Population data analysis in Northern Europe

Pranav Coimbatore
Dept. of Statistics
George Mason University
Fairfax, United States of America
pcoimbat@gmu.edu

***Abstract*—This paper aims to analyze the factors that influence the total population in the northern European region. These factors include birth rate, mortality rate, life expectancy, infant mortality rate, children per woman, growth rate which includes migratory population and natural increase or decrease in annual population. It is interesting to analyze the change in population and to visualize population data.**

***Keywords*—Birth rate, Mortality rate, Life expectancy, infant mortality rate, growth rate**

I. Introduction

The data set was sourced from, The French Institute for Demographic Studies or INED, a public research institute specialized in population studies[1]. The dataset can be downloaded in the form of a CSV file. The dataset contains data about total population estimations for the year 2021 for each country from the northern European region. The dataset has other factors like birth rate which gives the number of births per 1000 population, mortality rate which gives the number of deaths per year per 1000 population, life expectancy which gives the average length of life in years, infant mortality rate which gives the number of deaths of children under the age of 1 per 1000 years, number of children per woman, growth rate which gives the annual increase or decrease in the number of residents per 1000 population, and lastly the number of people aged 65 or above. The research questions that this study aims to answer is to find the factors that influence the total population in northern Europe. The study also aims to visualize the population distribution in each country by using choropleth maps.

II. METHODS

A. *Data Exploration and Vizualization*

Once the dataset was loaded into R, the data had to be cleaned, there were blank rows after the column headers which was removed, the last two rows in the dataset had the source of the dataset and they were removed as well. The column names were changed for simplicity. The columns were of character data type, and they were converted to numeric data type.

Then the dataset was ready for data analysis. Initially univariate plots were plotted to find out whether any two variables were correlated. The total population estimations were plotted on a choropleth map with an element of interactivity where one hovers over each country, the total population in thousands would be displayed. The choropleth map was constructed using the 'plotly' package[2][3]. From the choropleth map on the total population estimations, it was found that United Kingdom had the highest population.

B. *Multiple Linear Regression*

To find out the factors that influence the total population estimations, multiple linear regression was performed[5]. The idea was to find the best fit model to find out the variables that influence the total population estimations. There were seven variables in the dataset which were birth rate, mortality rate, infant mortality rate, life expectancy, children per woman, growth rate, and population of people aged over 65.

III. Results

Univariate analysis was performed on the total population as a response variable and the other variables as predictors. Fig.1 shows the relationship between total population estimates based on birth rate. From the graph it can observed that there is no discernible pattern and the data points do not fall along the straight line.

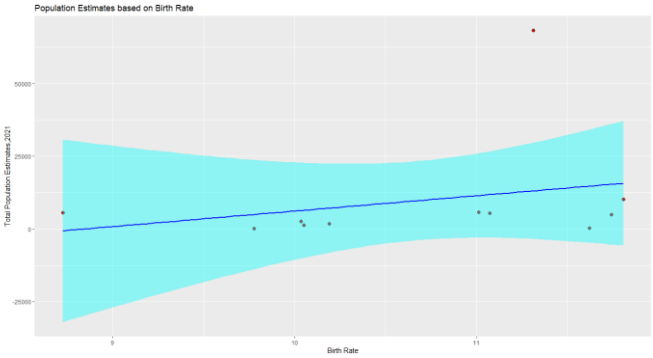


Figure 1 Scatter plot showing the relationship between total population estimates and birth rate

Fig.2. shows the relationship between total populations estimates and mortality rate. From the graph it can be observed that some data points fall on the straight line.

