
# Statistical Study on the analysis and Prediction of COVID-19 in India

- Course Code: MAT2001
- Course Title: Statistics for Engineers
- Semester: Winter 2020-21
- Slot: L43+L44
- Team member: S SURYA (20BCE1071)
- Course Faculty: Dr  JAGANATHAN B

## Abstract:

This study aims to show few visualisations to see the effect COVID-19 had along with the trends of the cases in India over the last 15 months. A prediction model based on Linear regression and Poisson regression model is presented.

The report also suggests various regression models for estimation of new COVID-19 cases based on various independent variables.

In December 2019, the novel and contagious virus belonging to the coronavirus COVID-19 outbreak started from the town of Wuhan, China. The virus also named as “severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)” being genetically close to the coronavirus that caused the SARS outbreak in 2003. The pandemic has become a major threat to the world and to India. In this report, various parameters has been plotted and various linear and Poisson regression models have been attempted to predict the new COVID 19 cases for each day.



## Key words:

Linear regression, Poisson regression, Predicted value, Adjusted R2.


## INTRODUCTION

Vaccination plays a vital role in eradicating and controlling the wide spread of diseases like smallpox, polio, measles, tetanus and leprosy throughout the world. It is also considered to be cost effective method in public health services and save millions of lives, especially the lives of children. Different vaccines could be suggested for any particular disease there by necessitating comparative vaccine trial to estimate the potency of vaccines. Numerous studies have been conducted in the estimation of vaccine efficacy using case-control and prospective cohort studies; and these studies offered varied information.

## DATA SOURCES AND DESCRIPTION

The data for this study is sourced from https://ourworldindata.org/. Data for “India” was taken up for this study.

The data is updated on a daily basis which is cumulative of all states in India. The data has 21 different data points per record. Some of the key data points that has been used for this study are Date, total_cases, new_cases, total_deaths, new_deaths, new_tests, total_tests, positive_rate tests_per_case, total_vaccinations , people_vaccinated, new_vaccinations). All Data Points are numeric except the Date field.

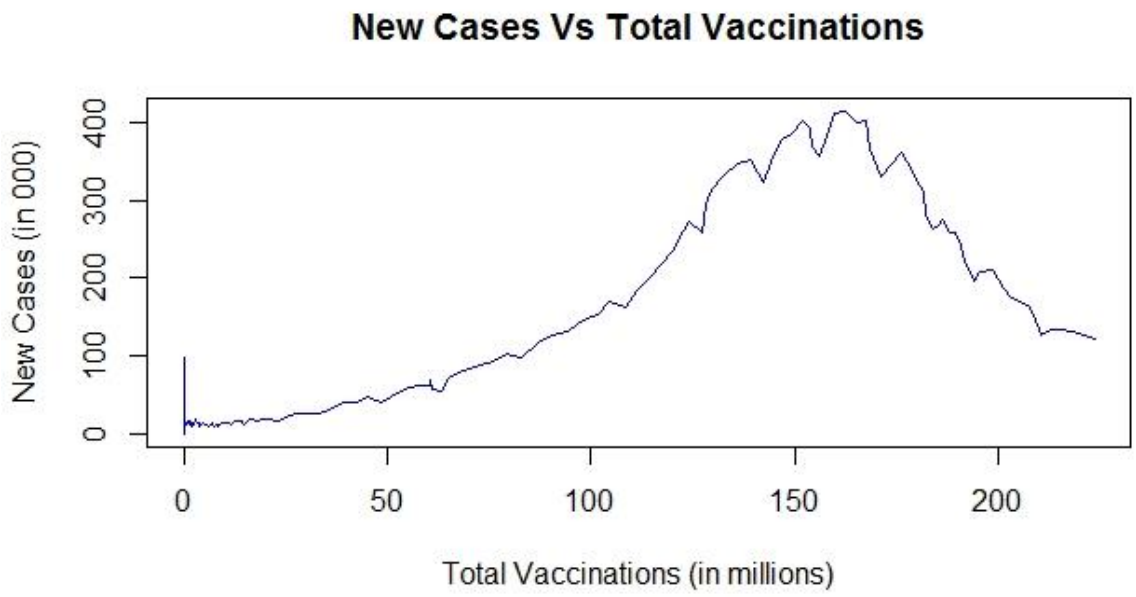
Data is updated daily and the data is available from 31st Jan 2020.

