Mathematical Modeling of COVID-19 Data in Nepal

RESEARCH QUESTION

Is there a positive relationship between the recovery amount of COVID-19 virus in terms of the number of newly confirmed cases?

INTRODUCTION

Severe Acute Respiratory Coronavirus-2(SARS-CoV-2), a novel β-coronavirus is the pathogen responsible for the coronavirus disease that we know. The novel coronavirus has spread across the globe, and at the time of writing this paper, there are 56,544,541 coronavirus cases with 1,353,918 attributable deaths around the world. From the beginning of the epidemic modeling, the description of the epidemics has been made using the classes such as the infected and recovered/removed individuals. By modeling the real-world data that we have obtained from Johns Hopkins University's Github repository, we would like to gain a deeper insight into the behavior of the Covid-19 virus. As we have observed since the beginning of this year, the virus has not exactly behaved in the way scientists/virologists predicted it to be; thus, we believe that it would be interesting to model the spread of the virus, using the various modeling tools and techniques we have learned in the class. For our paper, we have chosen the confirmed and recovered cases from Nepal to model the general behavior of the disease. The first case in Nepal was confirmed on 23rd January 2020 when a 31 year old student, who had returned to Kathmandu from Wuhan on 9th January 2020, tested positive for the virus. It was also the first recorded case in South Asia. This fact makes it even more interesting because the population density of the entire South Asian Region is really high; thus, it would be interesting to visualize the spread of the disease over an extended period of time. The data we have obtained contains the latest values for the different classes of the disease, so, in this project, we have been able to use the data from January until November to predict the relationships between different classes of the disease.

OBJECTIVE

To model the amount of recovery of SARS-Cov-2 as a function of newly reported cases of infection. Some of the underlying assumptions are:-

1. The birth rate and death rate are equal at a particular time throughout all districts of Nepal.
2. The total population of Nepal is categorized into three groups: infected people who are actively carrying the virus, people who recovered from the virus, and people who have not been infected regardless of whether they were in contact or not with an infected person.

FORMULATION AND METHODOLOGY

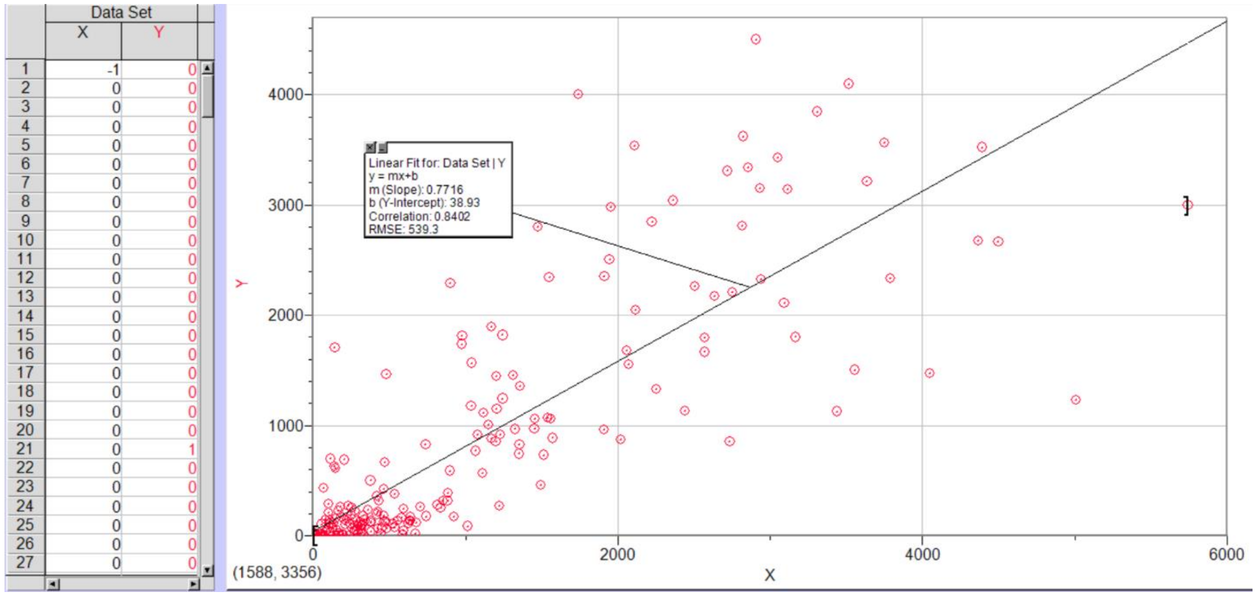
For this project, we are using the curve fitting criterion called the least-square criterion to determine the parameters of the functions that we are using to fit the data. After collecting the data, we separated the total recovered and total confirmed cases and wrote the data in a separate excel file. The daily newly recovered and confirmed cases was calculated by subtracting the previous count from the total count using the Excel formula. The descriptive statistics of each model was found out from SAS. After graphing the newly recovered cases vs newly confirmed cases, we used three models to fit the data:-

