Model

Here our response variable, y is difference of log values of total confirmed cases and Predictors are Time, Time$\hat{2}$ and Lockdown.

The linear model we consider is,

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$$

Here α is the intercept and β_1, β_2, β_3 are the co-efficients corresponding to the predictors. ϵ is the residual

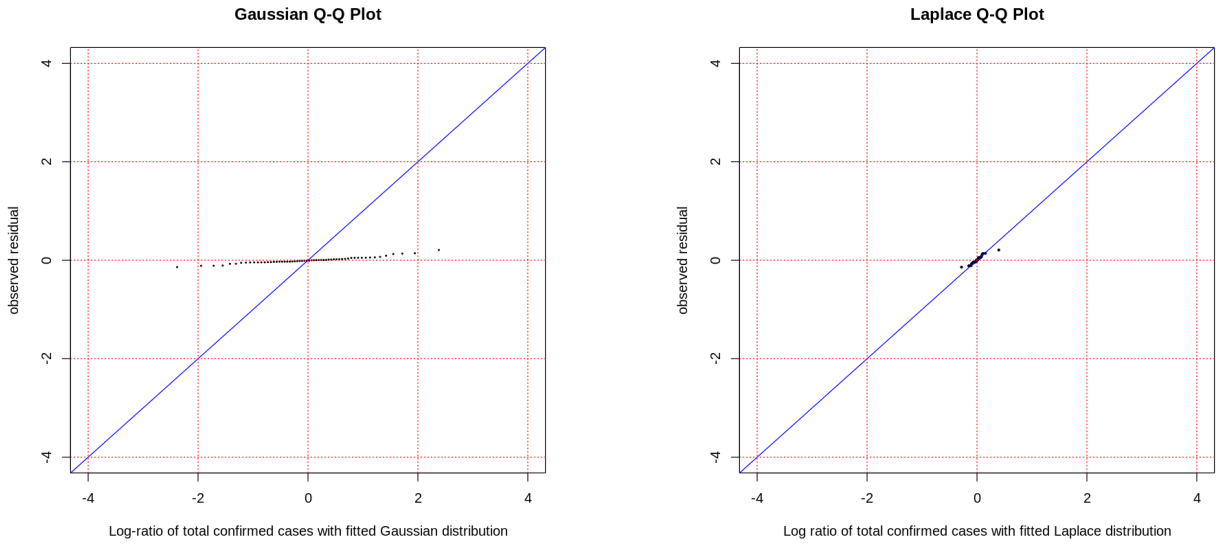
Linear Model Result(OLS)

```
Call:
lm(formula = ln_rt ~ Time + I(Time^2) + lock_down, data = India_data)

Residuals:
      Min        1Q    Median        3Q       Max
-0.139127 -0.036182 -0.009149  0.026402  0.208187

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.930e-02  3.066e-02   2.913  0.00520 **
Time         7.435e-03  3.115e-03   2.387  0.02054 *
I(Time^2)   -1.378e-04  4.166e-05  -3.307  0.00168 **
lock_down   -3.983e-02  3.870e-02  -1.029  0.30801
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.06466 on 54 degrees of freedom
Multiple R-squared:  0.2933,    Adjusted R-squared:  0.254
F-statistic:  7.47 on 3 and 54 DF,  p-value: 0.0002849
```

From the p-values we can conclude that Time and $Time^2$ has significant effect on response variable(log ratios) but p-value corresponding to lockdown variable is greater than 0.05 which means that lockdown doesn't have significant effect on response variable. From the R-square value we can also conclude that the fitting is not good. This analysis indicates that lockdown may reduce the disease progression - but it is not statistically significant.

Distribution of Residual

- **Shapiro-Wilk Test for Normality :** p-value - 0.0205
- **Q-Q plot :**



From Shapiro-Wilk test and looking at the Gaussian Q-Q plot, it is clear that the residual doesn't follow normal distribution. From the Laplacian Q-Q plot we can say that residuals follow laplace distribution.