Additionally, data from the John Hopkins Github () was also researched

## VISUALIZATIONS:

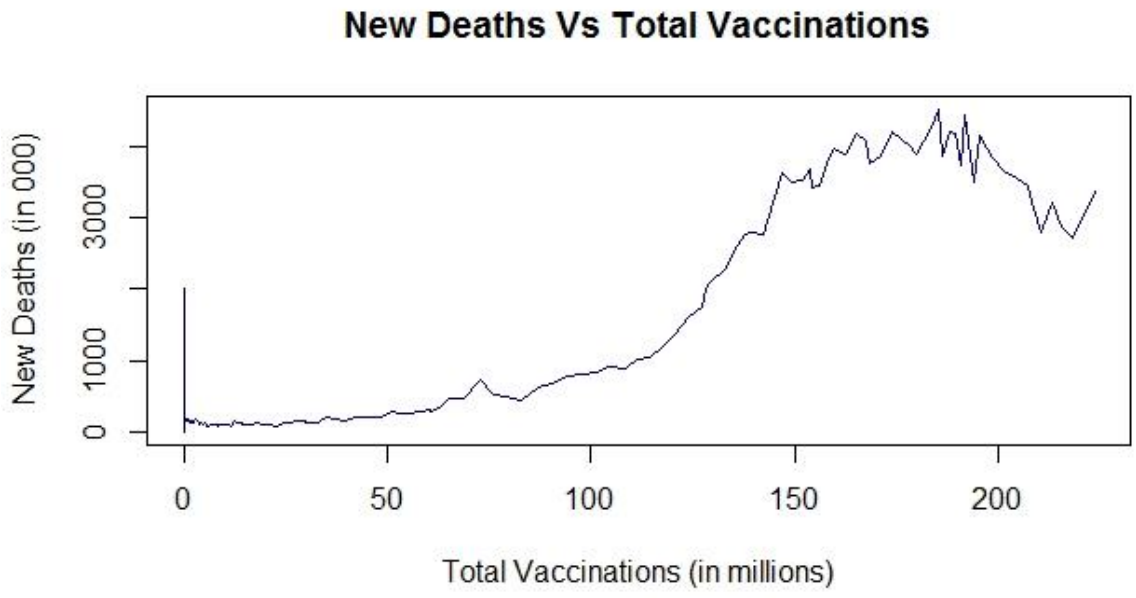
The data can be visualized with a few plots among different data points (parameters)

