Statistical Analysis of COVID cases in India

Subhranil Das¹ and Rashmi Kumari²
¹ Department of Electrical and Electronics Engineering, BIT Mesra
² Department of Computer Science Department, BIT Mesra
E-mail: subhranil.sdas2007u@gmail.com, rksingh210492@gmail.com

Abstract. The novel coronavirus SARS-COV-2 (COVID-19) has been considered as pandemic in the month of February as declared by World Health Organization. The cause of Coronavirus is due to group of RNA Viruses that mainly causes diseases in animals where this disease can be transferred to humans by droplets which ultimately results to death if it is not cured at early stages. But, in India, COVID cases has begun towards the end of February as the cases has been spiked in the month of March. In this paper, a complete analysis of Statistics for COVID cases has been demonstrated till the month of August. Using Statistical Measures, inferences are being drawn on the basis of COVID data on confirmed, recovered, active as well as death cases.

Keywords: COVID, Generalized Linear Models, Poisson Regression Model, RStudio

1. Introduction

The first signs of COVID-19 had started in Wuhan city in China in December 2019. Since this disease has been epidemic during initial months but it has become pandemic at the end of January 2020 with the spread of Corona Virus with more than 80 countries as declared by World Health Organization (WHO). WHO Emergency Committee has been declared global health emergency on 30th January 2020.

The Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) formerly known to be COVID-2019 has been emerged as a zoonotic agent that has been appeared as in the month of December 2019 which causes Novel COVID-19[1]. Hence, it has been resulted in pathogen which leads to critical cases in some conditions that requires a special assistance room in hospitals such as Intensive Care Units (ICU) for most of the cases[2][3][4][5][6][7]. The SARS-CoV-2, taxonomically, is currently part of the species of the SARS-related coronaviruses that belong to the subgenus Sarbecovirus. Together with the subgeneruses Embecovirus, Hibecovirus, Merbecovirus, and Nobecovirus, that are part of the genus Betacoronavirus (order Nidovirales; suborder Coronavirinae; family Coronaviridae; subfamily Coronavirinae) [8][9][10][11][12][13].

2. Methodology

In order to characterize the COVID-19 analysis of data, a methodology has been proposed for getting the acquired data from the R Software Version 3.6.3. Here, statistical measures such as Mean, Median, The statistical measures such as Mean, Median, Mode, Standard Deviation, Global Average etc. have been applied over confirmed cases, recovered cases and death cases. For finding the relationship between various cases, these statistical tools provide a complete information in region wise in India.
In this paper, Poisson Regression models have been utilized for modelling where counts of outcomes become a necessary criterion. There are two important outcomes of this model are Count data and Rate data. Count data are those data where it is discrete in nature as well as magnitude is non-negative that occurs within a certain period of time. Rate data is being defined as the data where the speed at which occurs within a time interval. Poisson Regression helps in analyzing count data as well as rate data in order to X values as a explanatory variables effect on Y values as a response variable.

Generalized Linear Models(GLM) are those models in where variables as a response that follows the normal distribution which makes contradictory to Linear regression models. Hence, non-linear relationship is being developed between the response and predictor variables which is expressed by the given expression.

\[
y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + e_i
\]  

where \( i = 1, 2, \ldots, n \). The response variable \( y_i \) is being modeled by a series of linear functions of predictor variables with error value.

3. Simulation Results

The Experimental results are being executed on the personal computer which has configuration of 16GB RAM, NVIDIA GEFORCE GTX as a graphics card. The platform being used for analysis of data which is being done on software name RStudio Version 3.6.3 as a software package installed free from the Internet that is free is cost and easily available. While executing the data, there are some few packages to be installed such as covid19.analytics, covid19.data(case='aggregated') standard library functions.

