ISSN No:-2456-2165

A Simple Regression Analysis on COVID-19 Using Global Data (Using Total and Recovered Cases as on 16 May 2020)

A. R. Muralidharan[1] ; Tena Manaye[2]
[1] Assistant Professor, [2] Lecturer- Department of Statistics,
College of Natural and Computational Sciences, Debre Berhan University, Ethiopia

Abstract:- In the world, the current scenario is on COVID-19, a pandemic, is very critical; the global is occupied by the COVID-19, a pandemic and the Public Health international concern (PHIEC) announces the outbreak of COVID-19. Most of the people are affected in various dimensions as social, psychological, economic, employment, education, services and so on. World is facing an ordeal due to COVID-19 and it increasingly continues to spread rapidly around the entire world. Daily the report on this COVID-19 cases were recorded in the classification of a data collected form countries as in the categories of confirmed cases, new cases, total deaths, new deaths, total recovered, active cases and so on. The way to lead this article is based on the daily report gathered from various sources, the outbreak of coronavirus disease 2019 gave a global impact on health issues of the people and it has had a deep impact on the world and our daily routines.

- ***Method***

The contemporary study is undertaken to analyse the correlation between the COVID -19 Total and recovered cases from the data dated 16 May 2020 from the website of Worldometer's COVID-19 for the Global data. This article studies an application of the Simple Regression Log- Linear Models (LLM) to analyse the correlation between the COVID - 19 from total and Recovered cases in the reference period. These outputs of this article were carried out by using SPSS 20. The outputs are showing significant difference at 5% level of significance.

- ***Conclusion***

The study results revealed that there is a significant positive correlation between the COVID - 19 total cases and recovered cases on 16 May 2020 in the world. This study reveals that the regression Model(LLM) as total cases of COVID-19 increases the Recovered cases of COVID-19 is also increase, with *p* = 0.000 (which is marginally significant at alpha=0.05). More precisely, it says that for a one Case of total increase in average case of recovered, the predicted Recover cases increases by 0.0000093 points holding the percent of total cases constant. The Q-Q plots show that there is a systematic deviation from normality. The normal Q-Q plot shows that almost all of the observed values from the data were normally distributed.

Keywords*:- COVID -19, Regression, Global COVID-19, Conformed Cases, Recovered Cases.*

I. INTRODUCTION

The Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV- 2), formerly known as the 2019 novel Coronavirus (2019-nCoV), is a newly emerging zoonotic agent that appeared in December 2019 and causes the Coronavirus Disease 2019 (COVID-19).

During the month of December 2019 this stated COVID-19 found in Wuhan city, China. This virus found in many of the pneumonia cases received in the southern part of China. The report gathered from the Municipal Health and Health commission quickly initiating the search regarding south China seafood city where twenty-seven cases were identified. On January 2020, China officially reported that outbreak of COVID-19 to the public where the cases found at Wuhan City (ECDC,2020) declared that " On 9 January 2020, China CDC reported that a novel coronavirus (SARA-CoV-2) had been detected as the causative agent for 15 out of the 59 pneumonia cases. On 10 January 2020, the primary novel coronavirus genome sequence was made publicly available. The sequence was deposited in the GenBank database.

The COVID-19 pandemic, which has already infected nearby 170 thousand people in 148 countries, leading to reach 6,500 deaths, has the potential to succeed in an outsized proportion of the worldwide population.

A. Data

The data is collected from internet source and it is a secondary data used for analysis on COVID impact on Global wide, the cases, death and recovered information were tabulated in Table 1.

Items	**Number of cases**		
Cases	4,648,566,		
Deaths	309001		
Recovered	1771074		
Currently infected	Mild (%)	Serious (%)	Total (%)
	45054(2)	2,523,437(98)	2568491(100)
Closed cases	Recovered or Discharged(%)	Death (%)	Total (%)
	1771074(85)	309001(15)	2080075(100)

Table 1:- Distribution of Cases found as on 16 May 2020
Source: Worldometer's COVID-19 data.(Thanks to Worldometer)

