REPORT ON DATA ANALYSIS PROJECT

The impact of stringency index and VRR on daily new deaths due to COVID-19 in Egypt, From February 2020 to May 2021

BY :

1-Peter Adel Sobhy

2-Eman Ahmed Abd El-Hameed

3-Mai Abd El salam Mohamed

4- Rana Mohamed Ibrahem

5-Hamza Mahmoud hussain

Under The Supervision of DR :

Abd Elhameed Elshabrawy

Table of Contents

Introduction :

In the midst of pandemic caused by severe acute respiratory syndrome–coronavirus 2 (SARS-CoV-2), now popularly referred to as coronavirus disease 2019 (COVID-19), many strategies have been adopted by governments as to how to fight this illness that has affected many lives. Non-pharmaceutical interventions (NPIs) are being implemented to eradicate the disease but these have just delayed and moderated the spread of the virus. A variety of regulations has been designed, among which include restriction of non-essential travel, imposition of social distancing, declaration of lockdowns on big cities, and banning mass gatherings, all supposedly to prevent the spread of the novel coronavirus. The COVID-19 disease was first identified in Wuhan, China, in early December 2019 with the early cases reported in the city resulting in its spread worldwide that brought negatively significant changes to the normal lives of people in different countries of the world, But we will conduct this study on Egypt only.

Problem Study:

The problem that will be discussed in this study is the death rate in Egypt within the precautionary measures taken to limit the spread of the virus among citizens, and whether these measures are taken to help reduce the possibility of death for patients or not.

Objective:

Knowing the impact of precautionary measures on the death numbers from the covid19

Hypotheses:

H0 = P > 0.05

Null hypothesis: There are no statistically significant differences after following the precautionary measures.

H1 = P < 0.05

Alternative hypothesis: There are statistically significant differences after following the precautionary measures

Variables :

1-New_deaths:

The number of dead people infected from COVID-19 every day.

2- Virus reproduction rate (VRR)

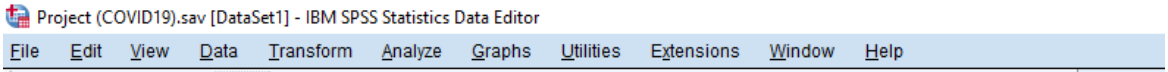
It is the rate of reproduction of the virus under the precautionary measures.

3- Stringency Index

What is the rate of following precautionary measures in the conditions of the spread of the virus?

The Capture of data:

The data used in this study is taken from the World Health Organization, and it includes all countries of the world, but as we mentioned before, we will do this study on Egypt only, so this is considered secondary data for us.



	Date	Y	X1	X2_index	X2	Y_adj
1	2020-03-15	0	1.55	11.11	1.00	1.00
2	2020-03-16	0	1.61	18.52	1.00	1.00
3	2020-03-17	2	1.61	18.52	1.00	1.00
4	2020-03-18	2	1.55	18.52	1.00	1.00
5	2020-03-19	0	1.56	29.63	1.00	1.00
6	2020-03-20	2	1.49	29.63	1.00	1.00
7	2020-03-21	2	1.45	40.74	1.00	1.00
8	2020-03-22	4	1.46	40.74	1.00	1.00
9	2020-03-23	5	1.43	40.74	1.00	1.00
10	2020-03-24	1	1.43	51.85	1.00	1.00
11	2020-03-25	1	1.45	84.26	4.00	1.00
12	2020-03-26	3	1.39	84.26	4.00	1.00
13	2020-03-27	6	1.40	84.26	4.00	1.00
14	2020-03-28	6	1.43	84.26	4.00	1.00
15	2020-03-29	4	1.44	84.26	4.00	1.00
16	2020-03-30	1	1.47	84.26	4.00	1.00
17	2020-03-31	5	1.53	84.26	4.00	1.00
18	2020-04-01	6	1.61	84.26	4.00	1.00
19	2020-04-02	6	1.67	84.26	4.00	1.00
20	2020-04-03	8	1.72	84.26	4.00	1.00
21	2020-04-04	5	1.70	84.26	4.00	1.00

Analysis

As a start for the analysis, all the variables were studied carefully in order to choose the most optimum method for analysis. The methods we used depended on the type of data, and the assumptions of the test itself. This section covers the descriptive analysis of all variables. Besides, it discusses all the steps that our team applied to the data to reveal the most output information that the data may hide.

Descriptive Analysis

The analysis includes three variables, one dependent variable "New Deaths" and two independent variables "VRR" and "Stringency Index". A descriptive analysis was studied for all variables. First, a descriptive study was made to the dependent quantitative variable "New Deaths" as shown in the next table.