Confirmed COVID cases are being displayed in Table 1 where data has been acquired from 188 countries. In this table, top 10 countries data are being depicted with global percentage data as well as last day change and data acquired on 19/08/2020. From this table, India stood Fourth among the confirmed cases with total of 673165 with last day change of 24850 cases in a single day.

| Country, Region Province, State | Totalts | Global Merc. Last DayChange | t-2 | t-3 | t-7 | t-14 | t-30 |
|--------------------------------|---------|-----------------------------|-----|-----|-----|------|------|
| US                             | 2889456 | 25.20                       | 45288| 52104| 54461| 39083| 26449| 23224|
| Brazil                         | 1577004 | 14.00                       | 37923| 42223| 48105| 40346| 15762| 10836|
| Russia                         | 675356  | 3.98                        | 60253| 67100| 67532| 67844| 77177| 8714|
| India                          | 67316   | 5.97                        | 24850| 22771| 20903| 19459| 14933| 9471|
| Peru                           | 299080  | 2.65                        | 3481 | 3595 | 3527 | 3430 | 3598 | 4202|
| Chile                          | 290847  | 2.59                        | 3758 | 3548 | 2498 | 4216 | 5602 | 4207|
| United Kingdom                 | 284900  | 2.53                        | 624  | 502  | 4395 | 622  | 2103 |      |
| Mexico                         | 252165  | 2.24                        | 694  | 6740 | 6742 | 4050 | 5342 | 4334|
| Spain                          | 250535  | 2.22                        | 0    | 442  | 444  | 201  | 234  | 214|
| Italy                          | 241419  | 2.14                        | 235  | 223  | 201  | 174  | 224  | 518|

Death cases in COVID are being displayed in Table 2 where cases count for India stands at eight position with total of 19268 deaths with last day with 613 deaths on the data collected on 19th August 2020. Also, data has been collected with reference to 2 days prior, 3 days before, 7 days
before, 14 days prior and 30 days before. It has been observed that there has been gradual increase in the number of cases starting from last 30 days.

### Table 2. Death COVID cases

| Country, Region, Province, State | Totals | % | LastDayChange | T-2 | T-3 | T-7 | T-14 | T-30 |
|---------------------------------|--------|---|---------------|-----|-----|-----|------|------|
| 1 US                            | 129676 | 4.57 | 242 | 631 | 698 | 240 | 255 | 970 |
| 2 Brazil                        | 64265  | 4.08 | 1091 | 1290 | 1525 | 552 | 615 | 1005 |
| 3 United Kingdom                | 441985 | 15.51 | 67 | 136 | 89 | 36 | 43 | 357 |
| 4 Italy                         | 34854  | 14.44 | 21 | 15 | 30 | 22 | 24 | 85 |
| 5 Mexico                        | 30366  | 12.04 | 523 | 654 | 679 | 267 | 1044 | 625 |
| 6 France                        | 27812  | 15.25 | 0 | 18 | 14 | 0 | 6 | 46 |
| 7 Spain                         | 28385  | 11.33 | 0 | 17 | 4 | 2 | 1 | 1 |
| 8 India                         | 19268  | 2.86 | 613 | 442 | 379 | 380 | 445 | 286 |
| 9 Iran                          | 11408  | 4.80 | 148 | 154 | 148 | 144 | 116 | 63 |
| 10 Peru                         | 10412  | 3.48 | 186 | 181 | 185 | 182 | 184 | 131 |

Recoverd COVID cases has been illustrated in Table 3 has been collected from 188 countries from 68 cities and 253 geographical locations. In this table, India has named 4th in the list of 10 countries with recovered 409083 cases with 14856 cases with previous day, 20032 cases recovered with 4 days prior, 14335 cases with 2 days before whereas 9468 cases recovered 14 days before as well as 4783 cases recovered from 1 month before. From this table, it suggests that the number of recovered cases has been steadily increased during the data acquired from 19th August.

### Table 3. Recovered COVID cases

| Country, Region, Province, State | Totals | LastDayChange | T-2 | T-3 | T-7 | T-14 | T-30 |
|---------------------------------|--------|---------------|-----|-----|-----|------|------|
| 1 Brazil                        | 990731 | 6138 | 6923 | 140090 | 18303 | 11339 | 11977 |
| 2 US                            | 894325 | 103921 | 8434 | 51976 | 5856 | 4673 | 6704 |
| 3 Russia                        | 446127 | 8972 | 8879 | 6041 | 5733 | 5118 | 8040 |
| 4 India                         | 409083 | 14856 | 14335 | 20032 | 12010 | 9468 | 4783 |
| 5 Chile                         | 257511 | 4108 | 4096 | 3804 | 4153 | 3960 | 3727 |
| 6 Iran                          | 198949 | 2503 | 2348 | 2611 | 2649 | 2207 | 2256 |
| 7 Mexico                        | 195724 | 6379 | 5588 | 5231 | 3923 | 3641 | 3983 |
| 8 Italy                         | 191944 | 477 | 384 | 366 | 307 | 440 | 1866 |
| 9 Peru                          | 189621 | 3769 | 3755 | 3852 | 3974 | 1050 | 2986 |
| 10 Germany                      | 181000 | 700 | 500 | 700 | 139 | 131 | 371 |
In this table, it has been observed that overall summary has been summarized in form of confirmed cases of worldwide totals, total deaths of 780908 with 3.53% deaths and recovered cases of 14116451. This table also illustrates that worldwide average of confirmed cases of 83221.63, deaths of 2935.74 cases and 55796.25 cases of recovered cases with 67.05%. Also, table illustrates that the number of confirmed cases is 438206.4, total deaths of 14100 with 3.22% and recovered cases of 255971.77 with 58.41% in terms of standard deviation. This data has been acquired from JHU/CCSE repository dated 19th August through means of RStudio platform (Version 3.6.3).