1. Linear fit
2. Power curve fit
3. Transformed least squares fit

Because we have the data values for approximately 300 days, we believe we can set up an accurate model that reflects the spread of the virus. The range of the data is from null values i.e. no cases to several hundred thousand cases in the recent months for one of the three estimated models. The spacing of the data points is equal to one day, as the dataset we have the daily number of cases for each class of the virus. After determining the parameter of the three given types of functions, we have calculated the absolute deviations from the fit to obtain the largest of the absolute deviation and the bounds on the absolute deviation. By using the largest of the absolute deviation and its bounds, we have concluded the best model that fits the behavior of the Covid-19 pandemic for a particular region i.e. Nepal.

CALCULATION AND FITTING

1. **Linear Fitting:**



From the dataset, we can observe that as the number of confirmed cases increases, the number of recovered cases increase as well. Thus, we can conclude that the number of recovered cases is proportional to the number of confirmed cases. Also, the k value i.e ratio of the data points is fairly close to each other, which suggests that there is potentially a linear relationship between the two variables.

Here, we consider fitting a model y = ax+b using the least square criterion. We call the least square estimate of the model f(x) = ax+b. Application of the criterion then requires the minimization of

$S = \sum_{i=1}^{m}[yi - f(xi)]$ ^ 2 = $\sum_{i=1}^{m}[yi - (axi + b)]$ ^ 2

After equating the partial differentiation of the above function with respect to a and b with 0, We obtain:-

$a = (m\, \Sigma x_i\, y_i - \Sigma x_i\, \Sigma y_i)/\ m\Sigma x_i^2 - (\Sigma x_i)^2$

$b = (\Sigma x_i^2\, \Sigma y_i - \Sigma x_i\, y_i\, \Sigma x_i)/\ m\Sigma x_i^2 - (\Sigma x_i)^2$

From SAS code and excel table,

$a = [(299*389464500) - (209776*173430)]/[(299*494151158) - (209776)^2]$

$= (1.165*10^{11} - 3.638*10^{10})/(1.478*10^{11} - 4.401*10^{10})$

= 0.772

b= $[(494151158*173430) – (389464500*209776)]/[(299*389464500) – (209776)^2]$

= $(8.570*10^{13} – 8.17*10^{13})/ (1.164*10^{11} – 4.401*10^{10})$

= 55.26

Therefore, the estimated model for linear fit is:-

y= 0.772x + 55.26

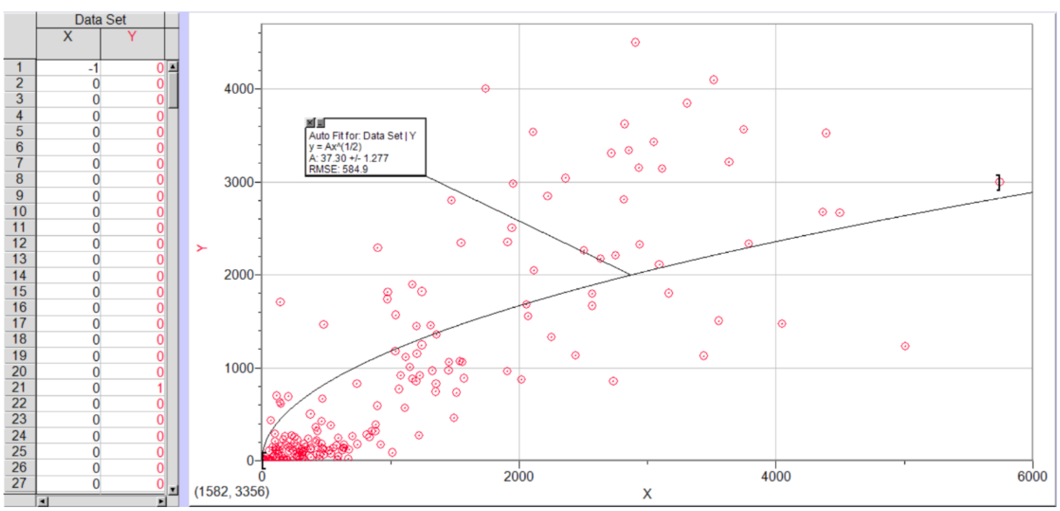
with the maximum deviation, dmax= 2692.436 from SAS code result.