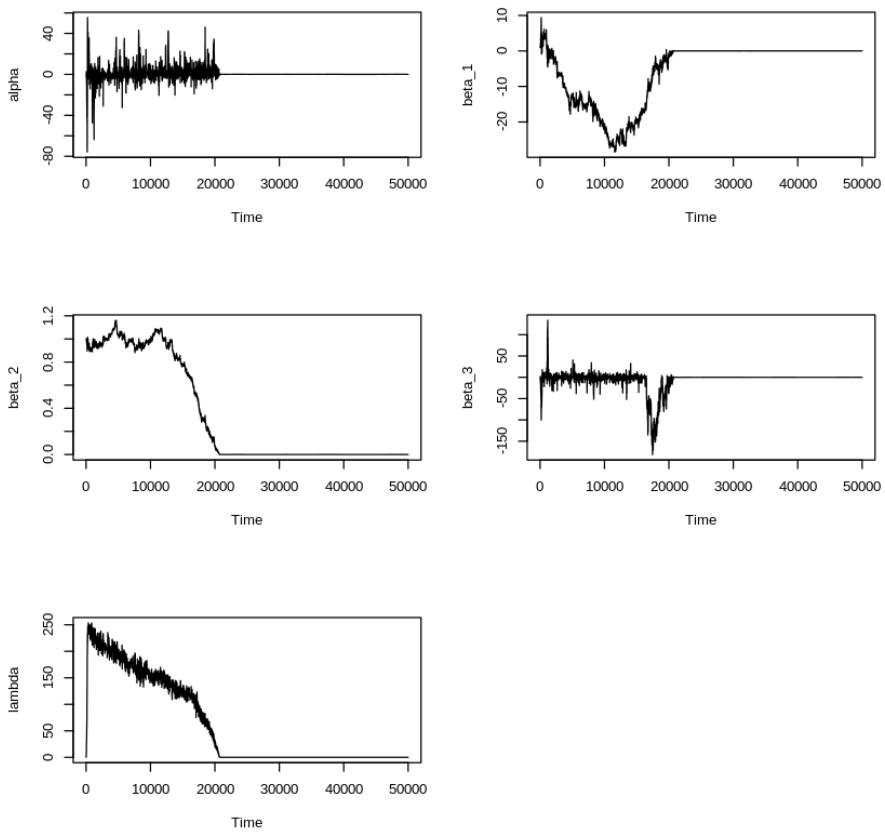
Bayesian Analysis

We have seen that the residuals doesn't follow Normal Distribution. They follow Laplace distribution with scale parameter λ and location parameter 0. So we can not take the conjugate prior Normal-Inverse Gamma.

Prior Distribution: We assume that α, β_1, β_2, β_3 follow Cauchy distribution with scale parameter 1 and location parameter 0 and λ follows Gamma Distribution.

Here we won't get the posterior easily. So we will use Metropolis-Hasting to get the posterior distribution of the parameters.

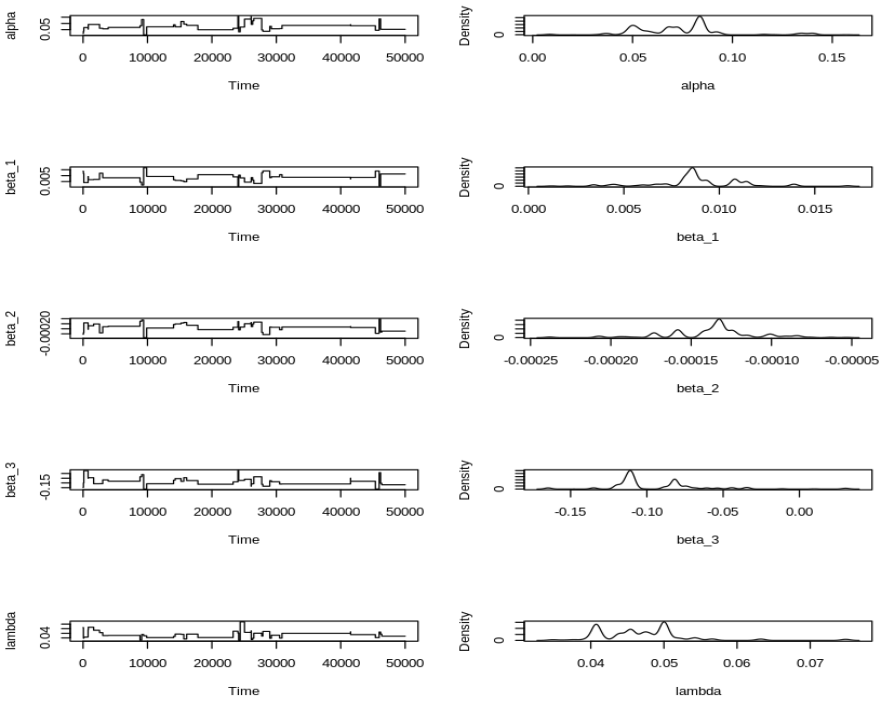
Trace Plot- Decide Burn-in period



The trace plot shows the sampled values of a parameter over time.This plot helps you to judge how quickly the MCMC procedure converges in distribution—that is, how quickly it forgets its starting values. In this case, it looks like there was a burn-in of about 21000 iterations, after which the MCMC sampler seems to mix well. So we now plot taking a burn-in of about 21000 iterations and conclude our result from the summary statistics of the posterior.