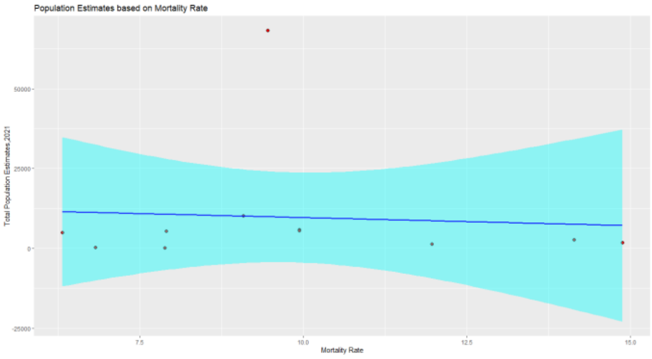


Figure 2 Scatter plot showing the relationship between total population estimates and estimates and mortality rate

Fig 3. Shows a scatter plot showing the relationship between total population estimates and life expectancy. It can be observed that there are outliers and data points deviate from the straight line.

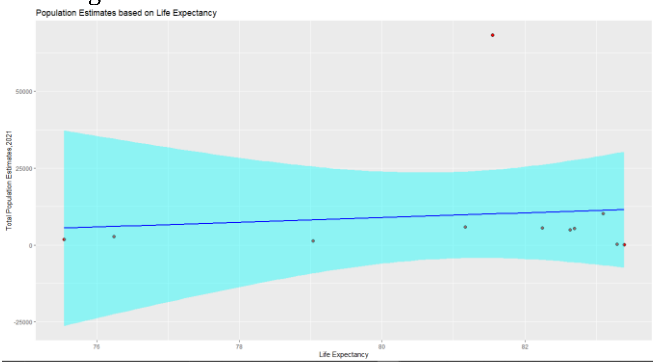


Figure 3 Scatter plot showing the relationship between total population estimates ad life expectancy

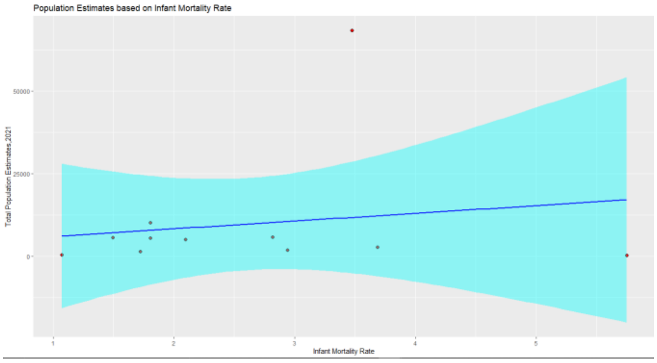


Figure 4 Scatter plot showing the relationship between total population estimates and infant mortality rate

The assumption was as infant mortality rate increases the population should decrease considerably. But there appears to be outliers and the total population does not appear to be affected by the variable.

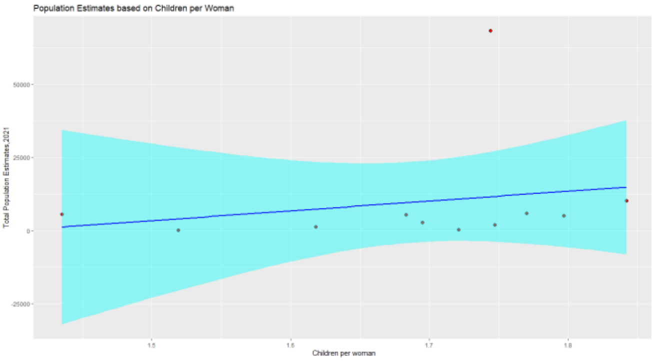


Figure 5 Scatter plot showing the relationship between total population estimates and children per woman

There appears to a slight increase in population with some data points, again outliers are present as shown in Fig.5. The points deviate from the straight line.

Fig.6. shows the relation between total population estimates and growth rate. It appears that the data points are closer to the straight line at the beginning then appear to deviate away from the line with outliers being present.