1. New Cases Vs Total Vaccinations



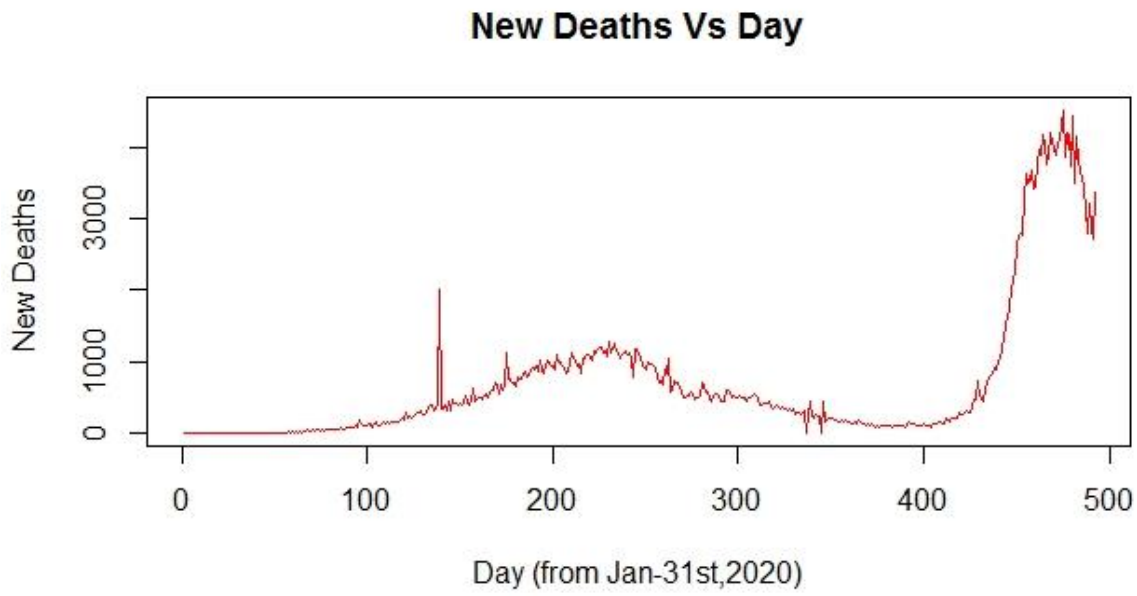

With increased vaccinations, the new cases started coming down after about 160 million vaccinations

2. New Deaths Vs Total Vaccinations

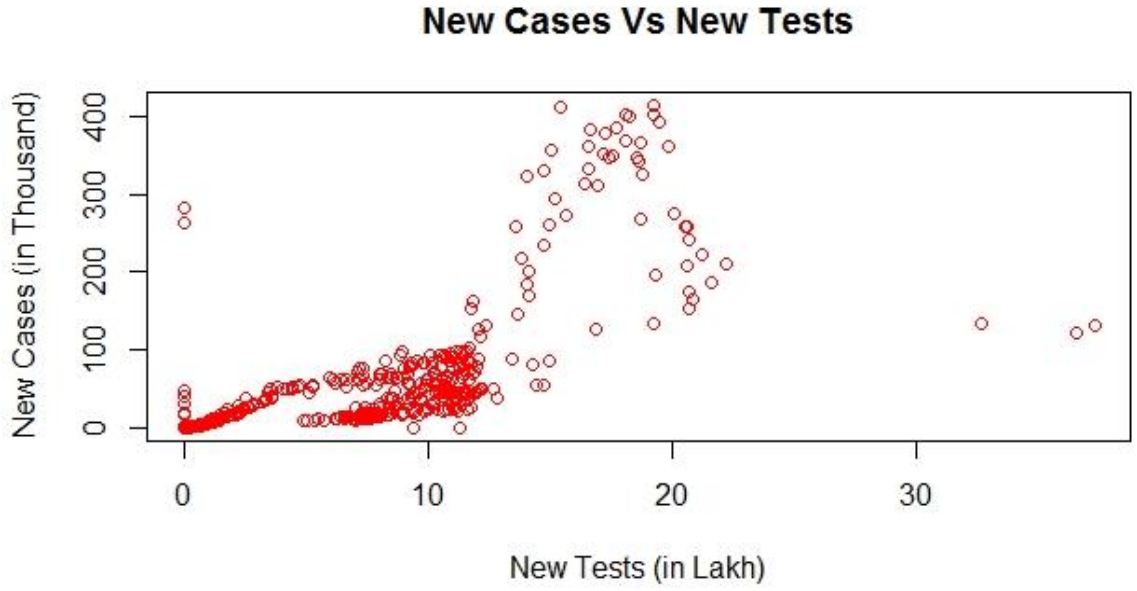


At around 175 million total vaccinations, the daily new Deaths peaked t around 3700 deaths.

