***COVID-19 Death Prediction Using Economic and Health Factors***

Contents

[**1.** **INTRODUCTION** 2](#_Toc92316625)

[**2.** **DATA ACQUISITION** 3](#_Toc92316626)

[**3.** **DATA EXPLORATION:** 4](#_Toc92316627)

[3.1 Reading the Data 4](#_Toc92316628)

[3.2 Identifying the Variables: 4](#_Toc92316629)

[3.2.1 Highest Deaths: 5](#_Toc92316630)

[3.2.2 Lowest Deaths: 5](#_Toc92316631)

[3.3 Missing Value Analysis: 6](#_Toc92316632)

[3.4 Data Normality: 6](#_Toc92316633)

[3.4.1 Normality Prediction 6](#_Toc92316634)

[3.4.2 Handling Normality: 8](#_Toc92316635)

[3.5 Outlier Handling: 9](#_Toc92316636)

[3.6 Correlation Analysis 10](#_Toc92316637)

[3.4.1. Dependent vs Independent Correlation: 10](#_Toc92316638)

[*3.4.1.b Spearman correlation:* 11](#_Toc92316639)

[3.4.2 Internal Correlation between Independent Variables: 12](#_Toc92316640)

[3.4.3 Partial Correlation: 12](#_Toc92316641)

[4. FACTOR ANALYSIS: 12](#_Toc92316642)

[4.1 First Round of Test - Kaiser-Meyer-Olkin Test: 13](#_Toc92316643)

[4.2.Second Round of test 14](#_Toc92316644)

[**5.** **REGRESSION MODEL:** 14](#_Toc92316645)

[5.1 Simple Linear Regression: 15](#_Toc92316646)

[5.2 Multiple Linear Regression: 16](#_Toc92316647)

[5.3 Linear Regression with Factor Analysis Components: 17](#_Toc92316648)

[5.3 a Partial Correlation: 18](#_Toc92316649)

[5.3 b Other than Principal Component Analysis (PCA) Variables 18](#_Toc92316650)

[5.4 Forward stepwise selection on Other than PCA 18](#_Toc92316651)

[5.5 Forward stepwise selection on Multiple Linear regression. 18](#_Toc92316652)

[**CONCLUSIONS** 19](#_Toc92316653)

[**REFERENCES:** 19](#_Toc92316654)

[**APPENDICES:** 19](#_Toc92316655)

# **INTRODUCTION**

COVID-19 is a universal widespread disease that is ruling the world since the year 2019. Almost every area of the global economy has suffered as an outcome of the COVID Virus. The consequences are particularly intense in terms of human health and economic operations. In addition, a large number of people died as a result of this sickness. So, using data analysis and machine learning schemes, a model is built to understand the pattern of COVID mortality count based on economic activities and the general health of various people living in different locations of the United Kingdom. The Covid Deaths data was used to train a regressor, which was then used to predict the death count based on the Economic and Health factor values provided. The model has also been subjected to various data exploration approaches, factor analysis, and hypothesis testing to improve its effectiveness. So, for a particular area, we can be able to Predict the Covid death count using our model. In this project, I am going to show through data analysis that how the COVID-19 Death is impacted due to the various Economic and Health Factors. In the near future, this model may be used to predict the death count of any other virus as long as economic and health indicators are provided.

# **DATA ACQUISITION**

The Input dataset used for this model is created by joining the COVID-19(Which has the death count of people affected due to COVID-19 based on regions), Economic, living type, and the health datasets (Sourced from Nomis Website) by using the “Geography/LA Code” as the Primary key. Please refer to the SQL Query in the Appendices Section.

Also, the Continuous variables are Standardised by using the Appropriate base Population variable (Per thousand Population). Please find the Abbreviation of the dependent and independent variables used for the model:

|  |  |  |
| --- | --- | --- |
| **Variable Name** | **Abbreviation** | **Data Type** |
| Total\_Deaths (Dependent Variable) | Covid-19 Death Counts | Num |
| Inemp\_FT\_limitedalot\_Dis | **Full-Time** Students Who have Total disability and |
| Also, a **lot of** limitations in Employment |
| Inemp\_FT\_limitedalittle\_Dis | **Full-Time** Students Who have Total disability and |
| Also, **little** limitations in Employment |
| Inemp\_FT\_notlimited\_Dis | **Full-Time** Students Who have Total disability and |
| Also, **no** limitations in Employment |
| Inemp\_PT\_limitedalot\_Dis | **Part-Time** Students Who have Total disability and |
| Also, a **lot of** limitations in Employment |
| Inemp\_PT\_limitedalittle\_Dis | **Part-Time** Students Who have Total disability and |
| Also, **little** limitations in Employment |
| Inemp\_PT\_notlimited\_Dis | **Part-Time** Students Who have Total disability and |
| Also, **no** limitations in Employment |
| Unemp\_limitedalot\_Dis | **Unemployed People** Who have Total disability and |
| Also, a **lot** of limitations in Employment |
| Unemp\_limitedalittle\_Dis | **Unemployed People** Who have Total disability and |
| Also, **little** limitations in Employment |
| Unemp\_notlimited\_Dis | **Unemployed People** Who have Total disability and |
| Also, **no** limitations in Employment |
| verybadHlth\_LTHP\_Age0to15 | People with **Bad** Long Term Health Problem in **Small Age** |
| verybadHlth\_LTHP\_Age16to49 | People with **Bad** Long Term Health Problem in **Teenage** |
| verybadHlth\_LTHP\_Age50to64 | People with **Bad** Long Term Health Problem in **Oldage** |
| verybadHlth\_LTHP\_Age65andover | People with **Bad** Long Term Health Problem in **Pensioners** |
| GenHlthDis\_LTHP\_Age0to15 | People with **General Disability** in **Small Age** and |
| also, Long Term Health Problem |
| GenHlthDis\_LTHP\_Age16to49 | People with General **Disability** in **Teenage** and |
| also, Long Term Health Problem |
| GenHlthDis\_LTHP\_Age50to64 | People with **General Disability** in **Oldage** and |
| also, Long Term Health |
| GenHlthDis\_LTHP\_Age65andover | People with **General Disability** in **Pensioners** and an also Long-Term Health |
| Single\_Living | Living alone in a Single Room |
| Partner\_Living | Living With Husband/Wife in a house |
| Joint\_Living | Living with Joint family members. |
| LA\_Code | Region Code | Char |
| LA\_name | Region Name |

Table 2.1 Variable Abbreviation and Type

# **DATA EXPLORATION:**

As part of data exploration in the data frame, the below steps have been executed:

## 3.1 Reading the Data

As the first step of data Exploration, the input dataset has been imported in the R studio using the read\_csv command. While importing the data, **‘file-encoding = 'UTF-8-BOM'** has to be used otherwise the first Column LA\_CODE will be prefixed with byte Order Mark (^….) and it will be difficult to use the variable name in such a way. Also using the **name. strings = '..'** keyword while importing fills the blank values if any with NA values. These are some good practices that can be followed for building a better model. Now, the data is available in the format of a data frame named COVID\_PRED for further analysis

## 3.2 Identifying the Variables:

The command **str(df)** is used to identify the internal structure of the COVID\_PRED data frame. After reviewing the variables, we have taken Total Deaths as the dependent/target variable and the other variables as the Independent/Predictor variables. All the variables are numeric datatype except LA\_Code and LA\_name fields which are characters. Also, our data frame size is 323\*23 which is 323 rows and 23 columns.

We could see that the character fields LA\_Code and LA\_name are just depicting the place where the death has happened. They will not impact much in the variance of the dependent variable. So, we will just try to take all the other variables except the character variables in a separate data frame named COVID\_PRED2. So, we can use this data frame for our further analysis of data.

Meanwhile, we can just infer the region which has the most COVID deaths based on the variables LA\_Name and LA\_code as follows:

### 3.2.1 Highest Deaths:

From the Covid\_Pred data frame, we have extracted the required variables LA\_code, LA\_name, and Total\_Deaths in a separate data frame named Covid\_reg. Then the death value is sorted in descending order so that the top death region comes first and the bottom death region goes at the bottom. Then the function head (Covid\_reg,n=2) is applied to find the top 2 regions. Please find the results below:

|  |
| --- |
| LA\_code LA\_name Total\_Deaths  E07000193 East Riding of Yorkshire 6.59  E07000076 Tendring 4.17 |

Table 3.2.1 Highest Death regions

From table 3.2.1, it is clear that the region “East riding of Yorkshire” and “Tendring” has the highest COVID deaths recorded. We can assume that the people living in these regions have weak immune systems or there may be many old and sick people. So, the government can focus first on these regions for providing safety by applying lockdown in the near future in case of any other virus variants.