Trace Plot and Plot of Posterior Distribution

(Considering Burn-in Period)



Posterior Summary

Here we check the mean, median, s.d and 95% credible interval for the parameters. Based on these we conclude the effect of the predictors on response variable.

	alpha	beta_1	beta_2	beta_3	lambda
median	0.0728	0.0086	-1e-04	-0.1089	0.0472
mean	0.0738	0.0088	-1e-04	-0.0946	0.0469
sd	0.0230	0.0024	0e+00	0.0307	0.0059
2.5%	0.0369	0.0034	-2e-04	-0.1349	0.0393
97.5%	0.1401	0.0139	-1e-04	-0.0113	0.0633

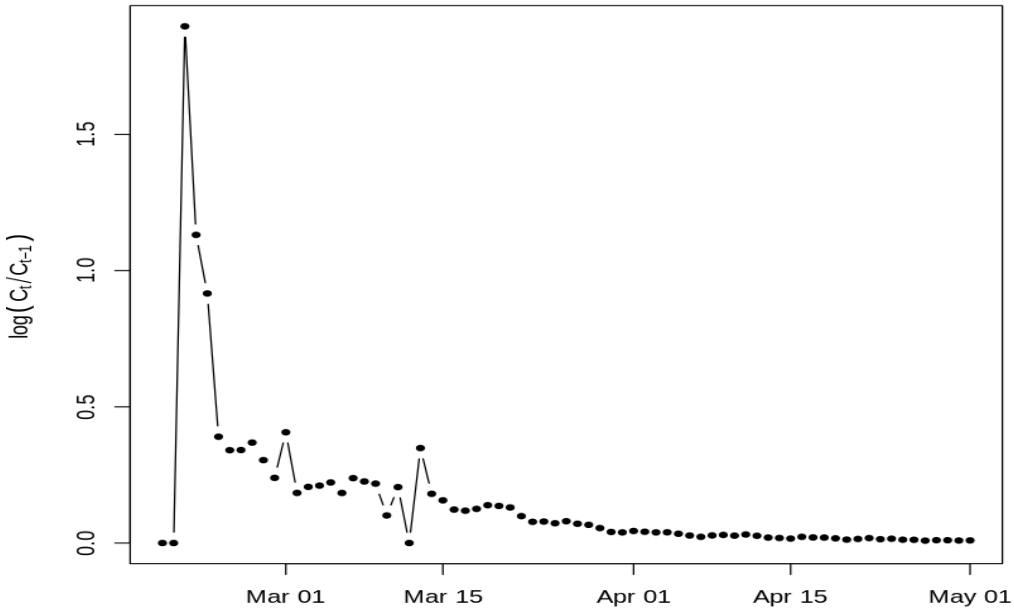
Interpretation

From the summary statistics of posterior of the parameters we can conclude whether a variable has a significant effect on the response variable or not. Here we can take the mean of the parameters as our estimates and can have a regression line. Based on the 95% credible interval from the summary statistics,we can say that all the x_1, x_2,x_3(Time, $Time^2$,lockdown) have significant effect on y as the the symmetric posterior quantile based credible intervals for the co-efficients exclude zero. Though from the summary it is clear that the effect of x's on y is not much. Co-efficient corresponding to lockdown comes out to be negative which does mean that when the lockdown value changes from 0 to 1, the y value decreases. So lockdown has a mild effect on total confirmed cases. So this says lockdown may reduce the disease progression. The result is kind of same as that we got by fitting the linear model(OLS) in terms of effect of the x's on y which also concludes that lockdown may reduce the disease progression though it was not statistically significant.

Analysis for Italy

Here the Data for Italy is considered from the date 18_{th} of February and till 1_{st} of May. The Lockdown started from 9_{th} of March.

Visualization



From 15_{th} March onwards there is a change in the graph. It indicates that the increase in the cases is happening more or less in same rate.

Model

Here our response variable, y is difference of log values of total confirmed cases and Predictors are Time, Time$\hat{2}$ and Lockdown.

The linear model we consider is,

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$$

Here α is the intercept and β_1, β_2, β_3 are the co-efficients corresponding to the predictors. ϵ is the residual

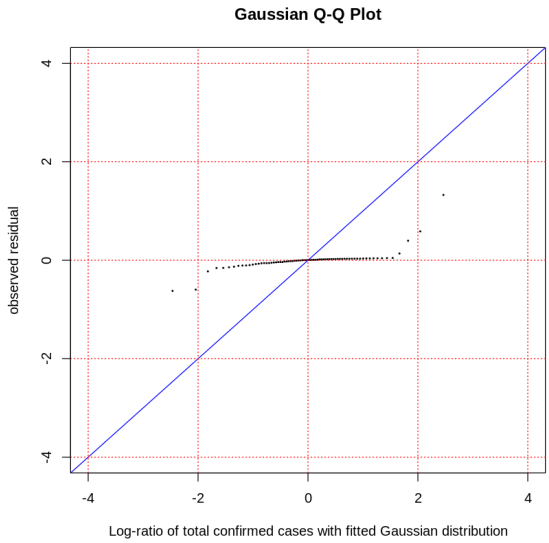
Linear Model Result(OLS)