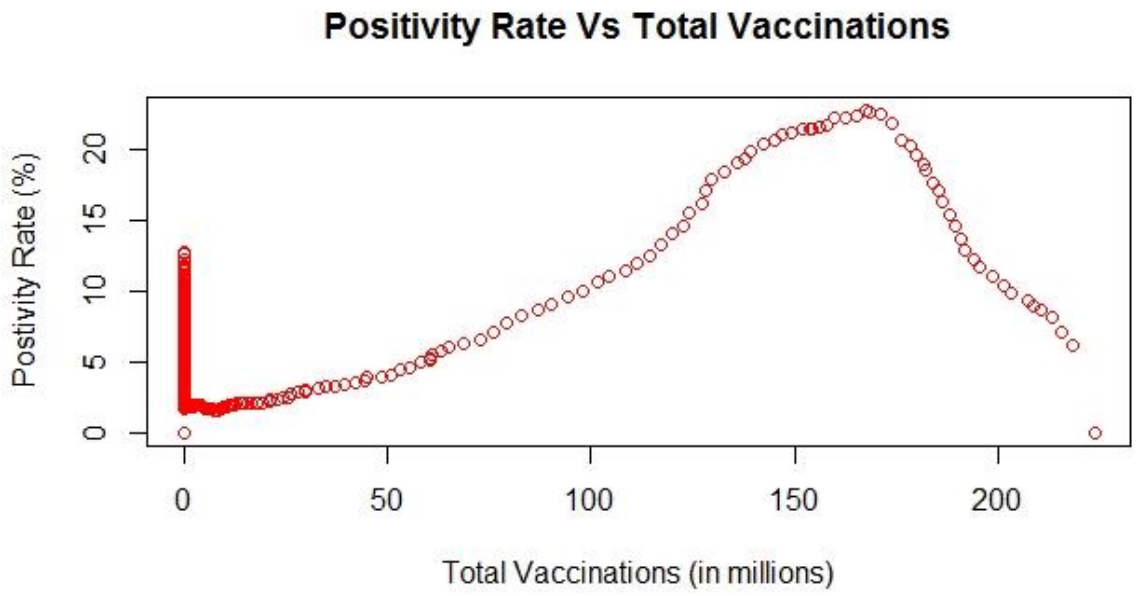
3. New Deaths Vs Cumulative Day




4. New Cases Indentified Vs New Tests that were done



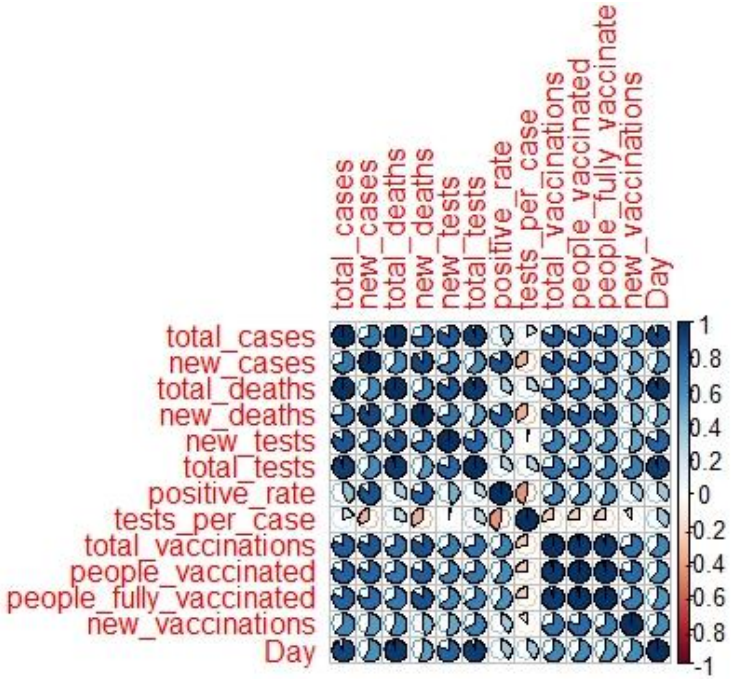
5. Positivity Rate Vs Total Vaccinations




With the increase in total vaccinations beyond 160 million, the positivity rate has also starting trending down.

## CORRELATIONS:

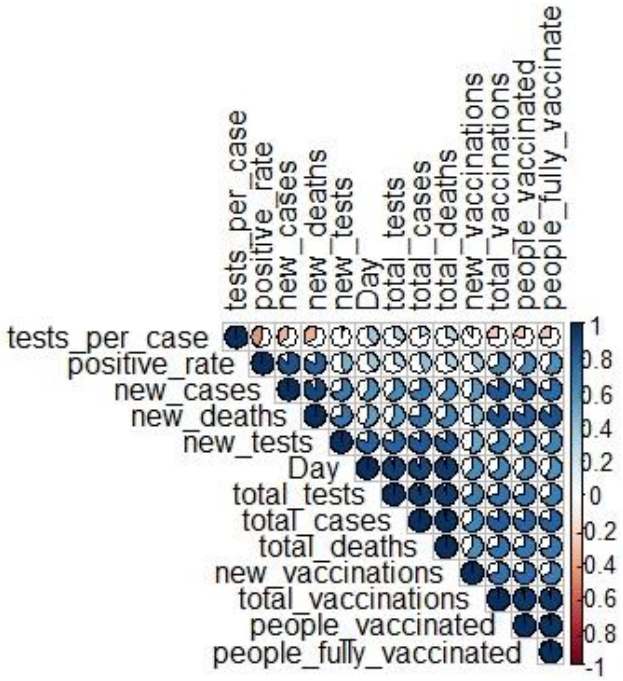
To identify the relationships between different data points, Correlation plot is done.