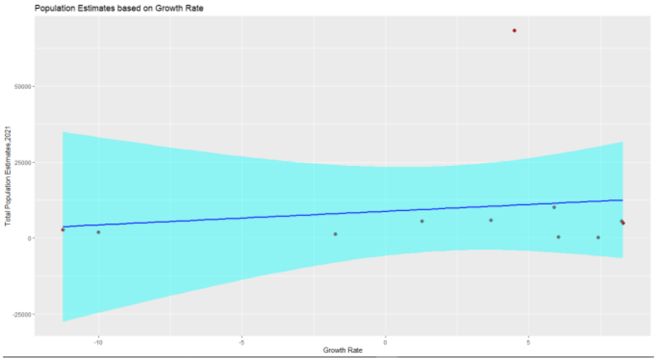


Figure 6 Scatter plot to show the relationship between total population estimates and growth rate

From fig.7. it appears that there is some sort of linear relationship between the number of people over the age of 65 and the total population estimates.

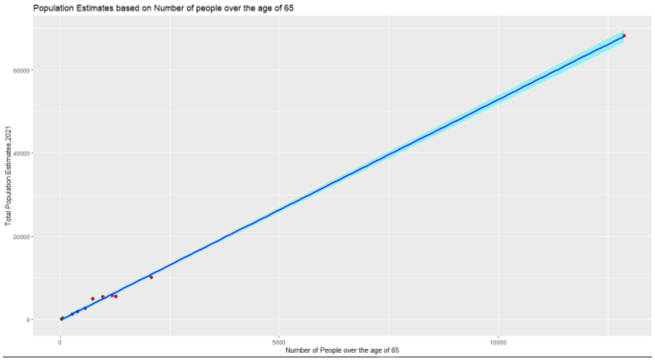


Figure 7 Scatter plot showing the relationship between total population estimates and the number of people over the age of 65

One of the research questions this study aims to answer is to view the data on a choropleth map to better understand how each country in the northern European region fares with each variable. Choropleth maps were constructed to find out how countries differ from each other.

Fig. 8. Shows that United Kingdom appears to have the highest population as the color intensity is the highest when compared to other countries.

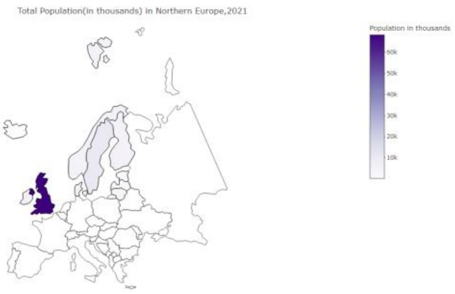


Figure 8 United Kingdom appears to have the highest population

Fig. 9. Shows that Sweden has the highest birth rate, that is the number of live births per 1000 population

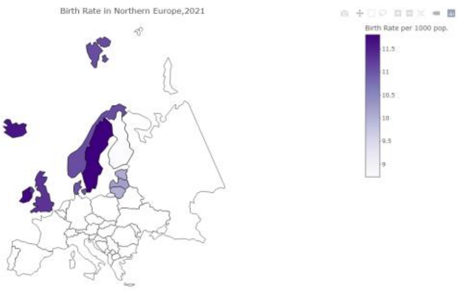


Figure 9 Sweden appears to have the highest birth rate

Fig 10. Shows that Latvia appears to have the highest mortality rate as indicated by the intensity of the color in the choropleth map.

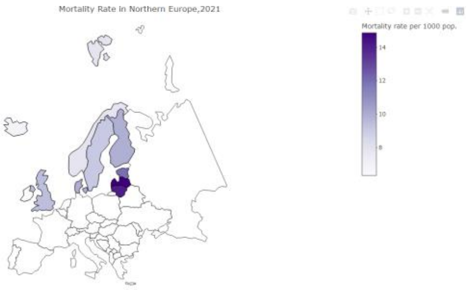

Mortality Rate in Northern Europe,2021
Mortality rate per 1000 pop.
14
12
10
8

Figure 10 Latvia appears to have the highest mortality rate

Fig. 11. Shows that Iceland appears to have the highest life expectancy among other countries in northern Europe.