✓ Descriptive Analysis

The analysis includes three variables, one dependent variable "New Deaths" and two independent variables "VRR" and "Stringency Index". A descriptive analysis was studied for all variables. First, a descriptive study was made to the dependent quantitative variable "New Deaths" as shown in the next table.

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
New Deaths	421	0	97	33.02	22.223	.672	.119	-.323	.237
Valid N (listwise)	421								

From the table, we can determine that the minimum and maximum daily new deaths are 0 and 97 respectively. The average is 33 deaths a day. The skewness is 0.672 and indicates that there is a possible abnormality in the data, and it will be checked again lately. The kurtosis is -0.323 (between -3 and 3) so it is accepted.
Then, we made the descriptive analysis for the Stringency Index variable after it was converted from quantitative to categorical, so that we can make it easier in analysis, and was renamed "Stringency Level" in which:
less than 60 = 1 (Minimum)
From 60.01 to 70 = 2 (Low)
From 70.01 to 80 = 3 (high)
More than 80.01 = 4 (maximum)

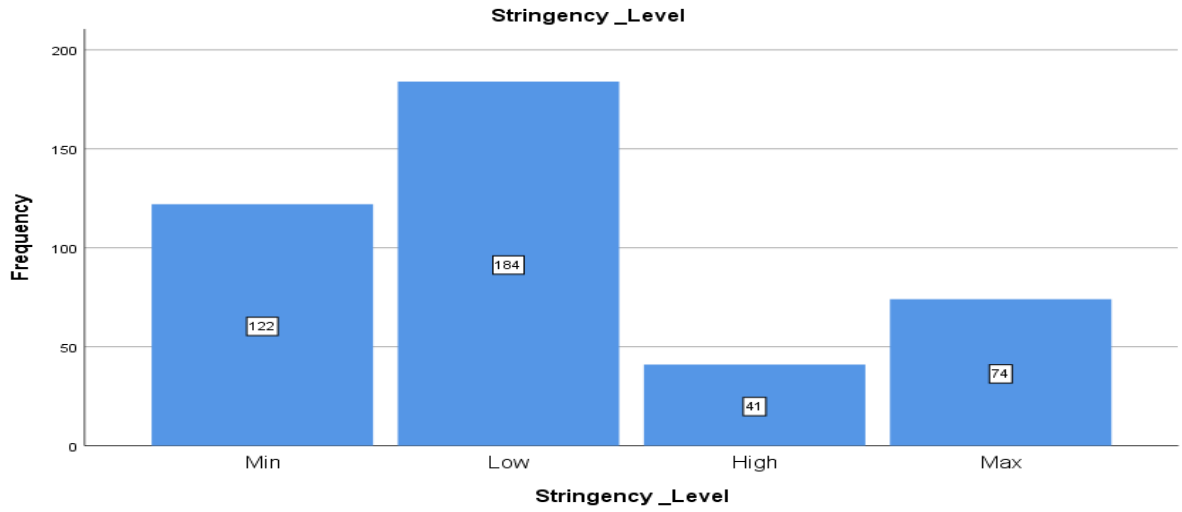
Statistics

Stringency _Level

N	Valid	421
	Missing	0

Stringency _Level

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Min	122	29.0	29.0	29.0
	Low	184	43.7	43.7	72.7
	High	41	9.7	9.7	82.4
	Max	74	17.6	17.6	100.0
	Total	421	100.0	100.0	



From the bar chart we can deduce that the more frequent stringency level is “low” in 184 out of 421 days. Next, the minimum level of stringency got the second rank with 122 days. Then, the maximum level was rank 3 with 74 days. And fourth rank was for the high level with only 41 days.

Finally, descriptive study was made to the second independent quantitative variable “VRR” as shown in the next table.

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
Reproduction_Rate	421	.45	1.72	1.0779	.22849	.023	.119	.289	.237
Valid N (listwise)	421								

From the table we can determine that the minimum and maximum daily virus reproduction rate are 0.454 and 1.72 respectively. The average is 1.0779 day. The skewness is 0.023 and indicates that the data is perfectly normal as it is so close to zero. The kurtosis is -0.289 (between -3 and 3) so it is accepted.

✓ Normality of Variables

Normality is one of the essential assumptions the quantitative variable should have. Therefore, Normality is checked for the two quantitative variables we have: “New Deaths” variable, and “VRR”.