From the above plot, some of the higher correlated values are as below:

1. new_cases Vs new_deaths
2. new_cases Vs positive_rate
3. new_cases Vs total_vaccinations
4. new_cases Vs  new_vaccinations
5. new_deaths Vs total_vaccinations
6. new_deaths Vs positive_rate

Another View of the Correlation Plot is as below




These above correlations shall  be used while deriving the Regression Models.

# I. REGRESSION MODELS

## Linear Regression Model:

Linear Regression model was developed for multiple relations and the details of the Adjusted R2 and Residual Standard Error is shown in the table below:

| Sl # | Dependent Variable | Independent Variables | Residual Std Error | Adjusted R2 |
|---|---|---|---|---|
| 1 | New Deaths | Total Vaccinations | 502.8 | 0.7501 |
| 2 | New Deaths | Total Vaccinations, New Tests and Total Cases | 456 | 0.7945 |
| 3 | New Cases | Total Vaccinations | 47690 | 0.7089 |
| 4 | New Cases | Total Vaccinations, New Tests and Total Cases | 43600 | 0.7567 |
| 5 | New Cases | Total Vaccinations, New Tests, Positive Rate and Day | 27610 | 0.9025 |

| | | | | |
|---|---|---|---|---|
| 6 | New Cases | Total Vaccinations, New Tests, People Vaccinated ,Positive Rate and Day | 27530 | 0.903 |

From the above table, it can be derived that most of the variance for “New Cases” are available through the 4 independent variables (Total Vaccinations, New Tests, Positive Rate and Day). Higher R-squared values indicate that the data points are closer to the fitted values. While higher R-squared values are good, they don’t tell you how far the data points are from the regression line. R-squared is a percentage.

The standard error of the regression provides the absolute measure of the typical distance that the data points fall from the regression line

## Screenshots for Linear Models

**1.** New Deaths Vs Vaccinations, New Tests, Total Cases

```
> print(summary(relation2))

Call:
lm(formula = new_deaths ~ total_vaccinations + new_tests + total_cases,
    data = train_data_numeric)

Residuals:
     Min       1Q   Median       3Q      Max
-2126.91  -172.00   -46.49   235.77  2344.00

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       1.720e+02  3.426e+01   5.021 7.21e-07 ***
total_vaccinations 1.601e-05  6.855e-07  23.350  < 2e-16 ***
new_tests          6.582e-04  6.378e-05  10.320  < 2e-16 ***
total_cases       -4.477e-05  7.147e-06  -6.264 8.26e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 456 on 488 degrees of freedom
Multiple R-squared:  0.7957,    Adjusted R-squared:  0.7945
F-statistic: 633.7 on 3 and 488 DF,  p-value: < 2.2e-16

> |
```

**2.** New Cases Vs Total Vaccinations, New Tests, Positive Rate and Day

```
Call:
lm(formula = new_cases ~ total_vaccinations + new_tests + positive_rate +
    Day, data = train_data_numeric)

Residuals:
    Min      1Q  Median      3Q     Max
-126936  -14373    2862   10274   96408

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)        -3.466e+04  3.230e+03 -10.731  < 2e-16 ***
total_vaccinations  5.804e-04  3.802e-05  15.266  < 2e-16 ***
new_tests           2.864e-02  3.747e-03   7.643 1.14e-13 ***
positive_rate       9.523e+05  3.322e+04  28.670  < 2e-16 ***
Day                 4.996e+00  1.523e+01   0.328    0.743
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 27610 on 487 degrees of freedom
Multiple R-squared:  0.9032,    Adjusted R-squared:  0.9025
F-statistic:  1137 on 4 and 487 DF,  p-value: < 2.2e-16
```