### 3.2.2 Lowest Deaths:

Then the function tail (Covid\_reg,n=2) is applied to find the bottom 2 regions. Please find the results below:

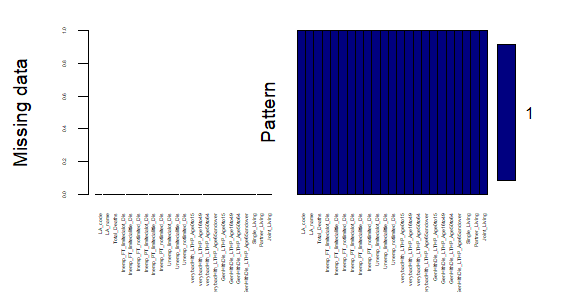
|  |
| --- |
| LA\_code LA\_name Total\_Deaths  E07000044 South Hams 0.54  E06000053 Isles of Scilly 0 |

Table 3.2.2 Lowest Death regions

From table 3.2.2, it is clear that the region “South Hams” and “Isles of Scilly” has the lowest COVID deaths recorded. We can assume that the people living in these areas are healthy. Also, we suspect that the region “Isles of Scilly” may be an outlier because of the Death Count value of Zero. There are chances that the death count is not recorded in that particular region. We can make sure while checking for outlier values. For these regions, the government can give the least importance to providing safety during the tough times in the near future as the other regions may be considered.

## 3.3 Missing Value Analysis:

The missing value is one of the main obstacles in the data frame that needs to be handled. So, we have used the **library VIM** Which creates a visualization graph projecting the percentage of missing values. Graph 3.3 shows the variables in the X-axis and the Count of the Null values in the y-axis. we could see that there are 100% no missing values in the given data frame. So, we can go ahead without performing any imputations in the data frame columns.



Graph 3.3 Missing values in all Variables

## 3.4 Data Normality:

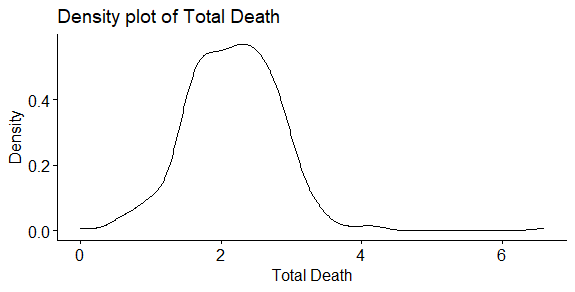
Normality is a concept that if we plot our data in form of a graph, the data points should be distributed evenly in form of a bell-shaped curve. This is one of the Assumptions that need to be checked while applying a regression model in our dataset.

### 3.4.1 Normality Prediction

There are many methods like QQ plot, Density plot, Box Plot, Shapiro test to predict whether the data is normal or not. We will look at how our data reacts to various tests in the following sections

#### Density Plot:

Density plot is one such method. If the density of the curve is in a bell shape, we can assume that the data is normally distributed. Now let us apply the density plot function in our dependent variable.



Graph 3.4 Density plot

#### b. Shapiro Test:

Even though we conclude that the data is not normal through density plot, we can confirm it through the Shapiro test as well. As the P-value is less than 0.05 the data is not normal.

|  |
| --- |
| #Normality using the test:  Shapiro.test(Covid\_Pred$Total\_Deaths) |

|  |
| --- |
| data: Covid\_Pred$Total\_Deaths  W = 0.95447, p-value = 1.826e-08 |

#### Boxplots :

Boxplots show whether the variables in a data frame fall in the same range. i.e., Each variable is picturized in form of boxes with a median in the middle, and also the 25th percentile and 75th percentile data will be plotted through lines. Also, the outlier values will be shown. Now let us see how the data frame Covid\_Pred2 looks:

We could see that the data frame is not in the range. Because the range values in the y-axis are not even. So, we have to normalize the data to continue our model.