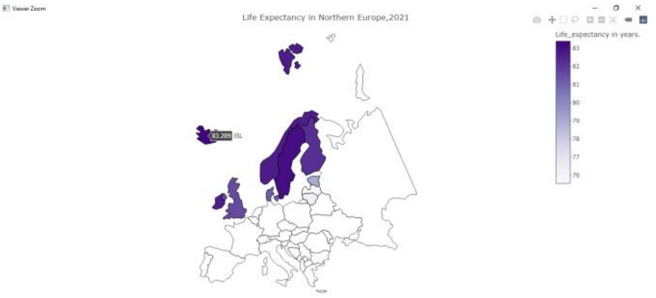

Viewer Zoom
Life Expectancy in Northern Europe,2021
Life_expectancy in years.
83
82
81
80
79
78
77
76
83.289 ISL

Figure 11 Iceland appears to have the highest life expectancy

Fig. 12. Shows that Lithuania appears to have the highest infant mortality rate

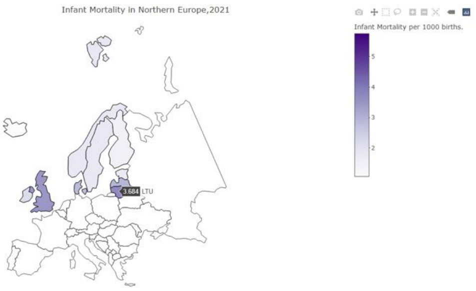

Infant Mortality in Northern Europe,2021
Infant Mortality per 1000 births.
5
4
3
2
3.684 LTU

Figure 12 Lithuania appears to have the highest infant mortality rate

Fig 13. Shows that women in Sweden have on an average of 2 children.

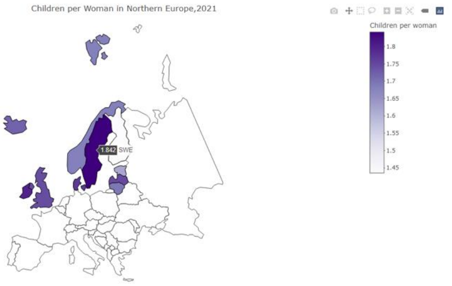


Figure 13 Choropleth map to show that women in Sweden have more than one child

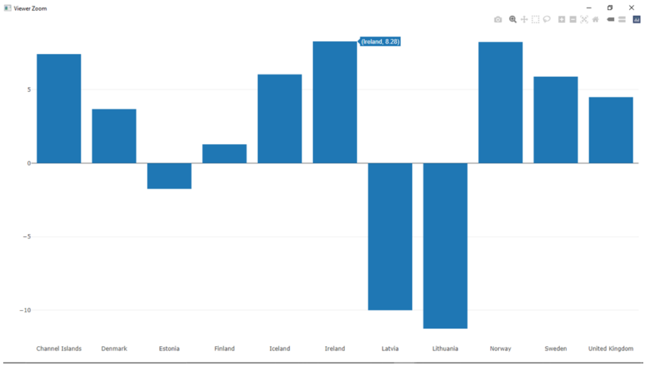


Figure 14 Shows that Ireland has the highest growth rate

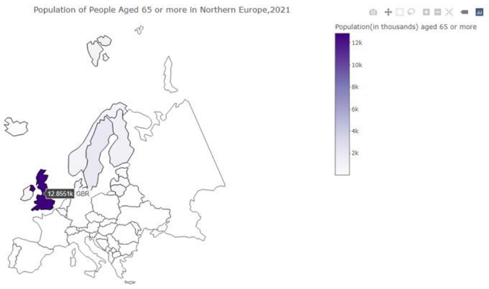

Population of People Aged 65 or more in Northern Europe,2021
Population(in thousands) aged 65 or more
12k
10k
8k
6k
4k
2k
12.8551k GBR