From the given data,

D=$\sqrt{(85883945.49)/299}$ = 535.95 ≤ Cmax

The bounds on Cmax are: 535.95 ≤ Cmax ≤ 2692.436.

2. Power Curve fitting



The behavior of the data points is such that it is increasing but it is not sharply increasing i.e. an increasing and concave down function could be used to model it. Also, the domain of the function is positive i.e. the number of recovered cases and the number of confirmed cases are both positive.

Here, we consider fitting a model $y = ax^{1/2}$ using the least square criterion. We call the least square estimate of the model $f(x) = ax^{1/2}$. Application of the criterion then requires the minimization of

$S = \sum_{i=1}^{m}[yi - f(xi)]$ ^ 2 = $\Sigma[y_i - ax_i^{1/2}]^2$

After equating the partial differentiation of the above function with respect to a with 0, We obtain:-

$a = (\Sigma x_i^{1/2} y_i)/\Sigma x_i$

Form the SAS code and excel table,

a= 7824050.039/209777= 32.279

Therefore, the estimated model for power fit is:-

$y = 32.279x^{1/2}$

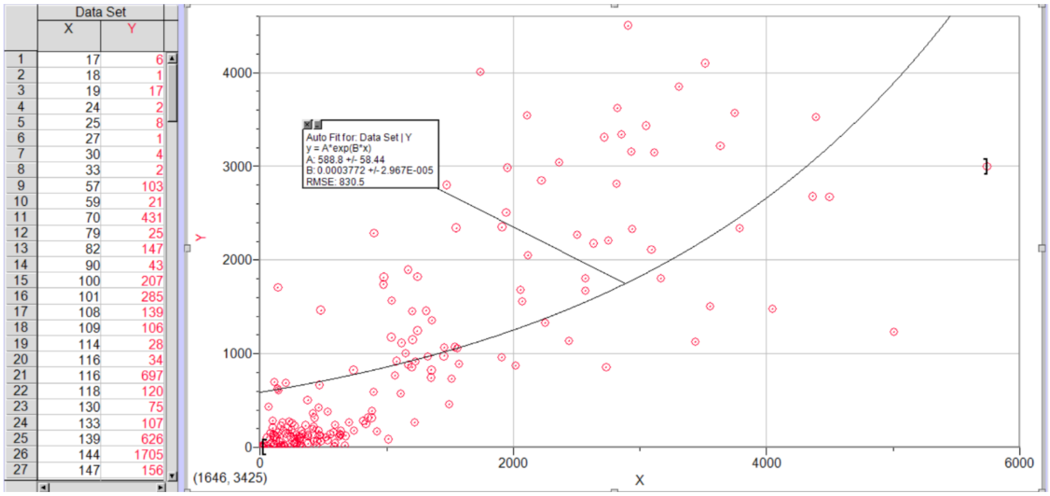
with the maximum deviation, dmax= 2488.381 from SAS code result.