## II. Poisson Regression Model:

Poisson Regression Model was tried with different combinations of input variables and the data is presented in the below table

| Sl # | Dependent Variable | Independent Variables | Null Deviance | Residual Deviance | AIC |
|---|---|---|---|---|---|
| 1 | New Cases | Total Vaccinations, New Tests and Total Cases | 45745162 | 6185803 | 6191146 |
| 2 | New Cases | Total Vaccinations, New Tests, Positive Rate and Day | 45745162 | 5012644 | 5017989 |
| 3 | New Cases | Total Vaccinations, New Tests, People Vaccinated ,Positive Rate and Day | 45745162 | 5011897 | 5017224 |

The null deviance shows how well the response variable is predicted by a model that includes only the intercept (grand mean).

Residual Deviance shows how well the response variable is predicted by the model that includes the independent variables also.

From the above table, it can be clearly seen that Residual Deviance decreased by a factor of nearly 10, thus indicating the importance of the weights for the independent variables.

The AIC (Akaike Information Criterion) Score provides a method for assessing the quality of the models through comparison of related models. It's based on the Deviance, but penalizes for making the model more complicated. Lower AIC scores are better, and AIC penalizes models that use more parameters. So if two models explain the same amount of variation, the one with fewer parameters will have a lower AIC score and will be the better-fit model.

**Screenshots of Poisson Model:**

1. Poisson Model (3 Independent Variables)

```
Call:
glm(formula = new_cases ~ total_vaccinations + new_tests + positive_rate,
    family = poisson(), data = train_data_numeric)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-442.53  -108.56   -40.39    49.10   634.84

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)        9.071e+00  5.689e-04 15944.3   <2e-16 ***
total_vaccinations -1.380e-09  5.606e-12  -246.1   <2e-16 ***
new_tests           8.854e-07  5.040e-10  1756.7   <2e-16 ***
positive_rate       1.209e+01  4.005e-03  3018.9   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 45745162  on 491  degrees of freedom
Residual deviance:  6185803  on 488  degrees of freedom
AIC: 6191146

Number of Fisher Scoring iterations: 5

>
```

2. Poisson Model (5 Independent Variables)

```
Call:
glm(formula = new_cases ~ total_vaccinations + new_tests + positive_rate +
    people_vaccinated + Day, family = poisson(), data = train_data_numeric)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-272.48   -97.02   -53.57    47.01   534.07

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)        8.340e+00  9.684e-04 8611.65   <2e-16 ***
total_vaccinations -4.683e-09  9.134e-12 -512.65   <2e-16 ***
new_tests           7.131e-07  5.522e-10 1291.25   <2e-16 ***
positive_rate       1.220e+01  4.185e-03 2915.03   <2e-16 ***
people_vaccinated  -2.143e-10  7.815e-12  -27.42   <2e-16 ***
Day                 3.439e-03  3.217e-06 1068.87   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 45745162  on 491  degrees of freedom
Residual deviance:  5011897  on 486  degrees of freedom
AIC: 5017244

Number of Fisher Scoring iterations: 5

>
```

## III Polynomial Regression Model:

Polynomial Regression model with different polynomial degrees for the independent variables were tried and the result is presented below:

| Sl # | Dependent Variable | Independent Variables | Residual Std Error | Adjusted R2 |
|---|---|---|---|---|
| 1 | New Cases | (Total Vaccinations)^2, New Tests, (Positive Rate)^2,People_vaccinated and Day | 14900 | 0.9716 |
| 2 | New Cases | (Total Vaccinations)^2, (New Tests)^2, (Positive Rate)^2,People_vaccinated and Day | 13790 | 0.9757 |
| 3 | New Cases | (Total Vaccinations)^3, (New Tests)^3, (Positive Rate)^3,People_vaccinated and Day | 12230 | 0.9809 |

With the polynomial models, there is an improvement in the Adjusted R2, as well as the Residual Error, in comparison to the Linear models.