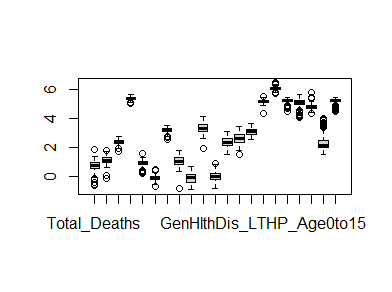
### 3.4.2 Handling Normality:

Normality of the data can be handled in many ways like Standardisation, normalization log transformations, and so on. This is one of the critical steps in exploring the data because abnormal data will result in creating a very bad model. To avoid that, we need to normalize the data according to the data points related to the data

#### a. Standardisation and Normalisation:

I tried applying the standardization technique by subtracting each value in the variable by mean and dividing it by standard deviation. But then the boxplot din gets altered much. Also, I tried applying the normalization by using the minimum and maximum value but it did not give any impact. Standardization and normalization piece of code has been attached in the R script.

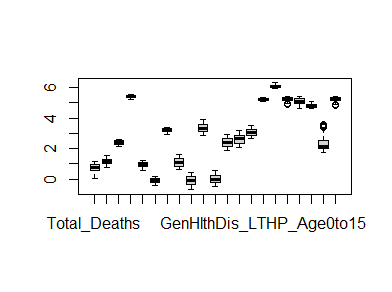
#### b. Log transformations:

 Log transformation is one of the techniques where the data points are replaced by their log value of them. So, on applying the log transformations, let us see how our boxplot looks in the below image:

3.4.2bLog Transformation Boxplot

We could see that the data has improved a lot but then there are outlier values. We will handle it in the below section.

## 3.5 Outlier Handling:

 Outlier is a variable whose value is not within the range. Linear regression is sensitive to outliers. Including those variables in the analysis will reduce the model’s efficiency. From image 3.4.2b it’s clear that we have outlier values in most of the variables. So, now removing them will affect the number of rows taken into consideration. As an alternate, we can impute the outlier values.

3.5 Outlier

We could see that most of the outlier values have been replaced with the 5th percentile and 95th percentile values. Comparative to image 3.4.2b, this image has improved a lot. Please find the code below:

|  |
| --- |
| #Outlier  Out1 <- function(x){  for (i in which(sapply(x, is.numeric))) {  quantiles <- quantile( x[,i], c(.05, .95 ), na.rm =TRUE)  x[,i] = ifelse(x[,i] < quantiles[1] , quantiles[1], x[,i])  x[,i] = ifelse(x[,i] > quantiles[2] , quantiles[2], x[,i])}  x}  # Replacing extreme values with percentiles  Covid\_out = out1(Covid\_log)  boxplot(Covid\_out) |

Function out1 has been created where the outlier value has been identified and replaced the lower outlier with the 5th percentile and the upper outlier with the 95th percentile. Here we pass the data frame with log value as input and generate a new data frame named Covid\_out.

## 3.6 Correlation Analysis

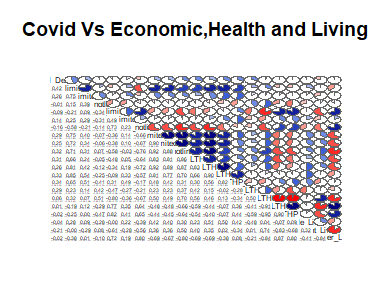
Finding the Correlation between the variables is very important because training the model with irrelevant variables will reduce the model’s efficiency. So, the below techniques have been applied to build an optimized model.

### 3.4.1. Dependent vs Independent Correlation:

Correlation can be calculated between the independent and dependent variables and also within the independent variable. The strongest variable is taken into consideration for further analysis. Usually, the Strongest Positive correlation value will be greater than 0.5. and strongest negative correlation value will be >-0.5

#### 3.4.1a Corrgram

The corrgram is a graph that displays the Correlation values between the dependent and the independent variables. Dark blue denotes the strong correlation and Red denotes the weak correlation. We have taken the data frame Covid\_out and applied the corrgram function.