From the given data,

$D=\sqrt{(101256557.5)/299} = 581.94 \leq$ Cmax

The bounds on Cmax are $581.94 \leq \text{Cmax} \leq 2488.38$

3. **Transformed least squares fit**



When we glanced through the data, it seemed to have an exponential dependency. As most of the natural processes have an exponential relationship, we predicted that the relationship between the number of confirmed cases and the number of recovered cases could also show a similar relationship. We used the graphical analysis software to fit the exponential function's curve; the datapoint seems to be fairly distributed around the curve. The exponential function we have fitted for the data points is increasing and passes through the point that is close to (0,1000). While it is true that the graph is not sharply increasing like an exponential, it still has some properties of an exponential function. Moreover, the range of the values on the y-axis is greater than 0, so that it also suffices one of the important properties of an exponential function.

Here, we consider fitting a model $y = ae^{bx}$ using the least square criterion. We call the least square estimate of the model $f(x) = ae^{bx}$. Application of the criterion then requires the minimization of

$$S = \sum_{i=1}^{m}[yi - f(xi)] \wedge 2 = \sum_{i=1}^{m}\left[yi - (ae^{bxi})\right] \wedge 2$$

We are using the given equation that is $y = ae^{bx}$ to a collection of data points. By taking the logarithm of both sides of the equation, we obtain the following:

$y = ae^{bx}$

or, $\ln(y) = \ln(ae^{bx})$

or, $\ln(y) = \ln(a) + \ln(e^{bx})$

or, ln(y) = ln(a) + bx ln€

or, ln(y) = ln(a) + bx

When the variables in lny versus x is plotted, the equation given above yield as straight line. On that particular graph, ln(a) is the intercept and b is the slope of the line. From the table of data points that we have transformed to fit the above equation, we use the following formula to obtain the required parameters for the exponential equation.

While we were transforming the data to fit the exponential fit, we noticed that there were several null values in our original dataset. Because the natural logarithm of zero values yields a math error, we cleaned the dataset such that we only analyzed the data points with non-zero values. This was a limitation of the model; however, we still could analyze the dataset from about 185 days, which gives enough information to create an accurate formula that models the data as we want.

The formula is:

$$b = \frac{185\sum(x)(\ln(y)) - (\sum x)(\sum lny)}{185\sum(lnx)^2 - (\sum lnx)^\wedge 2}$$

$$= \frac{(1858*1485520) - (209425*1049.68)}{185(494131865) - (209425)^\wedge 2}$$

= 1.156 * 10^-3

$$a = \frac{185\sum(x)^\wedge 2(\ln(y)) - (\sum x*lny)(\sum x)}{185\sum(lnx)^2 - (\sum lnx)^\wedge 2}$$

= 3010

Therefore, the estimated model for exponential fit is:-

y= $3010e^{(1.156*10^\wedge-3)x}$

with the maximum deviation, dmax= 2297426.49 from the SAS code result.

From the given data,

D=$\sqrt{(7475678898070)/185}$ = $4.041*10^{10} \leq$ Cmax

The bounds on Cmax are $4.041*10^{10} \leq$ Cmax $\leq$ 2297426.49.

Table of Deviation

Models	dmax	Sum of squared deviation
Linear	2692.436	85883945.49
Square root	2488.381	101256557.5
Exponential	2297426.49	7.47*10^12

Conclusion:

From the table, we observed that the linear fit has the lowest sum of the squared deviations; however, the power curve fit has the lowest maximum deviation. We would like to move forward with the power curve deviation because it captures the behavior of the datapoint in more detail than the linear graph, which says nothing more than the fact that the two variables are proportional to each other. Hence, from the data analysis and mathematical modeling of the dataset, we can conclude that the amount of recovered cases of COVID-19 approximates the square root of the amount of newly confirmed cases in Nepal at a particular time.

References

Giordano, Frank R. et. al. *A First Course in Mathematical Modeling*. Brooks/Cole Cengage Learning, 2014.