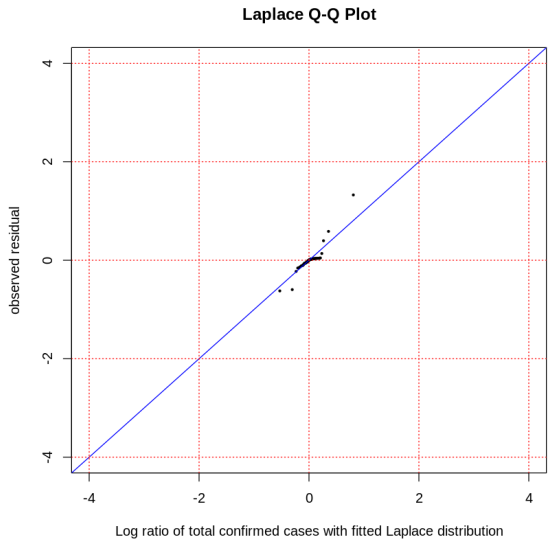
```
Call:
lm(formula = ln_rt ~ Time + I(Time^2) + lock_down, data = Italy_data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.62269 -0.05094  0.00514  0.02979  1.32670

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.496e-01  8.938e-02   7.268 4.34e-10 ***
Time        -2.714e-02  8.586e-03  -3.161  0.00233 **
I(Time^2)    2.521e-04  9.317e-05   2.706  0.00858 **
lock_down    6.809e-02  1.297e-01   0.525  0.60134
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2179 on 69 degrees of freedom
Multiple R-squared:  0.4159,    Adjusted R-squared:  0.3905
F-statistic: 16.38 on 3 and 69 DF,  p-value: 3.866e-08
```

From the p-values we can conclude that Time and $Time^2$ has effect on response variable(log-ratios) but p-value corresponding to lockdown variable is much greater than 0.05 which means that lockdown doesn't have significant effect on response variable. From the R-square value we can also conclude that the fitting is not good.

Distribution of Residual

- **Shapiro-Wilk Test for Normality :** p-value - 1.618e-13
- **Q-Q plot :**





From Shapiro-wilk test and looking at the Gaussian Q-Q plot, it is clear that the residual doesn't follow normal distribution. From the Laplacian Q-Q plot we can say that residuals follow laplace distribution.

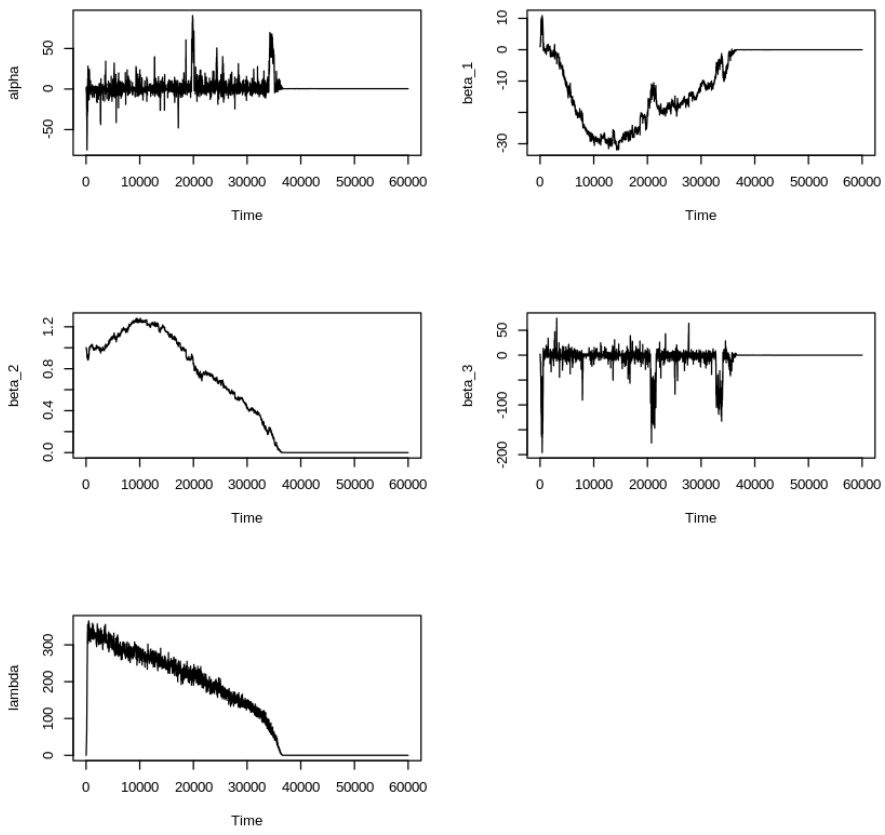
Bayesian Analysis

We have seen that the residuals doesn't follow Normal Distribution. They follow Laplace distribution with scale parameter λ and location parameter 0. So we can not take the conjugate prior Normal-Inverse Gamma.