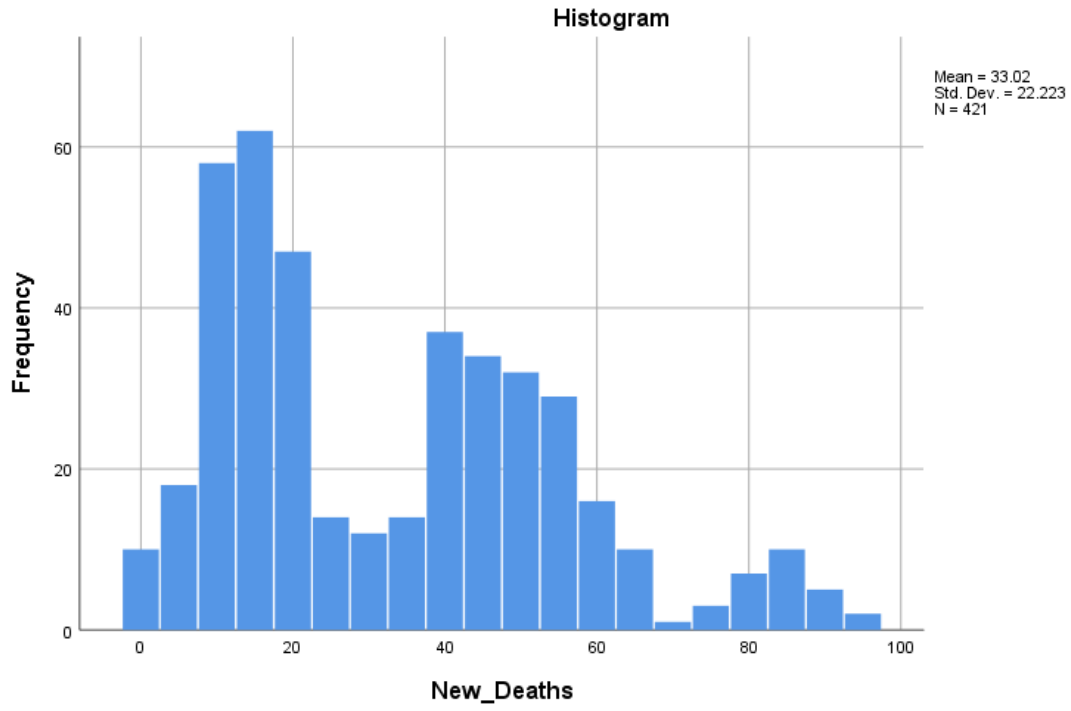
- New Deaths variable:

After checking the normality of the “New Deaths” variable, the following results revealed.

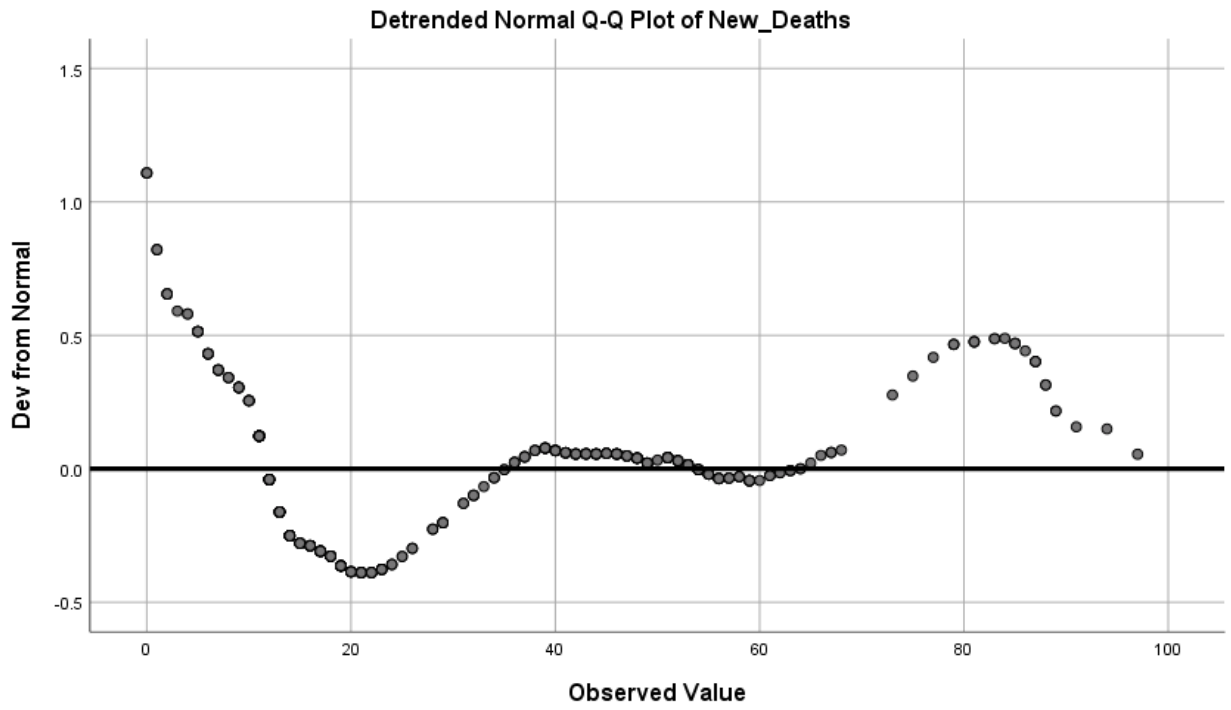
Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
New Deaths	.157	421	.000	.924	421	.000