Figure 15 Shows that the UK has the highest number of people living over the age of 65

Multiple Linear Regression

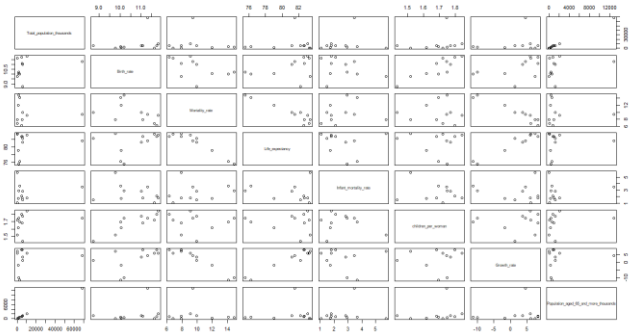
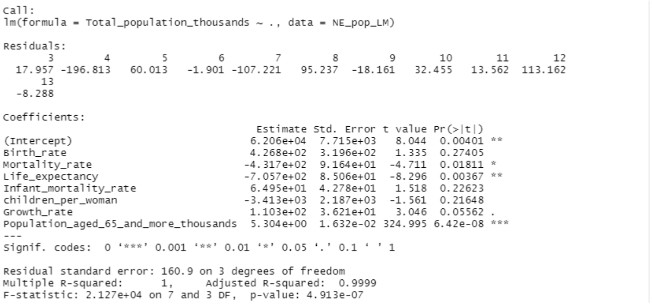


Figure 16 Scatter Plot matrix to show the correlation of variables



```
Call:
lm(formula = Total_population_thousands ~ ., data = NE_pop_LM)

Residuals:
       3        4        5        6        7        8        9       10       11       12 
  17.957 -196.813   60.013   -1.901 -107.221   95.237  -18.161   32.455   13.562  113.162 
      13 
  -8.288 

Coefficients:
                                       Estimate Std. Error t value Pr(>|t|)    
(Intercept)                           6.206e+04  7.715e+03   8.044  0.00401 ** 
Birth_rate                            4.268e+02  3.196e+02   1.335  0.27405    
Mortality_rate                       -4.317e+02  9.164e+01  -4.711  0.01811 *  
Life_expectancy                      -7.057e+02  8.506e+01  -8.296  0.00367 ** 
Infant_mortality_rate                 6.495e+01  4.278e+01   1.518  0.22623    
children_per_woman                   -3.413e+03  2.187e+03  -1.561  0.21648    
Growth_rate                           1.103e+02  3.621e+01   3.046  0.05562 .  
Population_aged_65_and_more_thousands 5.304e+00  1.632e-02 324.995 6.42e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 160.9 on 3 degrees of freedom
Multiple R-squared:      1,	Adjusted R-squared:  0.9999 
F-statistic: 2.127e+04 on 7 and 3 DF,  p-value: 4.913e-07
```

Figure 17 The multiple linear regression model of all the variables

Form the multiple linear regression model of all the variables, it can be observed that the birth rate has a p-value of 0.274 which is greater than the significance level of 0.05, hence birth rate does not appear to be significant on the total population estimates.

Mortality rate has a p-value of 0.018 which is less than the significance level of 0.05, hence it could be significant to the response variable.

Life expectancy has a p-value of 0.003 which is less than the significance level, hence the predictor variable might be significant. Infant mortality rate is 0.22 which is greater than the significance level and hence might not be significant.

Children per woman has a p-value of 0.21 which is greater than the significance level and hence might not be significant.

Growth rate has a p-value almost equal to the significance value. The variable could be significant.

The p-value of the number of people over the age of 65 is very low, it is 0.00000006 which is less than the significance level of 0.05, hence it could be significant. But the factors leading to such a low p-value should further investigated.