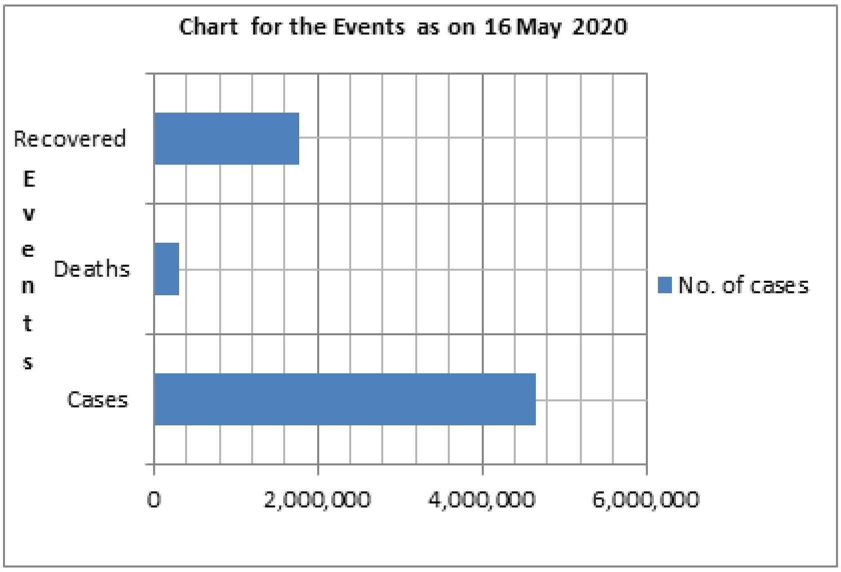

Fig 1:- Distribution of Events(Cases, Deaths and Recovered) as on 16 May 2020

The data used in this study is taken from Internet sources as on the date of 16, May 2020. At that time the Cases reached 4,648,566, deaths 309001, recovered 1771074, Currently Infected Patients 2568491 out of this 98 %(2,523,437) were mild condition and rest 2% (45054) were serious or critical and the closed cases were reached 2080075 out of this15% (309001) were death cases and 85% (1771074) were recovered or discharged (Table 1). The data includes that of two hundred and fifteen countries as it includes the coronavirus is cause of 213 countries and territories around the world and 2 international conveyances. The list of nations and territories and their continental regional classification is predicated on the United Nations records . The Source of data was collected from Worldometer's COVID-19 data. After observing the collected data, there were some no recovered cases found around in 5 countries and these countries were removed from the data and analysis made on the 210 countries(215 total countries- 5 removed= 210 the analysis made countries)

B. *Objective:*

The objective of this article is to find an equation on current COVID-19 with total case as independent and recovered cases as dependent on this data.

➢ *Specific objectives*

- To find the correlation between conformed and recovered cases
- To establish a model based on involved variables

II. METHOD FOR MODELLING

To carry the analysis, the straightforward bivariate linear model Yi = α + βXi + εi, there are four possible combinations of transformations involving logarithms: the linear case with no transformations, the linear-log model, the log-linear model, and the log-log model. In this study, the authors proposed the Log-Linear model then the model will be expressed as;

$$\log\hat{Y}_i = \alpha + \beta X_i$$

It to remember that authors are using natural logarithms. The reason to use logarithmically transforming variables in a regression model is to handle the scenario when a non-linear relationship exists between the independent and dependent variables. The logarithm will help to get an effective non-linear relationship within the linear model rather than the un-logged form in the model. Logarithmic transformations are an easy and simple means of regenerate the modeling as a highly skewed variable as the more approximately normal.

The contemporary study is undertaken to analyses the correlation between the COVID -19 total and recovered cases from the data (dated 16 May 2020) gathered from the website of Worldometer's COVID-19 for the Global data. This study involves the Simple Regression Linear Models to analyses the correlation between the COVID - 19 Total and Recovered cases in the above mentioned period. To analyses the correlation between the COVID - 19 total and recovered cases for the global data, the study used the following equation: Ln(y) = α + βx + ε Where, ln(y) is the natural log of Dependent Variable, x is COVID 19 total cases and α is constant, and β, is the coefficient parameter of the given model. The reason to use Log-linear regression is that the number of cases is large. Hence LLM makes use of large sample approximations, it requires large samples.

- *Variables*

Variables	Dependent/Independent	Variables code
Total cases (Total conformed cases)	Intendent (Predictor)	Total cases
Recovered cases	Dependent	Total recovered

Table 2:- Variable involved in the COVID-19 data

As we us the Log linear regression model, in this study it involved with two variables the variable total cases as a predictor and the recovered cases as dependent variable as it explained in Table 2.

III. DATA ANALYSIS

A. *Basic Statistics*

Basic or descriptive statistics are used to explain the basic idea about the collected data and we can't get any conclusion, here in this study the mean and SD are used to explain the central and dispersion measure of statistics for the given data.