Prior Distribution: We assume that α, β_1, β_2, β_3 follow Cauchy distribution with scale parameter 1 and location parameter 0 and λ follows Gamma Distribution.

Here we won't get the posterior easily. So we will use Metropolis-Hasting to get the posterior distribution of the parameters.

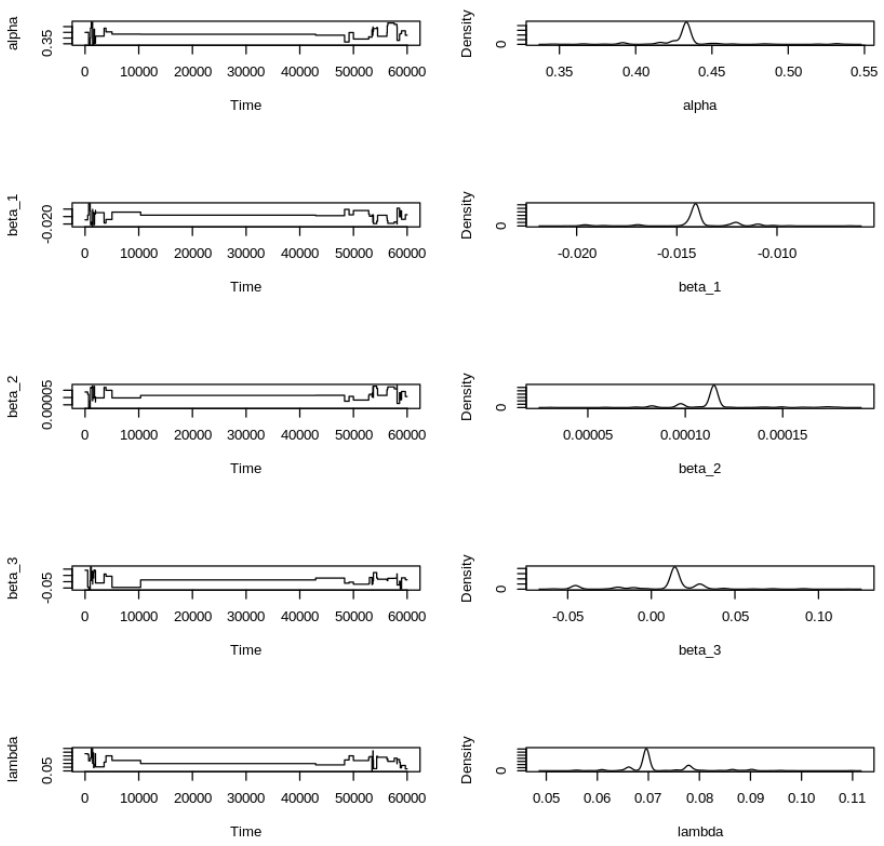
Trace Plot- Decide Burn-in period



The trace plot shows the sampled values of a parameter over time.This plot helps you to judge how quickly the MCMC procedure converges in distribution—that is, how quickly it forgets its starting values. In this case, it looks like there was a burn-in of about 38000 iterations, after which the MCMC sampler seems to mix well. So we now plot taking a burn-in of about 38000 iterations and conclude our result from the summary statistics of the posterior.

Trace Plot and Plot of Posterior Distribution

(Considering Burn-in Period)



Posterior Summary

Here we check the mean, median, s.d and 95% credible interval for the parameters. Based on these we conclude the effect of the predictors on response variable.

	alpha	beta_1	beta_2	beta_3	lambda
median	0.4331	-0.0141	1e-04	0.0140	0.0696
mean	0.4326	-0.0140	1e-04	0.0097	0.0724
sd	0.0263	0.0019	0e+00	0.0256	0.0077
2.5%	0.3657	-0.0196	1e-04	-0.0454	0.0609
97.5%	0.5196	-0.0102	2e-04	0.0724	0.0903