Graph 3.4.1 a Corrgram

|  |  |
| --- | --- |
| **Corrgram Variables** | **Total\_Death Correlation** |
| Inemp\_FT\_limitedalot\_Dis | 0.42 |
| verybadHlth\_LTHP\_Age65andover | 0.34 |
| verybadHlth\_LTHP\_Age50to64 | 0.33 |
| Unemp\_limitedalot\_Dis | 0.29 |
| Unemp\_notlimited\_Dis | 0.32 |

Table 3.4.1 b Corrgram Correlation values

Graph 3.4.1 shows the Pictorial representation and table 3.3.b shows the Variables holding the top five Correlation values. As we don’t have any variables greater than 0.5, we are taking the top value variable. We could see that the variables Inemp\_FT\_limitedalot\_Dis, verybadHlth\_LTHP\_Age65andover, and verybadHlth\_LTHP\_Age50to64 are having a strong correlation with the dependent variable.

We will apply spearman as well to make sure that the correlation results are consistent.

#### 3.4.1.b Spearman correlation:

Spearman test is also applied on the COVID\_out data frame and the results are projected in table 3.4.1.b.a. The variables with strong correlation have been highlighted.

|  |  |  |  |
| --- | --- | --- | --- |
| **Spearman dep vs indep** | **Value** | **Correlation** | **Correlation Result** |
| Inemp\_FT\_limitedalot\_Dis | 6.13E-16 | 0.4296123 | Strong Correlation |
| Inemp\_FT\_limitedalittle\_Dis | 7.81E-07 | 0.2730945 | Weak Correlation |
| Unemp\_limitedalot\_Dis | 8.134E-08 | 0.2948382 | Weak Correlation |
| Unemp\_limitedalittle\_Dis | 1.12E-06 | 0.2670125 | Weak Correlation |
| Unemp\_notlimited\_Dis | 7.03E-09 | 0.3159303 | Strong Correlation |
| verybadHlth\_LTHP\_Age0to15 | 3.68E-09 | 0.3202068 | Strong Correlation |
| verybadHlth\_LTHP\_Age16to49 | 1.94E-07 | 0.2842861 | Weak Correlation |
| verybadHlth\_LTHP\_Age50to64 | 6.28E-11 | 0.3539635 | Strong Correlation |
| verybadHlth\_LTHP\_Age65andover | 7.66E-13 | 0.3848423 | Strong Correlation |
| GenHlthDis\_LTHP\_Age0to15 | 1.04E-06 | 0.2423963 | Weak Correlation |

Table 3.4.1.b Spearman – Dependent vs Independent

We could see that the variables Inemp\_FT\_limitedalot\_Dis, verybadHlth\_LTHP\_Age50to64, and verybadHlth\_LTHP\_Age65andover have a great relationship with COVID\_Death count.

#### 3.4.1.c Hypothesis Testing:

Here the Hypothesis is “Whether the independent variable and the dependent variable are related or not”. So, the null hypothesis is “There is no remarkable connection between the variables” and the alternate hypothesis is “There is a remarkable connection between the variables”.

So Based on the above hypothesis rule, we can state that the highlighted variables in table 3.4.1.b are related to the Covid death as the P-value is less than 0.05. i.e., We are rejecting the null hypothesis (Rejecting that they are not related. Accepting that the variables are related).

### 3.4.2 Internal Correlation between Independent Variables:

Now, we will find the internal correlation between the variables selected from the Spearman correlation.

|  |  |  |
| --- | --- | --- |
| **Spearman Internal Correlation** | Unemp\_notlimited\_Dis | verybadHlth\_LTHP\_Age65andover |
| Inemp\_FT\_limitedalot\_Dis | 2.29e-10 P-value  0.3443984 correlation | 2.2e-16 is P-value 0.5076628 Correlation |
| Unemp\_notlimited\_Dis | - | 2.768e-08 P value 0.3029954 Correlation |

Table 3.4.2. Internal correlation

The variable verybadHlth\_LTHP\_Age65andover has a good correlation with Inemp\_FT\_limitedalot\_Dis and Unemp\_notlimited\_Dis and also the P-value is less than 0.05 so we can accept the alternate hypothesis by assuming that the variable verybadHlth\_LTHP\_Age65andover is significant.

Now we need to decide on selecting either the Inemp\_FT\_limitedalot\_Dis or Unemp\_notlimited\_Dis variable so we need to do a Partial correlation by keeping Unemp\_notlimited\_Dis as the Control variable and again Inemp\_FT\_limitedalot\_Dis as the Control variable.