The diagnostic plots were plotted for the model[6][7]. Form fig. 18. It appears that The red line is not close to be being a horizontal line centered at 0.The pattern is not consistent with the assumption that the errors are independent identically distributed samples from a normal distribution with mean 0. There are 3 numbered points, one near the top and two below 0. These appear to be outliers

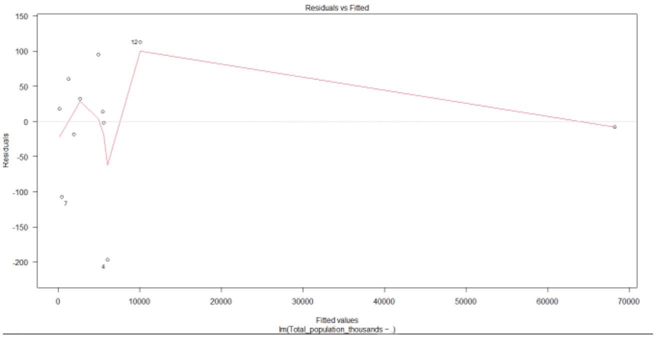


Figure 18 The residual vs fitted values plot for the model with all variables

From fig. 19. It appears that the top right of the Q-Q plot shows two outliers numbered 8 and 12, similarly the bottom left also shows an outlier numbered 4. The points deviate from the diagonal straight line which indicates that the data may not be normally distributed.

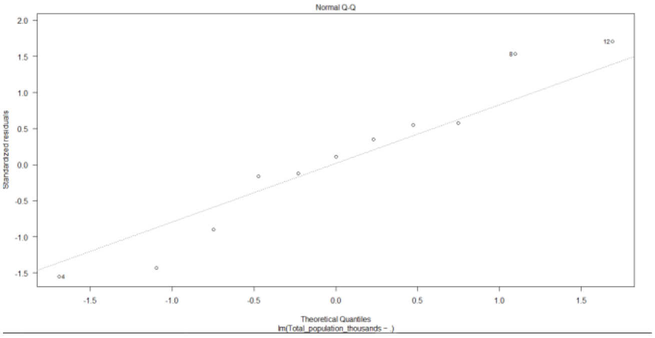


Figure 19 Q-Q plot for the model with all variables

From Fig. 20. the plot ideally should have been roughly flat red line at y = 1. The inconsistency in the line indicates a big departure from the identically distributed assumption

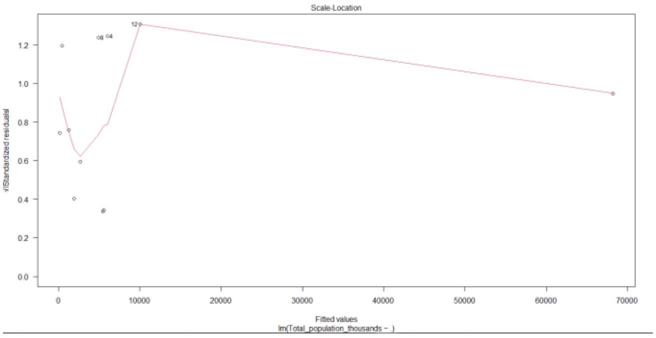


Figure 20 Scale-location plot for the model with all variables

From fig. 21. It looks like there are 3 data points outside the cooks' distance, which indicates that they may be influential points.

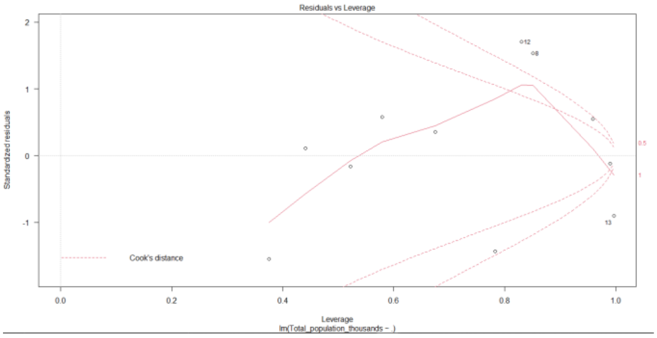
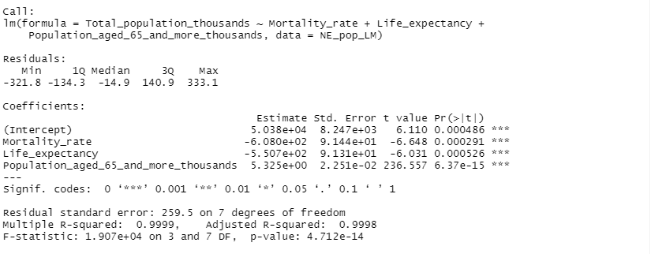