Table 4. Overall Summary

| Time Series Worldwide TOTS | ts-confirmed | ts-deaths | ts-recovered | 3.53% | 63.77% |
|----------------------------|--------------|-----------|--------------|-------|--------|
| 22136954                   | 780908       | 14116451  | 3.53%        | 63.77%|

| Time Series Worldwide AVGS | ts-confirmed | ts-deaths | ts-recovered | 3.53% | 67.05% |
|---------------------------|--------------|-----------|--------------|-------|--------|
| 83221.63                  | 2935.74      | 55796.25  | 3.53%        | 67.05%|

| Time Series Worldwide SDS  | ts-confirmed | ts-deaths | ts-recovered | 3.22% | 58.41% |
|----------------------------|--------------|-----------|--------------|-------|--------|
| 438206.4                   | 14100        | 255971.77 | 3.22%        | 58.41%|

In India, the first case of COVID has been detected in the month of February. By using GLM Poisson model, the number of cases has been increased from 0 to 15 by using log from zero day to up to 200 days. The calculated coefficients of GLM Poisson model are 7.1666 and 0.037 which is marked by red colour line as shown in figure1 where the coefficients calculated from exponential model are 0.865 and 0.077 highlighted by blue line. From this figure, it has been clearly understood that the rising lines of GLM and exponential model has steady characteristics from day 0 to till now.

In this second figure, the cases of COVID started prominently on 21st April and cases rises upto 500 in next month and goes exponentially 5,00,000 cases in the month of June and touches the mark of 25,00,000 cases in August making overall rank of 4th in all over the world. The calculated exponential GR and GLM Poisson GR is around 1.08 and 1.04 respectively.
Figure 1. GLM Poisson ratio and Exp GR ratio for cases in India
The number of changes of cases in each day are illustrated in the figure 2 starting from the month of February to August. It has been noticed that the number of daily changes has been very low in the sating month of February to gradually and steadily increases to 2000 in March. Then, in April month the number has been increased to 30,000 and in August month the daily changes have topped maximum of 64,000 till the recent data available to us. The line has been marked with red color and then marked with green and afterwards, it is marked with blue color (highest changes).

In figure 2(b), the growth rate of India is more and less constant starting from 24/01/2020 to till now 14/08/2020. In this figure indicates that the growth rate is around 5% during the initial phases during the month of January, February and March but the growth rate becomes steadier less than 5% during the next four months.
In this figure 3(a), red color signifies the number of deaths from zero to a certain number during February to the month of August. The violet color represents the confirmed cases where the cases is 2 in the month of February and exponentially rises in the cases during the month of June, July and August where the number reaches to 30,00,000. The blue color indicates the recovered cases where the number is very low during the months of February, April and May but the number reaches exponentially to 20,00,000 as it follows the same as confirmed cases during next two months. But, the number of active cases is steadily increases from zero in the month of April to 7,50,000 cases during the month of August.

In figure 3(b), all the lines such as red line, violet line, black line indicates the number of death cases, recovered cases and confirmed cases respectively. Here, in this figure it is being observed that during initial months such as February, March the lines are distorted manner but next five months the lines becomes more in prominent manner with clear demarcation of lines. The confirmed cases (14.6) is much higher than recovered cases(12.2) whereas the number of death cases is significantly lower at a value of 10 at the end of the month August.
4. Results and Discussion

The statistical analysis of data is being represented with GLM Poison’s Ratio as well as Exp Ratio as one of the powerful tools in statistics which helps in analyzing the results obtained from the graphs plotted above. In this work, using RStudio Software Version 3.6.3 data has been acquired through JHU/CCSE Repository where the number of confirmed cases, number of death cases, number of recovered cases as well as full summary of results have been shown in previous tables as mentioned above. Using the coefficients of GLM based model and exponential model, growth rates as well as the number of changes of each day can be interpreted and have been displayed in form of graphs where the time period have been taken as month of February to the month of August.