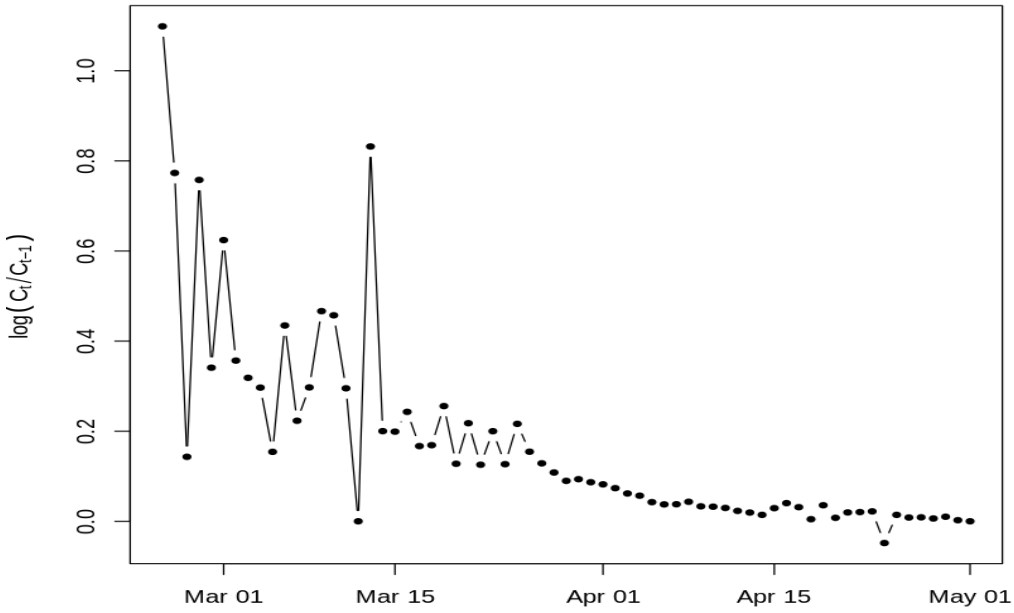
Interpretation

From the summary statistics of posterior of the parameters we can conclude whether a variable has a significant effect on the response variable or not. Here we can take the mean of the parameters as our estimates and can have a regression line. Based on the 95% credible interval from the summary statistics,we can say that the x_1, x_2(Time, $Time^2$) have effect on response variable as the the symmetric posterior quantile based credible intervals for the co-efficients exclude zero. From the summary, looking at the values of the co-efficients it is clear that the effect of x's on y is not much. There is a mild effect of the predictors on response. But for lockdown the symmetric posterior quantile based credible interval includes zero and the interval ranges from -0.04 to 0.07. It may have a mild effect on response but based on these result we conclude that the effect of lockdown is not significant.

Analysis for Spain

Here the Data for Italy is considered from the date 24_{th} of February and till 1_{st} of May. The Lockdown started from 13_{th} of March.

Visualization



From 25_{th} March onwards there is a change in the graph. It indicates that the increase in the cases is happening more or less in same rate.

Model

Here our response variable, y is difference of log values of total confirmed cases and Predictors are Time, Time$\hat{2}$ and Lockdown.

The linear model we consider is,

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$$

Here α is the intercept and β_1, β_2, β_3 are the co-efficients corresponding to the predictors. ϵ is the residual

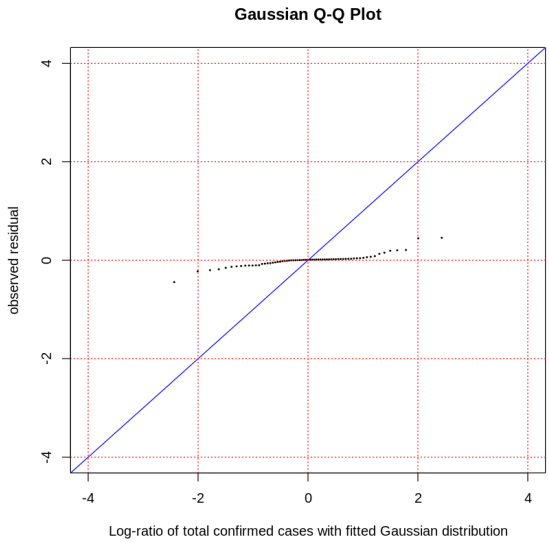
Linear Model Result(OLS)