Variables	Mean	SD	Total
Cases	20795.46	109570.3	210
Recovered	5.8771	2.72	210
Pearson correlation Coefficient(p Value)	0.877 (0.00) (P<0.05) Significance at 5%		

Table 3:- Descriptive statistics

From Table 3, out of 210 countries, the total number of conformed cases of COVID-19 with the mean and Standard deviation as 20795.46 and 109570.3 respectively. And the other variable as involved in the study is Recovered cases after natural log as on May 2020 with a mean and standard deviation of 5.8771 and 2.72 respectively.

B. *Regression Analysis*

To check the relation between the total cases and recovered cases by using the Pearson's correlation the relation between the involved variables gives the coefficient of 0.875 as we may conclude that there is a high positive correlation between the cases and recovered cases also the coefficient shows the significance at 5% level of significance.

Measure	Values			
R	0.877			
R Square	0.769			
Adjusted R square	0.768			
Standard Error of the Estimate	52624.183			
Change Statistics	**R Square change**	**F change**	**Degrees of freedom (df1, df2)**	**Significance F change**
	0.769	710.492	1,208	0.000

Table 4:- Model summary

Table 4, the output of Model summary and overall fit statistics displayed and from this table ,the Adjusted R Square of this study model is 0.768 with the R square is 0.768, this implies that the given model explains 76.7 percent of the variance in the data. Also from R square, 76.8% can be explained and which is very large explained. The F change value is 710.492 with 1, 208 degrees of freedom, and also shows the evidence to a significantly differ at a 5% level.

Model	Sum of Squares	Degrees of freedom (df)	Mean Square	F	Significance
Regression	217.911	1	217.911	33.958	0.000
Residual	1334.740	208	6.417	---	-----
Total	1552.652	209	----	---	----

Table 5:- Analysis of Variance

Table 5 gives an analysis of ANOVA (Regression) ,this table indicates that the model predicts the dependent variable significantly well. From the table 5, it is observed that the F= 33.958 with a P value of 0.000 (*p* < 0.000, which is less than 0.05). This indicates the statistical significance of the give model is a good fit for the data. Also the overall interpretation is the model statistically significantly predicts the outcome variable.

Model 1	Coefficients			t	Sig.	95% confidence interval for B (Bounds)		Collinearity statistics	
	Un standardized		Standardized						
	B	SE	Beta			Lower	Upper	Tolerance	VIF
(Constant)	5.683	.178		31.939	0.00	5.333	6.034		
Total cases	0.0000096	0.00	375	5.827	0.00	0.00	0.00	1.000	1.000

Table 6:- Coefficients

From table 6, it provides the necessary information to predict cases from COVID-19, as well as determine whether total cases contribute statistically significantly to the given model. Furthermore, we can use the values in the coefficient table under the "Unstandardized Coefficients" column, as shown in table 4. Here α = 5.683 and β= 0.0000093 of the given model. Based on these coefficients, now it is easy to present the regression equation as Ln (Recover) = 5.683 + 0.0000093(total cases). These estimated coefficients gives the connection between the independent variables and dependent variable as the quantity of increase in cases that might be predicted by a 1 unit increase within the predictor. Thus this regression model suggests that as total cases of COVID-19 increase the recovered cases of COVID-19 are also increase, with p = 0.000 (which is marginally significant at alpha=0.05). More precisely, it says that for one Case of the total increase in average case of recovered, the predicted Recover cases increase by 0.0000093 points holding the percent of total cases constant.

Model1 Dimension	Eigenvalue	Condition Index	Variance Proportions	
			(Constant)	Total cases
1	1.187	1.000	0.41	0.41
2	0.813	1.208	0.59	0.59

Table 7:- Collinearity Diagnostics

By using collinearity diagnostics one can confirm that there are serious problems with Multicollinearity. Thus the table 7 deals with diagnostics, the first column show the dimensions as in our study we have dimensions as 2 and it is equivalent to a factor analysis, to make an attempt on determine dements with independent information. Here there were two dimensions.

The next column of the collinearity diagnostics table is Eigen value, the value close to zero implies that the predictors are high correlated and that small changes in the data values may lead to large changes in the estimated of the coefficients. Here the second dimension is near to the value zero. Since “near to” is somewhat imprecise it’s better to the condition index for the diagnosis. The condition implies are computed because the square roots of the ratios of the most important eigenvalue to every successive eigenvalue. Less than the value of 15 indicates a possible problem with non-collinearity. In this study the table shows that both the dimensions are less than 15 and it implies that, a possible problem with non-collinearity.