a. Lilliefors Significance Correction


Histogram
Mean = 33.02
Std. Dev. = 22.223
N = 421
Frequency
0
20
40
60
0
20
40
60
80
100
New_Deaths


Normal Q-Q Plot of New_Deaths
Expected Normal
4
2
0
-2
0
30
60
90
120
Observed Value



From the Histogram and significance of the test, we can deduce that the new deaths variable is not normal. In addition, many points are far from the normal line in the two graphs, which also indicates that the variable is not normally distributed.

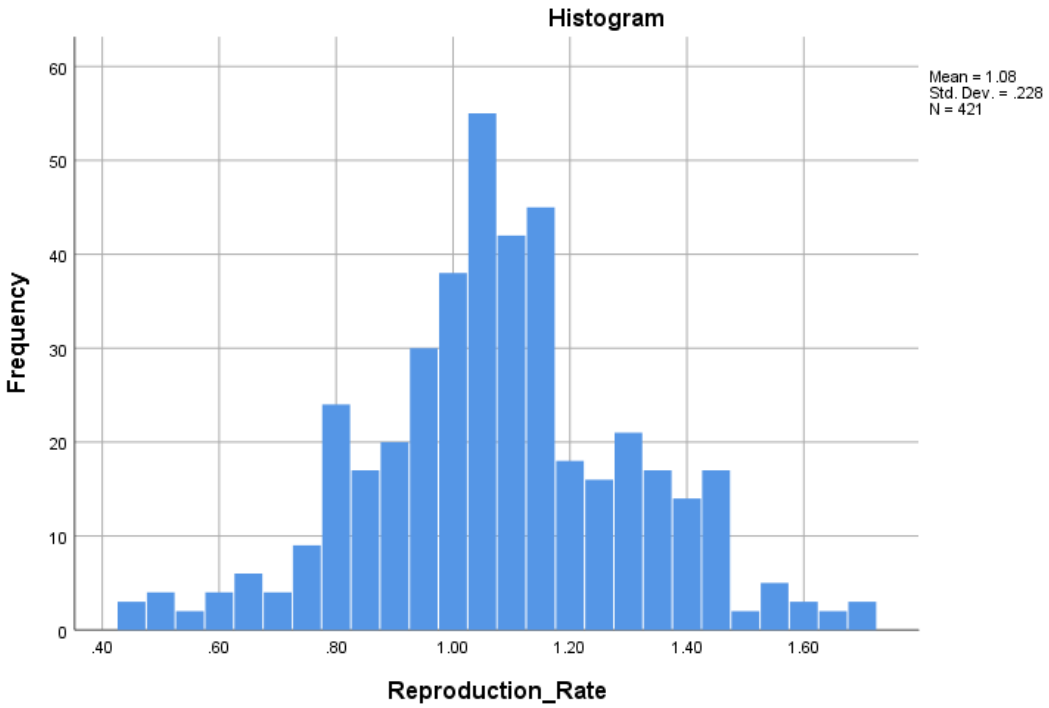
- VRR variable:

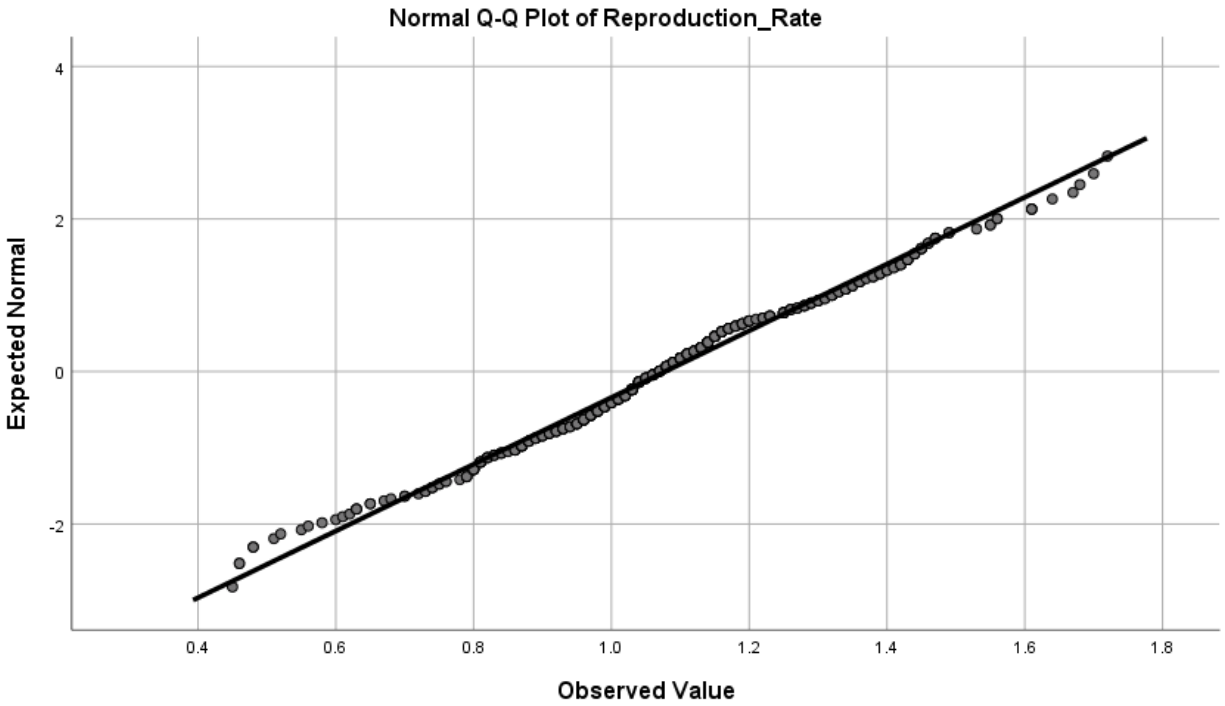
After checking the normality of the “VRR” variable, the following results revealed.