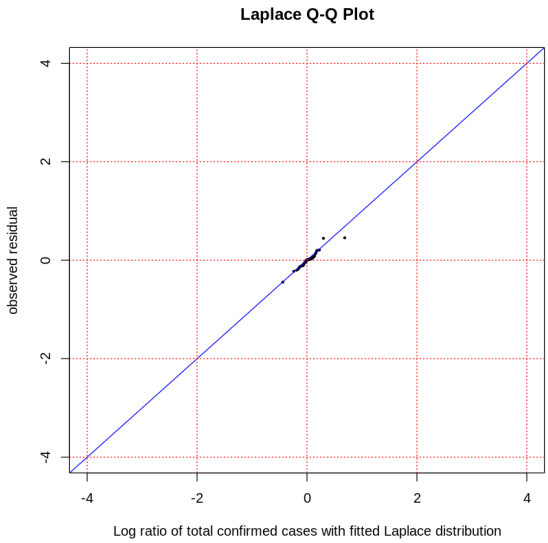
```
Call:
lm(formula = ln_rt ~ Time + I(Time^2) + lock_down, data = Spain_data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.44413 -0.05410  0.01118  0.02669  0.45631

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.866e-01  5.570e-02  12.326  < 2e-16 ***
Time        -3.409e-02  5.885e-03  -5.792 2.39e-07 ***
I(Time^2)    3.255e-04  6.971e-05   4.669 1.63e-05 ***
lock_down    1.971e-01  8.153e-02   2.418   0.0185 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1305 on 63 degrees of freedom
Multiple R-squared:  0.6785,    Adjusted R-squared:  0.6632
F-statistic: 44.31 on 3 and 63 DF,  p-value: 1.595e-15
```

From the p-values we can conclude that Time, $Time^2$ and lockdown all have effect on response variable(log-ratios) as the p-values corresponding to the variables are less than 0.05 . From the R-square value we can also conclude that the fitting is moderately good.

Distribution of Residual

- **Shapiro-Wilk Test for Normality :** p-value - 1.246e-06
- **Q-Q plot :**





From Shapiro-wilk test and looking at the Gaussian Q-Q plot, it is clear that the residual doesn't follow normal distribution. From the Laplacian Q-Q plot we can say that residuals follow laplace distribution.

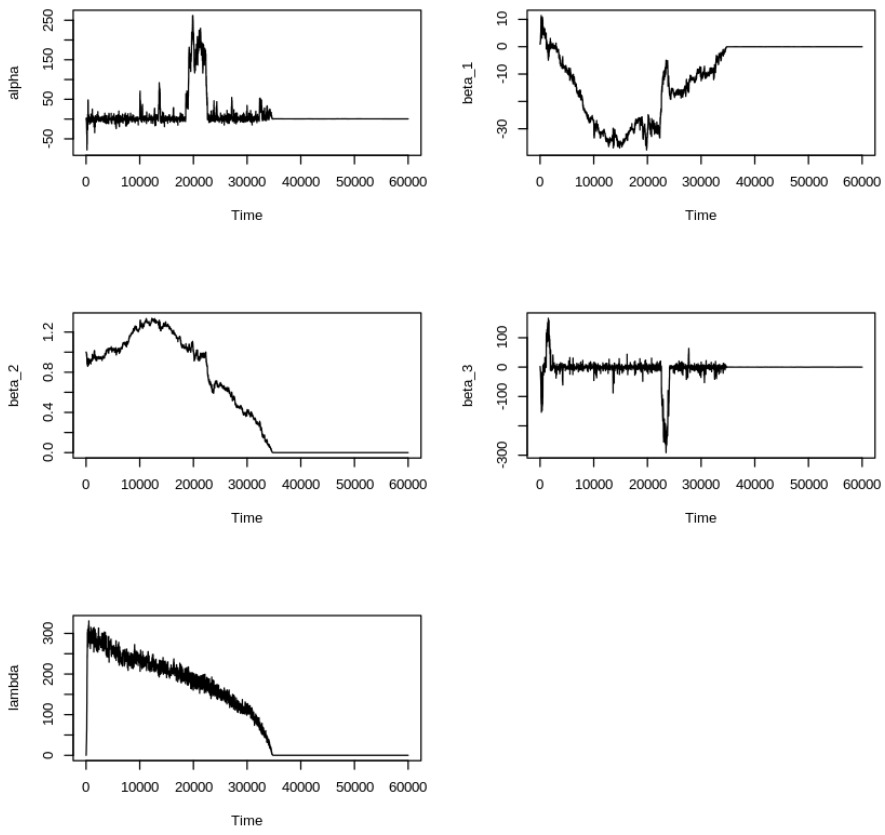
Bayesian Analysis

We have seen that the residuals doesn't follow Normal Distribution. They follow Laplace distribution with scale parameter λ and location parameter 0. So we can not take the conjugate prior Normal-Inverse Gamma.

Prior Distribution: We assume that α, β_1, β_2, β_3 follow Cauchy distribution with scale parameter 1 and location parameter 0 and λ follows Gamma Distribution.

Here we won't get the posterior easily. So we will use Metropolis-Hasting to get the posterior distribution of the parameters.