### 3.4.3 Partial Correlation:

We are calculating the Partial correlation between the variables using the Pearson method. Please find the results below:

Unemp\_notlimited\_Dis as Control Variable:

|  |
| --- |
| estimate p.value statistic n gp Method  1 0.03384232 0.5451189 0.6057369 323 1 pearson |

Inemp\_FT\_limitedalot\_Dis as Control variable:

|  |
| --- |
| estimate p.value statistic n gp Method  1 0.2885458 1.363136e-07 5.390963 323 1 pearson |

As the p-value is tremendous in the first case, we are not going to consider the variable Unemp\_notlimited\_Dis and at the same time, P-value is less than 0.05 in the second case so we are going to consider the variable Inemp\_FT\_limitedalot\_Dis.

# FACTOR ANALYSIS:

Factor analysis is a technique used in machine learning models to find the number of essential components that need to be applied in the model to make it a robust one.

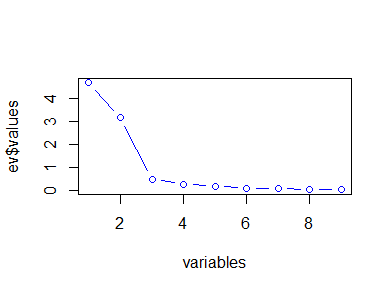
## First Round of Test - Kaiser-Meyer-Olkin Test:

The Kaiser-Meyer-Olkin statistics (KMO test) is used to check whether the variables that are available as part of the data frame are sufficient to carry out the factor analysis process. If the value is greater than 0.8, we are good to proceed with the model with the fact that the data is sufficient.

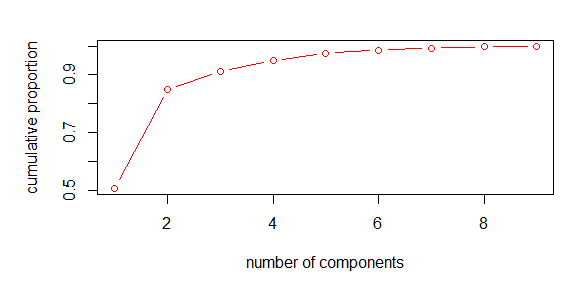
So, we are applying the test to all the variables available in the data frame. Please refer to the result below:

|  |
| --- |
| MSA for each item =  Inemp\_FT\_limitedalot\_Dis Inemp\_PT\_limitedalot\_Dis  0.87 0.89  Unemp\_limitedalot\_Dis Unemp\_notlimited\_Dis  0.76 0.81  verybadHlth\_LTHP\_Age50to64 verybadHlth\_LTHP\_Age65andover  0.70 0.59  GenHlthDis\_LTHP\_Age50to64 GenHlthDis\_LTHP\_Age65andover  0.74 0.66  Joint\_Living 0.73 |

Almost all the variables have an MSA value greater than 0.70 except verybadHlth\_LTHP\_Age65andover. So, we can proceed with the remaining variables for further analysis.

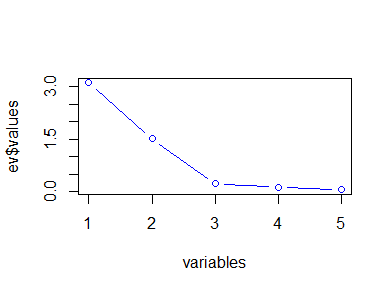
We have also used the nFactors library to determine the number of components to consider. The scree plot is a plot while considering the relation between the components and their eigenvalues. So, based on the graph’s data point where it becomes stable, we can select the number of components. Please find the scree plot below:

The number of components that can be considered is 3. Using Eigenvalues and correlation matrix, we have the below output in factor analysis:



As we could see that the 90% variance is covered in 3 factors. We are considering only 3 factors to the model.

## 4.2. Second Round of test

 Now we can weed out some factors which has low MSA value and continue our KMO test and Eigen Plot and decide the number of factors to consider.

Now we could see that the number of components should be 3 that needs to be applied to the model.

# **REGRESSION MODEL:**

Linear regression is a machine learning model that is widely used by many engineers for predicting the continuous variables. Our problem statement is predicting the death count of COVID. So, we have implemented the Linear regression in different forms and observed their performance, and found the best model to predict the dependent variable.