Johns Hopkins University, P 2020. COVID-19 Data Repository by the Center for Systems Science and Engineering(CSSE)., DOI: https://github.com/CSSEGISandData/COVID-19

```
/*
 *
 * Task code generated by SAS Studio 3.8
 *
 * Generated on '05/10/2020 23:07'
 * Generated by 'u58230396'
 * Generated on server 'ODAWS01-USW2.ODA.SAS.COM'
 * Generated on SAS platform 'Linux LIN X64 3.10.0-1062.9.1.el7.x86_64'
 * Generated on SAS version '9.04.01M6P11072018'
 * Generated on browser 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/88.0.4324.19
 * Generated on web client 'https://odamid-usw2.oda.sas.com/SASStudio/main?locale=en_GB&zone=GMT-05%253A00&ticket=ST-402715-n
 *
 */

ods noproctitle;
ods graphics / imagemap=on;

proc means data=WORK.IMPORT chartype mean std min max n vardef=df;
    var 'd=Y-Yi'n;
run;
```

Analysis Variable : d=Y-Yi d=Y-Yi				
Mean	Std Dev	Minimum	Maximum	N
-16.7266846	537.4856367	-2692.44	2672.70	298

```
/*
 *
 * Task code generated by SAS Studio 3.8
 *
 * Generated on '05/10/2020 23:12'
 * Generated by 'u58230396'
 * Generated on server 'ODAWS01-USW2.ODA.SAS.COM'
 * Generated on SAS platform 'Linux LIN X64 3.10.0-1062.9.1.el7.x86_64'
 * Generated on SAS version '9.04.01M6P11072018'