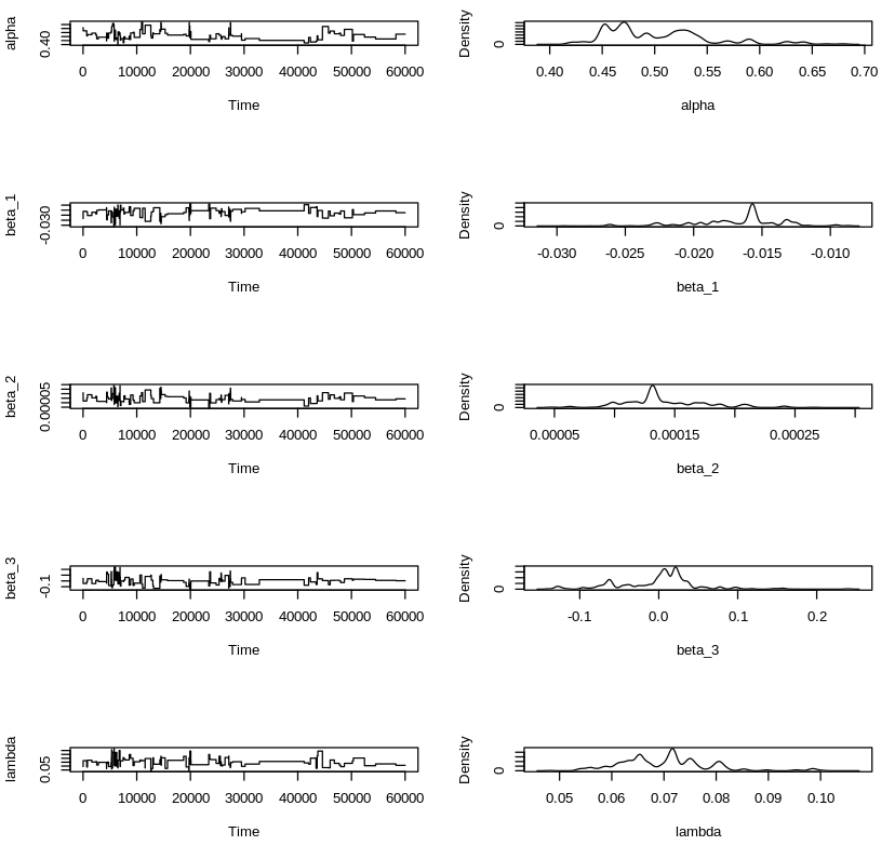
Trace Plot- Decide Burn-in period



The trace plot shows the sampled values of a parameter over time.This plot helps you to judge how quickly the MCMC procedure converges in distribution—that is, how quickly it forgets its starting values. In this case, it looks like there was a burn-in of about 36000 iterations, after which the MCMC sampler seems to mix well. So we now plot taking a burn-in of about 36000 iterations and conclude our result from the summary statistics of the posterior.

Trace Plot and Plot of Posterior Distribution

(Considering Burn-in Period)



Posterior Summary

Here we check the mean, median, s.d and 95% credible interval for the parameters. Based on these we conclude the effect of the predictors on response variable.

	alpha	beta_1	beta_2	beta_3	lambda
median	0.4331	-0.0141	1e-04	0.0140	0.0696
mean	0.4326	-0.0140	1e-04	0.0097	0.0724
sd	0.0263	0.0019	0e+00	0.0256	0.0077
2.5%	0.3657	-0.0196	1e-04	-0.0454	0.0609
97.5%	0.5196	-0.0102	2e-04	0.0724	0.0903

Interpretation

From the summary statistics of posterior of the parameters we can conclude whether a variable has a significant effect on the response variable or not. Here we can take the mean of the parameters as our estimates and can have a regression line. Based on the 95% credible interval from the summary statistics,we can say that the x_1, x_2(Time, $Time^2$) have effect on response variable as the the symmetric posterior quantile based credible intervals for the co-efficients exclude zero. From the summary, looking at the values of the co-efficients it is clear that the effect of x's on y is not much. There is a mild effect of the predictors on response. But for lockdown the symmetric posterior quantile based credible interval includes zero and the interval ranges from -0.1 to 0.1. It may have a mild effect on response but based on these result we conclude that the effect of lockdown is not significant.

Comments

Here we have tried to do a simple Bayesian analysis considering a linear model. There are lot of things we didn't consider here. We can not just simply capture the effect of lockdown by this analysis. There are persons who were infected earlier before lockdown but tested positive after lockdown started. Heree we have not considered a period of 14 days after the lockdown. In these 14 days the persons who are tested positive mostly got infected before lockdown started. So this is not taken into consideration which is vital in concluding the significance of lockdown. Also the testing rate is also factor. Initially the testing rate was not that high due to a lot of factors. There are other variables also which has an effect on response. We didn't consider all of those. So we need to consider a lot of other factors to strongly conclude whether lockdown works or not.