5. Conclusion

The results inferred from the data acquired from cases concerned with India has played a prominent analysis by using measures such as mean, median, standard deviation for providing whole summary of total cases in the world. It gives a clear picture of statistical analysis done through RStudio Environment.

Acknowledgment

Rashmi Kumari, PhD student as well as Co-author of this paper is very grateful to Subhranil Das, Main author of this paper for his endeverous and stupendou s effort for carrying out this research paper on COVID. As there is no funding from various agencies, he has made his own research on COVID cases in India as he was one of active cases while going through the phase of COVID. His valuable research has made a timely effort and bringing the latest happening of COVID around India. I am very thankful for my colleague for tremendous effort on his side to make the paper within the stipulated period of time.

References

[1] Bonilla-Aldana, D. K., Dhama, K., & Rodriguez-Morales, A. J. (2020). Revisiting the one health approach in the context of COVID-19: a look into the ecology of this emerging disease. *Adv Anim Vet Sci*, 8(3), 234-237.
[2] Zhu, N., Zhang, D., Wang, W., Li, X., Yang, B., Song, J., ... & Niu, P. (2020). A novel coronavirus from patients with pneumonia in China, 2019. *New England Journal of Medicine*.
[3] Chan, J. F. W., Yuan, S., Kok, K. H., To, K. K. W., Chu, H., Yang, J., ... & Tsioi, H. W. (2020). A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster. *The Lancet*, 395(10223), 514-523.
[4] Chen, N., Zhou, M., Dong, X., Qu, J., Gong, F., Han, Y., ... & Yu, T. (2020). Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. *The Lancet*, 395(10223), 507-513.
[5] Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., ... & Cheng, Z. (2020). Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *The lancet*, 395(10223), 497-506.
[6] Lancet, T. (2020). Emerging understandings of 2019-nCoV. *Lancet (London, England)*, 395(10221), 311.
[7] Bastola, A., Sah, R., Rodriguez-Morales, A. J., Lal, B. K., Jha, R., Ojha, H. C., ... & Morita, K. (2020). The first 2019 novel coronavirus case in Nepal. *The Lancet Infectious Diseases*, 20(3), 279-280.
[8] Bonilla-Aldana, D. K., Villamil-Gómez, W. E., Rabaan, A. A., & Rodríguez-Morales, A. J. (2020). Una nueva zoonosis viral de preocupación global. *Iatreia*, 33(2), 107-110.

[9] Millán-Oñate, J., Rodríguez-Morales, A. J., Camacho-Moreno, G., Mendoza-Ramírez, H., Rodríguez-Sabogal, I. A., & Álvarez-Moreno, C. (2020). A new emerging zoonotic virus of concern: the 2019 novel Coronavirus (SARS CoV-2). *Infectio*, 24(3), 187-192.

[10] Ksiazek, T. G., Erdman, D., Goldsmith, C. S., Zaki, S. R., Peret, T., Emery, S., ... & Rollin, P. E. (2003). A novel coronavirus associated with severe acute respiratory syndrome. *New England Journal of Medicine*, 348(20), 1953-1966.

[11] Li, Q., Guan, X., Wu, P., Wang, X., Zhou, L., Tong, Y., ... & Xing, X. (2020). Early transmission dynamics in Wuhan, China, of novel coronavirus–infected pneumonia. *New England Journal of Medicine*.

[12] Phan, L. T., Nguyen, T. V., Luong, Q. C., Nguyen, T. V., Nguyen, H. T., Le, H. Q., ... & Pham, Q. D. (2020). Importation and human-to-human transmission of a novel coronavirus in Vietnam. *New England Journal of Medicine*, 382(9), 872-874.

[13] Pongpirul, W. A., Pongpirul, K., Ratnarathon, A. C., & Prasitsirikul, W. (2020). Journey of a Thai taxi driver and novel coronavirus. *New England Journal of Medicine*, 382(11), 1067-1068.