 * Generated on browser 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/88.0.4324.19
 * Generated on web client 'https://odamid-usw2.oda.sas.com/SASStudio/main?locale=en_GB&zone=GMT-05%253A00&ticket=ST-402715-n
 *
 */

ods noproctitle;
ods graphics / imagemap=on;

proc means data=WORK.IMPORT chartype mean std min max n sum vardef=df;
    var 'd^2'n;
run;
```

Analysis Variable : d^2 d^2					
Mean	**Std Dev**	**Minimum**	**Maximum**	**N**	**Sum**
288201.16	891849.87	0.1324960	7249211.61	298	85883945.49

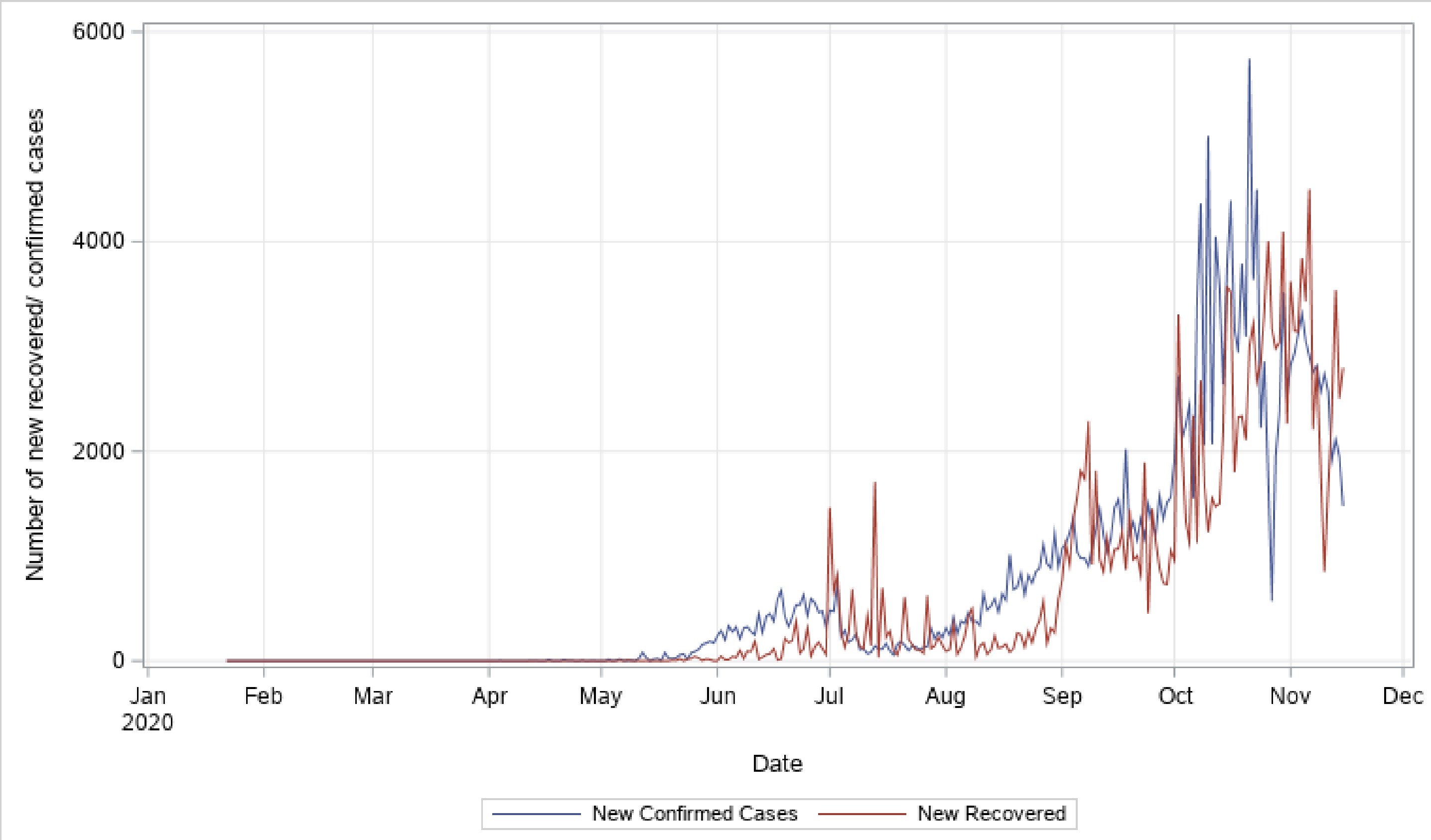
```
/*
 *
 * Task code generated by SAS Studio 3.8
 *
 * Generated on '05/10/2020 23:21'
 * Generated by 'u58230396'
 * Generated on server 'ODAWS01-USW2.ODA.SAS.COM'
 * Generated on SAS platform 'Linux LIN X64 3.10.0-1062.9.1.el7.x86_64'
 * Generated on SAS version '9.04.01M6P11072018'
 * Generated on browser 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/88.0.4324.19
 * Generated on web client 'https://odamid-usw2.oda.sas.com/SASStudio/main?locale=en_GB&zone=GMT-05%253A00&ticket=ST-402715-n
 *
 */

ods noproctitle;
ods graphics / imagemap=on;

proc means data=WORK.IMPORT1 chartype min max n sum vardef=df;
    var 'Y-Yi=d'n X ZY;
run;
```