Figure 21 Residual vs Leverage plot for the model with all variables

Another model was built, with only three variables mortality rate, life expectancy and the number of people aged over 65. The p-value of mortality was found out to be 0.000291 which is less than the significance level of 0.05. Hence the predictor could be significant.

The p-value of life expectancy was found out to be 0.000526 which is less than the significance level, hence the predictor could be significant.

The p-value of the number of people over the age of 65 was found out to be 6.37×10^{-15}. The model is heavily influenced by the variable of the number of people aged over 65 as indicated by the low p-value. The model has a residual standard error of 259.5 and an adjusted R-squared value of 0.9998 which means that 99% variance in the measure of population estimates can be predicted by mortality rate, life expectancy and the number of people aged 65 or above.



```
Call:
lm(formula = Total_population_thousands ~ Mortality_rate + Life_expectancy +
    Population_aged_65_and_more_thousands, data = NE_pop_LM)

Residuals:
   Min     1Q Median     3Q    Max 
-321.8 -134.3  -14.9  140.9  333.1 

Coefficients:
                                       Estimate Std. Error t value Pr(>|t|)    
(Intercept)                           5.038e+04  8.247e+03   6.110 0.000486 ***
Mortality_rate                       -6.080e+02  9.144e+01  -6.648 0.000291 ***
Life_expectancy                      -5.507e+02  9.131e+01  -6.031 0.000526 ***
Population_aged_65_and_more_thousands  5.325e+00  2.251e-02 236.557 6.37e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 259.5 on 7 degrees of freedom
Multiple R-squared:  0.9999,	Adjusted R-squared:  0.9998 
F-statistic: 1.907e+04 on 3 and 7 DF,  p-value: 4.712e-14
```

Figure 22 Another model was built showing with mortality rate, life expectancy and number of people aged over 65

The diagnostic plots were plotted for the model.

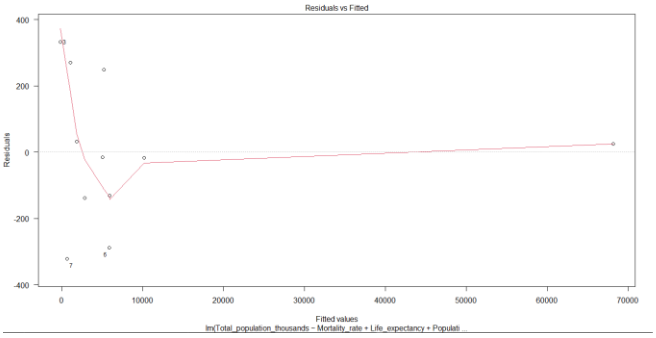


Figure 23 Residual vs fitted values plot for the model with mortality rate, life expectancy and the number of people aged 65 and above

From the fig 23. It can be observed that the red line is not close to be being a horizontal line centered at 0. Rather it dips below 0. There are outliers numbered 3, 7 and 6. Overall the pattern is not consistent with the assumption that the errors are independent identically distributed samples from a normal distribution with mean 0.

From fig. 24. It appears that there are outliers at the top right numbered 3 and at the bottom left with data points being numbered 7 and 6. The points are not consistent with the straight line and appear to deviate away from it, which may indicate that the data may not be normally distributed.