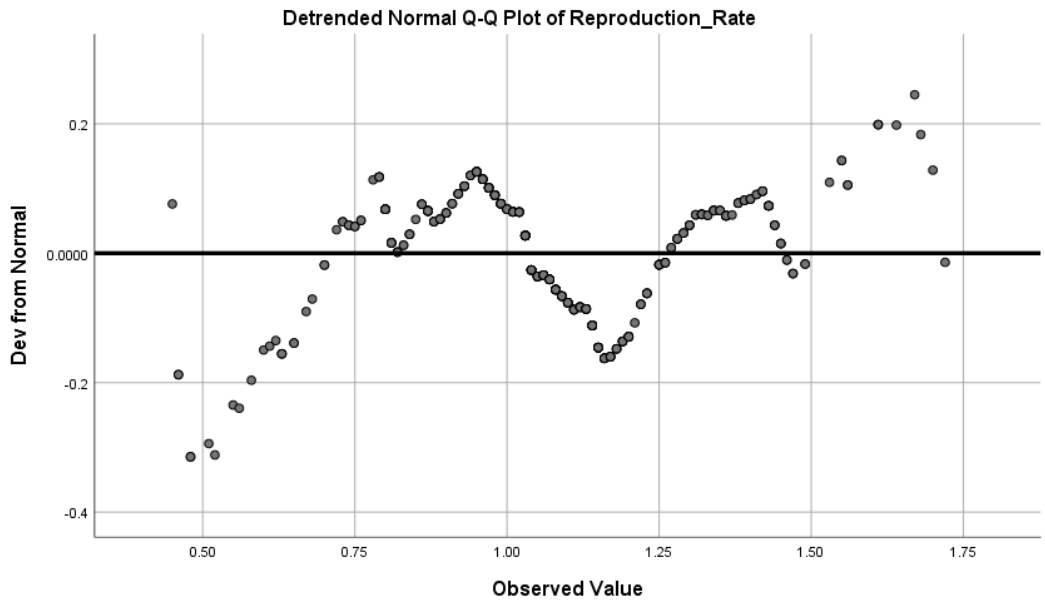
Tests of Normality

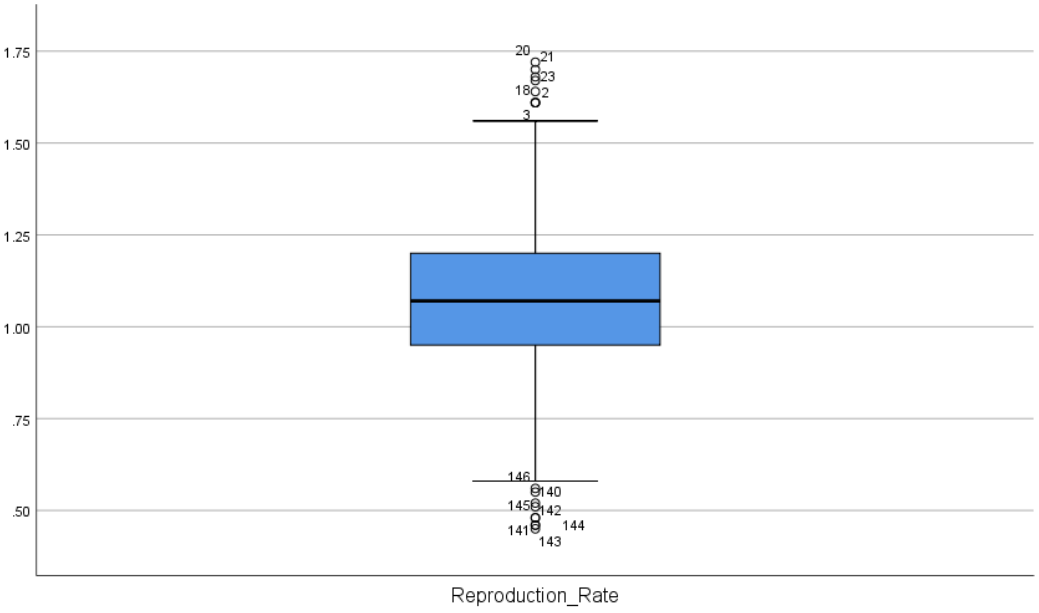
	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Reproduction_Rate	.068	421	.000	.991	421	.010

a. Lilliefors Significance Correction


Histogram
Mean = 1.08
Std. Dev. = .228
N = 421
Frequency
60
50
40
30
20
10
0
.40
.60
.80
1.00
1.20
1.40
1.60
Reproduction_Rate


Normal Q-Q Plot of Reproduction_Rate
Expected Normal
4
2
0
-2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
1.8
Observed Value





From the Histogram and significance of the test, we can deduce that the VRR variable is not normal. In addition, many points are far from the normal line in the two graphs, which also indicates that the variable is not normally distributed. The box plot shows that there are many outliers too.

Form the previous study, it is clear that the variables are not normal. The ordinal logistic regression was chosen to complete the analysis because the normality is not one of its assumptions.

To use the ordinal logistic regression model, the dependent variable must be ordinal. So, we convert the dependent variable "new deaths" from a quantitative variable to ordinal, in which:

New deaths	Burden	Rank
Less than 25	least	1
From 26 to 50	Moderate	2
From 51 to 75	Important	3
More than 76	Critical	4

Then the test of parallel line was made to the data and it was significant. The given data cannot be analyzed by ordinal logistic regression because the test will lose the property of order. And as an alternative to ordinal logistic regression, three binary logistic regression models were applied to the data.

✓ Binary Logistic Regression

Binary logistic regression is used when the dependent variable is binary. So, the data was classified to three parts according to the "New Deaths" variable categories, in which:
The first model includes the categories "Least" and "Moderate" burden of disease. The first model includes the categories "Moderate" and "Important" burdens. And the third model includes "important" and "Critical", taking stringency level 4 as a reference to all other stringency levels.

First, binary logistic regression was tested between "least" and "moderate" categories of "New Deaths". The following results are revealed.

Case Processing Summary

Unweighted Cases		N	Percent
Selected Cases	Included in Analysis	324	100.0
	Missing Cases	0	.0
	Total	324	100.0
Unselected Cases		0	.0
Total		324	100.0

a. If weight is in effect, see classification table for the total number of cases.

This table indicates that there are no missing cases.

Dependent Variable Encoding

Original Value	Internal Value
Least	0
Moderate	1

This table indicates that the moderate category is the reference category.