Variable	Label	Minimum	Maximum	N	Sum
Y-Yi=d	Y-Yi=d	-1410.41	2488.38	297	-30804.79
X	X	0	5743.00	297	209777.00
ZY	ZY	0	242910.41	297	7824050.04

```
/*
 *
 * Task code generated by SAS Studio 3.8
 *
 * Generated on '05/10/2020 23:37'
 * Generated by 'u58230396'
 * Generated on server 'ODAWS01-USW2.ODA.SAS.COM'
 * Generated on SAS platform 'Linux LIN X64 3.10.0-1062.9.1.el7.x86_64'
 * Generated on SAS version '9.04.01M6P11072018'
 * Generated on browser 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/88.0.4324.19
 * Generated on web client 'https://odamid-usw2.oda.sas.com/SASStudio/main?locale=en_GB&zone=GMT-05%253A00&ticket=ST-402715-n
 *
 */

ods noproctitle;
ods graphics / imagemap=on;

proc means data=WORK.IMPORT2 chartype min max n sum vardef=df;
    var 'y-yi'n '(y-yi) squared'n 'X squared *lnY'n 'X*lnY'n X lnY Y;
run;
```

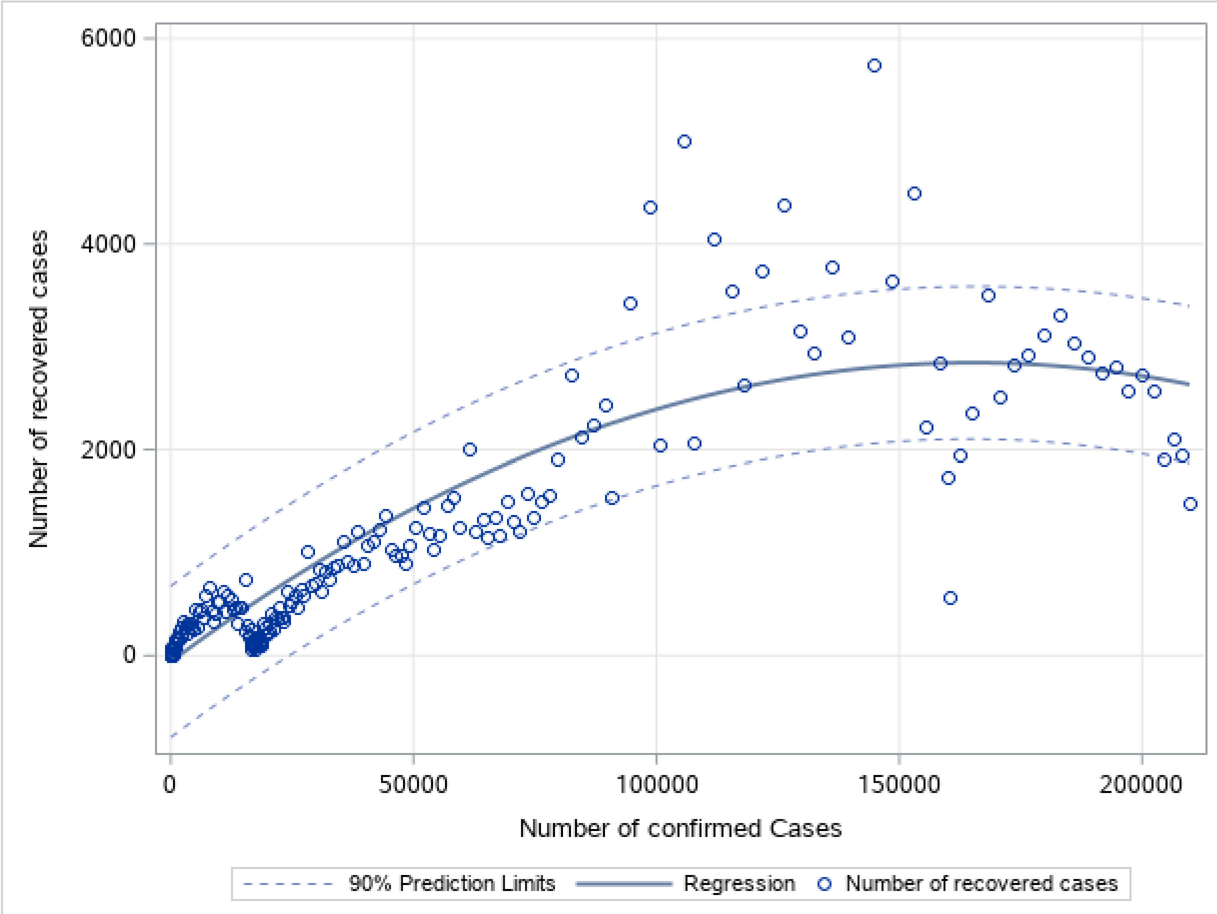
Variable	Label	Minimum	Maximum	N	Sum
y-yi	y-yi	-2297426.50	-1850.17	184	-9070050.31
(y-yi) squared	(y-yi) squared	3423144.77	5.2781685E12	184	7.4756789E12
X squared *lnY	X squared *lnY	0	264022402	184	3743029975
X*lnY	X*lnY	0	45972.91	184	1485520.19
X	X	17.0000000	5743.00	184	209425.00
lnY	lnY	0	8.4118327	184	1049.68
Y	Y	1.0000000	4500.00	184	173405.00

```
/*
 *
 * Task code generated by SAS Studio 3.8
 *
 * Generated on '02/10/2020 23:03'
 * Generated by 'u58230396'
 * Generated on server 'ODAWS01-USW2.ODA.SAS.COM'
 * Generated on SAS platform 'Linux LIN X64 3.10.0-1062.9.1.el7.x86_64'
 * Generated on SAS version '9.04.01M6P11072018'
 * Generated on browser 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/88.0.4324.19
 * Generated on web client 'https://odamid-usw2.oda.sas.com/SASStudio/main?locale=en_GB&zone=GMT-05%253A00&https%3A%2F%2Fodam
 *
 */

ods graphics / reset width=8in height=4.7in imagemap;

proc sort data=WORK.IMPORT1 out=_SeriesPlotTaskData;
    by Date;
run;

proc sgplot data=_SeriesPlotTaskData;
    series x=Date y='New Confirmed Cases'n /;
    series x=Date y='New Recovered'n /;
    xaxis grid;
    yaxis grid label="Number of new recovered/ confirmed cases";
run;

ods graphics / reset;

proc datasets library=WORK noprint;
    delete _SeriesPlotTaskData;
    run;
```