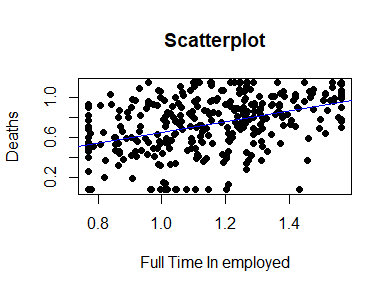
## 5.1 Simple Linear Regression:

Simple Linear regression is a method in which only one significant variable is fitted to the model. But this model is not advisable because it will lead to the overfitting of the data. Also, in real life, we cannot rely on only one factor for predicting the variable. We can use the topmost significant variable and see how the model works.

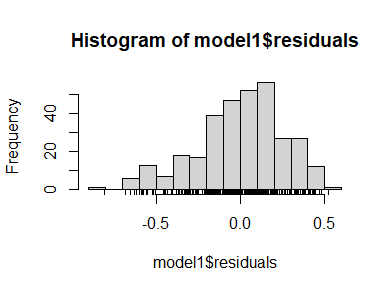
Please find the model results below:

|  |
| --- |
| Residual standard error: 0.2586 on 321 degrees of freedom  Multiple R-squared: 0.174, Adjusted R-squared: 0.1714  F-statistic: 67.63 on 1 and 321 DF, p-value: 4.962e-15 |

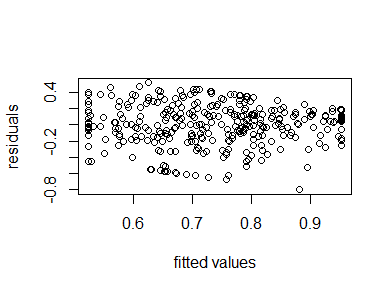
**Residual standard error** describes the standard deviation of the error points. The **minimum** value makes the better model. Model 1 gives minimum error.

**R-squared** shows the extent to which the input variable explains the output variables. **A value near 1** makes the better model. But this model has a low R-squared value so this model is not acceptable.

The above graph shows how well the data is fit in the line. But the model did not fit all the points.

**Histogram:**

**Scatter plot for Residuals:**



## 5.2 Multiple Linear Regression:

Multiple Linear Regression is a model in which instead of using a single variable. We will be using multiple variables. i.e., all variables of the model. So, we have implemented it.

Please find the model results below:

|  |
| --- |
| Residual standard error: 0.2251 on 302 degrees of freedom  Multiple R-squared: 0.4111, Adjusted R-squared: 0.3721  F-statistic: 10.54 on 20 and 302 DF, p-value: < 2.2e-16 |

This model has a good R-squared value and also the Residual error is less as well. P-value is less than 0.05 so we can accept the alternate hypothesis. Here the hypothesis is “Variation in the independent variable is affecting the dependent variable”. The null hypothesis is that there is no relation and the alternate is that variation is affecting the dependent variable. So, in this case, we are accepting the alternate hypothesis that the change in the independent variable is affecting the dependent variable.

**KS test:**

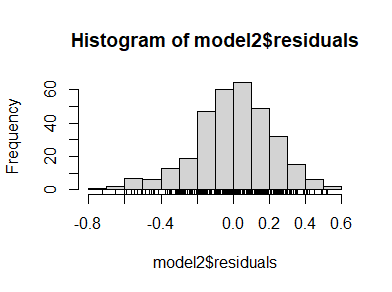
|  |
| --- |
| data: model2$residuals  D = 0.047805, p-value = 0.4515  alternative hypothesis: two-sided |