**Screenshots for the Polynomial model output:**

```
> print(summary(relation8))

Call:
lm(formula = new_cases ~ poly(total_vaccinations, 2) + new_tests +
    poly(positive_rate, 2) + people_vaccinated + Day, data = train_data_numeric)

Residuals:
   Min     1Q Median     3Q    Max
-70741  -7401   1588   5483  71143

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                  4.151e+04  2.109e+03  19.683  < 2e-16 ***
poly(total_vaccinations, 2)1 5.732e+05  6.458e+04   8.877  < 2e-16 ***
poly(total_vaccinations, 2)2 -2.112e+05 1.815e+04 -11.639  < 2e-16 ***
new_tests                    5.041e-02  2.259e-03  22.311  < 2e-16 ***
poly(positive_rate, 2)1      1.151e+06  2.136e+04  53.881  < 2e-16 ***
poly(positive_rate, 2)2      5.055e+05  1.667e+04  30.318  < 2e-16 ***
people_vaccinated           -1.707e-04  6.272e-05  -2.722  0.00673 **
Day                         -6.426e+01  9.807e+00  -6.553 1.44e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14900 on 484 degrees of freedom
Multiple R-squared:  0.972,     Adjusted R-squared:  0.9716
F-statistic:  2400 on 7 and 484 DF,  p-value: < 2.2e-16

> |
```

```
> relation12<- lm(formula = new_cases~poly(total_vaccinations,3)+poly(new_tests,3)+
poly(positive_rate,3)+people_vaccinated+Day, data=train_data_numeric)
> print(summary(relation12))

Call:
lm(formula = new_cases ~ poly(total_vaccinations, 3) + poly(new_tests,
    3) + poly(positive_rate, 3) + people_vaccinated + Day, data = train_data_numer
c)

Residuals:
   Min     1Q Median     3Q    Max
-44103  -6476   -747   3352  63609

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                   6.806e+04  3.477e+03  19.574  < 2e-16 ***
poly(total_vaccinations, 3)1  6.191e+05  7.953e+04   7.785 4.33e-14 ***
poly(total_vaccinations, 3)2 -2.014e+05  1.598e+04 -12.603  < 2e-16 ***
poly(total_vaccinations, 3)3 -2.295e+05  2.510e+04  -9.144  < 2e-16 ***
poly(new_tests, 3)1           6.541e+05  2.584e+04  25.318  < 2e-16 ***
poly(new_tests, 3)2          -3.689e+04  2.849e+04  -1.295 0.196067
poly(new_tests, 3)3          -8.401e+04  2.497e+04  -3.364 0.000829 ***
poly(positive_rate, 3)1       9.051e+05  3.051e+04  29.664  < 2e-16 ***
poly(positive_rate, 3)2       3.738e+05  1.817e+04  20.572  < 2e-16 ***
poly(positive_rate, 3)3      -6.889e+03  1.424e+04  -0.484 0.628832
people_vaccinated            -7.809e-05  5.298e-05  -1.474 0.141189
Day                          -3.380e+01  1.289e+01  -2.622 0.009010 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12230 on 480 degrees of freedom
Multiple R-squared:  0.9813,    Adjusted R-squared:  0.9809
F-statistic:  2289 on 11 and 480 DF,  p-value: < 2.2e-16
```

## CONCLUSION

This study presented a Statistical study of the COVID-19 outbreak situation in India. The cases are rising very fast and they need aggressive control strategies. Broadly, 12 different models that fell into three different prediction methodologies were presented, to predict the new cases of COVID-19. Also different visualization graphs were presented which showed the pattern of COVID-19 proliferation in India. The Polynomial regression model based on third degree of independent variables provided the best Adjusted R2 (0.9809) and Residual Standard error (12230) for estimating the new cases.

# REFERENCES


1. Covid-19 data from ourworldInData.org (https://github.com/owid/covid-19-data/tree/master/public/data/)

2. An introduction to the Akaike information criterion (https://www.scribbr.com/statistics/akaike-information-criterion/)

3. Standard Error of the Regression vs. R-squared (https://statisticsbyjim.com/regression/standard-error-regression-vs-r-squared/)

4. John Hopkins University: Novel Coronavirus (COVID-19) Cases, provided by JHU CSSE Accessed from https://github.com/CSSEGISandData/COVID-19


.