6000
4000
2000
0
Number of new recovered/ confirmed cases
Jan 2020
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Date
New Confirmed Cases
New Recovered

```
/*
 *
 * Task code generated by SAS Studio 3.8
 *
 * Generated on '02/10/2020 22:33'
 * Generated by 'u58230396'
 * Generated on server 'ODAWS02-USW2.ODA.SAS.COM'
 * Generated on SAS platform 'Linux LIN X64 3.10.0-1062.9.1.el7.x86_64'
 * Generated on SAS version '9.04.01M6P11072018'
 * Generated on browser 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/88.0.4324.19
 * Generated on web client 'https://odamid-usw2.oda.sas.com/SASStudio/main?locale=en_GB&zone=GMT-05%253A00&ticket=ST-368310-M
 *
 */

ods graphics / reset width=6.4in height=4.8in imagemap;

proc sgplot data=WORK.IMPORT;
    scatter x='Number of confirmed Cases'n y='Number of recovered cases'n /;
    xaxis grid;
    yaxis grid;
run;

ods graphics / reset;
```

```
/*
 *
 * Task code generated by SAS Studio 3.8
 *
 * Generated on '02/10/2020 22:40'
 * Generated by 'u58230396'
 * Generated on server 'ODAWS02-USW2.ODA.SAS.COM'
 * Generated on SAS platform 'Linux LIN X64 3.10.0-1062.9.1.el7.x86_64'
 * Generated on SAS version '9.04.01M6P11072018'
 * Generated on browser 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/88.0.4324.19
 * Generated on web client 'https://odamid-usw2.oda.sas.com/SASStudio/main?locale=en_GB&zone=GMT-05%253A00&ticket=ST-394829-G
 *
 */

ods graphics / reset width=6.4in height=4.8in imagemap;

proc sgplot data=WORK.IMPORT;
    reg x='Number of confirmed Cases'n y='Number of recovered cases'n / nomarkers
        degree=2 cli alpha=0.10;
    scatter x='Number of confirmed Cases'n y='Number of recovered cases'n /;
    xaxis grid;
    yaxis grid;
run;

ods graphics / reset;
```


6000
4000
2000
0
Number of recovered cases
0
50000
100000
150000
200000
Number of confirmed Cases
90% Prediction Limits
Regression
Number of recovered cases

```
/*
 *
 * Task code generated by SAS Studio 3.8
 *
 * Generated on '02/10/2020 22:51'
 * Generated by 'u58230396'
 * Generated on server 'ODAWS01-USW2.ODA.SAS.COM'
 * Generated on SAS platform 'Linux LIN X64 3.10.0-1062.9.1.el7.x86_64'
 * Generated on SAS version '9.04.01M6P11072018'
 * Generated on browser 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/88.0.4324.19
 * Generated on web client 'https://odamid-usw2.oda.sas.com/SASStudio/main?locale=en_GB&zone=GMT-05%253A00&https%3A%2F%2Fodam
 *
 */

ods noproctitle;
ods graphics / imagemap=on;

proc pls data=WORK.IMPORT method=pls plots;
    model 'Number of recovered cases'n='Number of confirmed Cases'n;
run;
```

Data Set	WORK.IMPORT
Factor Extraction Method	Partial Least Squares
PLS Algorithm	NIPALS
Number of Response Variables	1
Number of Predictor Parameters	1
Missing Value Handling	Exclude
Number of Factors	1

Number of Observations Read	299
Number of Observations Used	299

Percent Variation Accounted for by Partial Least Squares Factors				
	Model Effects		**Dependent Variables**	
Number of Extracted Factors	**Current**	**Total**	**Current**	**Total**
1	100.0000	100.0000	75.2570	75.2570