**VIF Results:**

|  |
| --- |
| Covid\_out$Inemp\_FT\_limitedalot\_Dis Covid\_out$Inemp\_FT\_limitedalittle\_Dis  12.039626 5.504115  Covid\_out$Inemp\_FT\_notlimited\_Dis Covid\_out$Inemp\_PT\_limitedalot\_Dis  6.636110 5.835728  Covid\_out$Inemp\_PT\_limitedalittle\_Dis Covid\_out$Inemp\_PT\_notlimited\_Dis  1.850936 7.516503  Covid\_out$Unemp\_limitedalot\_Dis Covid\_out$Unemp\_limitedalittle\_Dis  17.717357 7.205873  Covid\_out$Unemp\_notlimited\_Dis Covid\_out$verybadHlth\_LTHP\_Age0to15  23.978501 6.388388  Covid\_out$verybadHlth\_LTHP\_Age16to49 Covid\_out$verybadHlth\_LTHP\_Age50to64  31.681599 36.538238  Covid\_out$verybadHlth\_LTHP\_Age65andover Covid\_out$GenHlthDis\_LTHP\_Age0to15  19.353383 4.564581  Covid\_out$GenHlthDis\_LTHP\_Age16to49 Covid\_out$GenHlthDis\_LTHP\_Age50to64  30.495142 17.406206  Covid\_out$GenHlthDis\_LTHP\_Age65andover Covid\_out$Single\_Living  38.263069 4.245644  Covid\_out$Partner\_Living Covid\_out$Joint\_Living  24.943855 10.489707 |

Variance Inflation Factor finds the multicollinearity of data. The greater value affects the data.

Histogram:



## 5.3 Linear Regression with Factor Analysis Components:

Here, we have implemented the multiple regression model with the factor analysis components alone.

Please find the model results below:

|  |
| --- |
| Residual standard error: 0.2542 on 318 degrees of freedom  Multiple R-squared: 0.2093, Adjusted R-squared: 0.1994  F-statistic: 21.05 on 4 and 318 DF, p-value: 2.07e-15 |

We could see that this model is not good as it has a low R-squared value.

## 5.3 a With Partial Correlation components:

Here, we have implemented the multiple regression model with the results of partial correlation components.

Please find the model results below:

|  |
| --- |
| Residual standard error: 0.2577 on 319 degrees of freedom  Multiple R-squared: 0.1844, Adjusted R-squared: 0.1767  F-statistic: 24.04 on 3 and 319 DF, p-value: 4.709e-14 |

We could see that this model is not good as it has a low R-squared value.

## 5.3 b Other than Principal Component Analysis (PCA) Variables

Here, we have implemented the multiple regression model with components other than the ones selected in model 5.3.

Please find the model results below:

|  |
| --- |
| Residual standard error: 0.2293 on 305 degrees of freedom  Multiple R-squared: 0.3828, Adjusted R-squared: 0.3485  F-statistic: 11.13 on 17 and 305 DF, p-value: < 2.2e-16 |

Comparatively this model is good. This ranks second to the multiple linear regression model.

## 5.4 Forward stepwise selection on Other than PCA

Forward stepwise selection is a process where the components are added one by one in a step-wise manner. Here we have taken the components which are not used in PCA. Please find the model results below:

|  |
| --- |
| Residual standard error: 0.238 on 319 degrees of freedom  Multiple R-squared: 0.3046, Adjusted R-squared: 0.298  F-statistic: 46.57 on 3 and 319 DF, p-value: < 2.2e-16 |

This also performed well.

5.5 Forward stepwise selection on Multiple Linear regression.

We have implemented Forward stepwise but with the multiple regression model. Please find the model results below:

|  |
| --- |
| Residual standard error: 0.2353 on 317 degrees of freedom  Multiple R-squared: 0.3247, Adjusted R-squared: 0.314  F-statistic: 30.48 on 5 and 317 DF, p-value: < 2.2e-16 |

This model also performed well.

**Inference:**

Finally, we can infer that the Multiple Regression model is the best robust model as it has satisfied the assumptions of linear regression.

# **CONCLUSIONS**

In this paper, the Covid death prediction model has been trained with various regression models using the Economic and Health factors of people located in different regions of the United Kingdom (UK) and finally could see that the multiple Regression model with all the variables gives the best results. Also, Full-time employee is mostly affected by COVID Deaths as they are very much exposed to the Publics. At the same time, the aged people are affected next to them due to their weak immune systems. so, in the near future, we can concentrate on these people first to avoid deaths. We can further optimize the model by collecting more data and variables which contribute to the Covid deaths. Also, this model can be used in the prediction of other virus variants as well.

# **REFERENCES:**

1. Prediction of COVID-19 severity using laboratory findings on admission: informative values, thresholds, ML model performance Yauhen Statsenko,1 Fatmah Al Zahmi,2,3 Tetiana Habuza,4 Klaus Neidl-Van Gorkom,1 Nazar Zaki

# **APPENDICES:**



# **R Code:**