Block 1: Method = Enter

Iteration History[a,b,c,d]

Iteration		-2 Log likelihood	Coefficients: Constant	Virus Reproductive Rate	Stringency Level(1)	Stringency Level(2)	Stringency Level(3)
Step 1	1	269.956	1.854	-2.347	2.195	-.842	-.152
	2	258.795	3.452	-3.748	2.628	-1.461	-.249
	3	258.218	3.972	-4.168	2.727	-1.691	-.298
	4	258.215	4.013	-4.200	2.735	-1.711	-.303
	5	258.215	4.013	-4.200	2.735	-1.711	-.303

a. Method: Enter

b. Constant is included in the model.

c. Initial -2 Log Likelihood: 424.955

d. Estimation terminated at iteration number 5 because parameter estimates changed by less than .001.

From this table we notice that the -2 log likelihood (the un explained data) decreases from 269.956 to 258.215 through 5 iterations. This means that the predictor variables explain variability compared to empty model but not that much.

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	166.740	4	.000
	Block	166.740	4	.000
	Model	166.740	4	.000

Here we reject that the right hand side of the equation is equal to zero

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	258.215[a]	.402	.551

a. Estimation terminated at iteration number 5 because parameter estimates changed by less than .001.

The value of Nagelkerke R Square is acceptable if they are ranged from 0.2 to 0.4, but as they reach 1 the model is assumed to has high predictive power.

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	44.897	8	.000

Hosmer and Lemeshow test the model goodness of fit , in which we reject the null hypothesis (goodness of fit) unfortunately.

Classification Table[a]

Observed		Predicted: New Deaths (least, Moderate)		Percentage Correct
		Least	Moderate	
Step 1 New Deaths (least, Moderate)	Least	193	13	93.7
	Moderate	41	77	65.3
Overall Percentage				83.3

a. The cut value is .500

From this table we deduce that the analysis could explain 193 times in case of the “least” category with a percentage of 93.7%, and also 77 times in case of “moderate” category with a percentage 65.3%.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Virus Reproductive Rate	-4.200	.837	25.151	1	.000	.015

	Stringency Level			83.221	3	.000	
	Stringency Level(1)	2.735	.496	30.368	1	.000	15.403
	Stringency Level(2)	-1.711	.561	9.303	1	.002	.181
	Stringency Level(3)	-.303	.597	.258	1	.612	.738
	Constant	4.013	1.159	11.998	1	.001	55.339

a. Variable(s) entered on step 1: Virus Reproductive Rate, Stringency Level.

This final table shows the variables in the equation "log(p/1-p) = b0 + b2*x2 + b3*x3+b4*x4"
We can conclude that following:

- For everyone unit increase in "Virus Reproductive Rate", the log odds of least number of New deaths versus moderate decreases by 4.2 (accepted logically)
- Stringency level with rank of 1, versus rank of 4, increases the log odds of the least versus moderate category of new deaths by 2.735 (not accepted logically)
- Having Stringency level with rank of 2, versus rank of 4, decreases the log odds of the least versus moderate category of new deaths by 1.711 (accepted logically)
- There is no significant difference having Stringency level with rank of 3, versus rank of 4 to change the log odds of the least versus moderate category of New deaths

<u>Secondly, binary logistic regression was tested between "moderate" and "important" categories of "New Deaths". The following results are revealed.</u>

Case Processing Summary

Unweighted Cases		N	Percent
Selected Cases	Included in Analysis	190	100.0
	Missing Cases	0	.0
	Total	190	100.0
Unselected Cases		0	.0
Total		190	100.0

This table indicates that there is no missing cases.

Dependent Variable Encoding

Original Value	Internal Value
Moderate	0
Important	1

This table indicates that the important category is the reference category.

Block 1: Method = Enter

Iteration History[a,b,c,d]

Iteration		-2 Log likelihood	Coefficients Constant	Virus Reproductive Rate	Stringency Level(1)	Stringency Level(2)	Stringency Level(3)
Step 1	1	231.117	-2.629	.470	1.477	2.486	.650
	20	227.925	-21.873	.500	20.662	21.705	19.495

From this table we notice that the -2 log likelihood (the un explained data) decreases from 231.117 to 227.925 through 20 iterations. This means that the predictor variables explain variability compared to empty model but not that much.

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	24.223	4	.000
	Block	24.223	4	.000
	Model	24.223	4	.000

Here we reject that the right hand side of the equation is equal to zero

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	227.925[a]	.120	.163

The value of Nagelkerke R Square is acceptable if they are ranged from 0.2 to 0.4, here is quite low.

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	32.330	8	.000

Hosmer and Lemeshow test the model goodness of fit , in which we reject the null hypothesis (goodness of fit) unfortunately.