Statistics(Residuals)	Minimum	Maximum	Mean	Standard deviation	N
Predicted value	5.6833	19.5183	5.8771	1.02110	210
Std predicted value	-0.190	13.359	0.000	1.000	210
SE of predicted value	0.175	2.347	0.194	0.153	210
Adjusted predicted value	5.6860	60.9506	6.0722	3.82289	210
Residual	-6.81824	4.81102	0.00000	2.52712	210
Std. Residual	-2.62	1.899	0.000	0.998	210
Stud. Residual	-7.160	1.905	-0.021	1.101	210
Deleted residual	-48.25054	4.84153	-0.19507	4.16443	210
Stud deleted residual	-8.229	1.917	-0.026	1.138	210
Mahal. Distance	0.000	178.471	0.995	12.322	210
Cook’s Distance	0.000	155.768	0.745	10.749	210
Cenered Leverage Value	0.000	0.854	0.005	0.059	210

Table 8:- Residuals statistics

In table 8, shows the residual statistics, summarise the nature of the residuals and predicted values in the given model. It is important to viewing it so you can get a better understanding of the spread of values that the model predicts and the range of error within the model. As the maximum value of Cook’s distance in our sample is 155.768 which far more than the value of 1 which may be a cause for concern. Thus there is no any problem arises in our data .

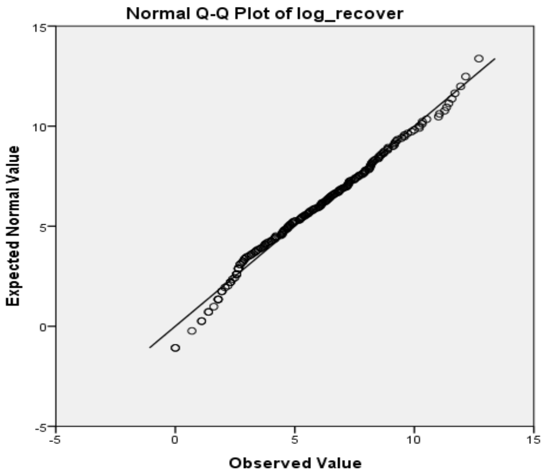

Fig 2:- Normal Q-Q Plot

Figure 2, A Normal Q-Q (or Quantile-Quantile) Plot compares the observed quantiles of the data with dots and with the quantiles that the expectation will be about existence of Normality by using solid lines . Here the given data is approximately normally distributed because the points were close to the solid line.

IV. CONCLUSION

This study is based on secondary data collected from Internet source, there 215 countries available and 5 countries not having any recovery record on the 16 May 2020, here the analysis were involved with 210 countries only. The predictor variable is total cases and the recovered case as considering independent variable for this analysis. This study use simple log- linear egression and the results reveales that there is a significant high positive correlation between the COVID - 19 total cases and recovered cases on 16 May 2020 in the world. Also this study reveals that the regression Model(LLM) as total cases of COVID-19 increases the Recovered cases of COVID-19 is also increase, with $p = 0.000$ (which is marginally significant at alpha=0.05). More precisely, it says that for a one Case of total increase in average case of recovered, the predicted Recover cases increases by 0.0000093 points holding the percent of total cases constant. The Q-Q plots shows, the given data is approximately normally distributed because the points were close to the solid line.

ACKNOWLEDGMENTS

The authors of this article wish to express their great thanks and regards to the website owners of Worldometers , for utilizing their data as on the reference dated data.

REFERENCES

[1]. ACHA Guidelines Preparing for COVID-19(2020) ACHA Guidelines

[2]. Data source https://www.worldometers.info/coronavirus/

[3]. Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2013). *Multivariate data analysis: Advanced diagnostics for multiple regression* [Online supplement] Retrieved from http://www.mvstats.com/Downloads/Supplements/Advanced_Regression_Diagnostics.pdf

[4]. IBM (n.d.). *Collinearity diagnostics*. Retrieved August 19, 2019, from https://www.ibm.com/sup

[5]. Snee, R. D. (1983). Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. *Journal of Quality Technology, 15*, 149-153. doi:10.1080/00224065.1983.11978865

[6]. WHO (2020) Coronavirus disease (COVID-19) Situation Report – 116, Data as received by WHO from national authorities by 10:00 CEST, 15 May 2020

[7]. Richard Baldwin et al., (2020) Economics in the Time of COVID-19CEPR Press Centre for Economic Policy Research