Classification Table[a]

	Observed	Predicted

			New Deaths (moderate,important)		Percentage
			Moderate	Important	Correct
Step 1	New Deaths (moderate,important)	Moderate	93	25	78.8
		Important	38	34	47.2
	Overall Percentage				66.8

From this table we deduce that the analysis could explain 93 times in case of the “Moderate” category with a percentage of 78.8%, and also 34 times in case of “Important” category with a percentage 47.2%.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Virus Reproductive Rate	.500	.858	.340	1	.560	1.650
	Stringency Level			11.409	3	.010	
	Stringency Level(1)	20.662	11600.351	.000	1	.999	941026755.613
	Stringency Level(2)	21.705	11600.351	.000	1	.999	2669202346.403
	Stringency Level(3)	19.495	11600.351	.000	1	.999	292928299.157
	Constant	-21.873	11600.351	.000	1	.998	.000

a. Variable(s) entered on step 1: Virus Reproductive Rate, Stringency Level.

This final table shows the variables in the equation ”log(p/1-p) = b0 + b2*x2 + b3*x3+b4*x4” , but all the independent variables are not significant.

<u>Finally, binay logistic regression is between “important” and “critical” categories of the new deaths variable</u>

Case Processing Summary

Unweighted Cases[a]		N	Percent
Selected Cases	Included in Analysis	97	100.0
	Missing Cases	0	.0
	Total	97	100.0
Unselected Cases		0	.0
Total		97	100.0

This table indicates that there is no missing cases.

Dependent Variable Encoding

Original Value	Internal Value
Imortant	0
Critical	1

This table indicates that the Critical category is the reference category.

Iteration History[a,b,c,d]

Iteration		-2 Log likelihood	Coefficients Constant	Virus Reproductive Rate	Stringency Level(1)	Stringency Level(2)
Step 1	1	57.888	3.336	-1.513	-3.842	-3.106
	20	44.837	8.442	-5.407	-24.471	-4.951

From this table we notice that the -2 log likelihood (the un explained data) decreases from 57.888 to 44.837 through 20 iterations. This means that the predictor variables explain variability compared to empty model with high efficiency.

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	65.873	3	.000
	Block	65.873	3	.000
	Model	65.873	3	.000

Here we reject that the right hand side of the equation is equal to zero

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	44.837[a]	.493	.724

a. Estimation terminated at iteration number 20 because maximum iterations has been reached. Final solution cannot be found.

The value of Nagelkerke R Square is acceptable if they are ranged from 0.2 to 0.4, and when it approach 1 that means that it has high power.

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	3.609	7	.824

Hosmer and Lemeshow test the model goodness of fit , in which we have to accept that null hypothesis (goodness of fit) this time.

Classification Table[a]

Observed			Predicted		
			New Deaths(important,critical)		Percentage Correct
			Imortant	Critical	
Step 1	New Deaths(important,critical)	Imortant	71	1	98.6
		Critical	8	17	68.0
	Overall Percentage				90.7

a. The cut value is .500

From this table we deduce that the analysis could explain 71 times in case of the “Important” category with a percentage of 98.6%, and also 17 times in case of “Critical” category with a percentage 47.2%.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Virus Reproductive Rate	-5.407	3.575	2.287	1	.130	.004
	Stringency Level			15.214	2	.000	
	Stringency Level(1)	-24.471	6453.085	.000	1	.997	.000
	Stringency Level(2)	-4.951	1.269	15.214	1	.000	.007
	Constant	8.442	3.875	4.745	1	.029	4637.339

a. Variable(s) entered on step 1: Virus Reproductive Rate, Stringency Level.

This final table shows the variables in the equation “log(p/1-p) = b0 + b2*x2 + b3*x3+b4*x4”
We can conclude the following:

- For every one-unit change in “Virus Reproductive Rate”, the log odds of Important number of new deaths versus Critical decreases by 5.407 (accepted logically)
- There is no significance that having Stringency level with rank of 1, versus rank of 4, changes the log odds of the Important versus Critical category of new deaths.
- Having Stringency level with rank of 2, versus rank of 4, decreases the log odds of the moderate versus important category of new deaths by 8.442 (accepted logically)
- There is no Stringency Level (3) in the table because all the “important” and “critical” new deaths burden has level (1) or (2) or (4).

Conclusion:

In view of the research, we found that there is a link between Stringency level and VRR and the sew deaths as the stringency level is getting higher the deaths decreases until the level of 3 then there is no difference between level 3 and 4 as the result showed that but it's important to notice the VRR as it also increases the new deaths as there is newly affected people are, so to decrease the new deaths we have to work in rising the Stringency level to decrease also the VRR which these variables lead to having less new deaths