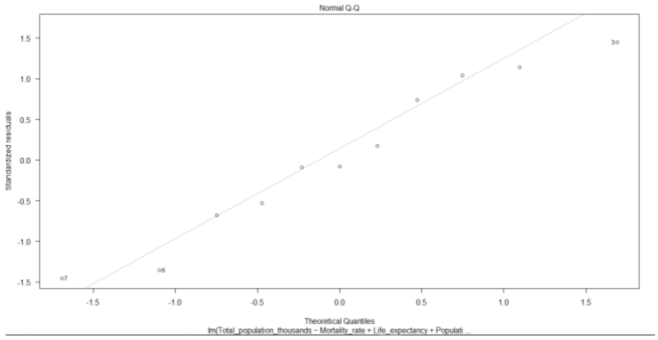


Figure 24 Q-Q plot for the model with mortality rate, life expectancy and the number of people aged 65 and above

Like the previous model, the scale-location plot as indicated by the fig.25. shows that the model violates the assumption that errors are independent and are identically distributed samples from a normal distribution with mean 0.

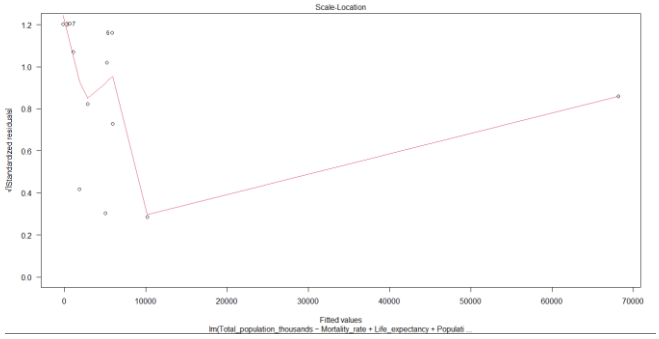


Figure 25 Scale-location for the model with mortality rate, life expectancy and the number of people aged 65 and above

From fig.26. it can be observed that there appears to be a large leverage in the model. As indicated by the p-value, the variable of the number of people aged over 65 could have a strong influence on the coefficients in the model. This can be observed by the data point numbered 13 which lies outside the cook’s distance.

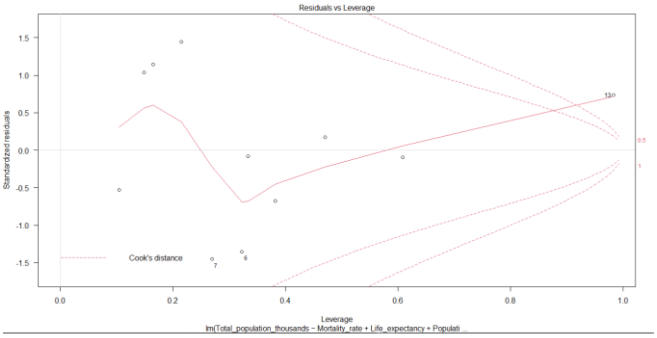


Figure 26 Residual vs leverage plot for the model with mortality rate, life expectancy and the number of people aged 65 and above

Further models were built to obtain a lower residual standard error, but the linearity assumptions were failing. With all the models built it appeared that one variable was having a high leverage on the coefficients in the model.

IV. CONCLUSION

This study aimed to analyze the how the data related to different variables in the dataset vary overreach country, this was achieved through building choropleth maps to determine how each country in northern Europe fared compared to each other. A regression model was built to analyze the factors which influenced the total population estimates. The number of people aged 65 and above had a large influence on the coefficients. The limitation of this model would be the lack of a larger dataset which might have shown other dependencies.

REFERENCES

[1] "All countries," Ined.fr. [Online]. Available: https://www.ined.fr/en/everything_about_population/data/all-countries/?lst_continent=908&lst_pays=924. [Accessed: 10-Dec-2021].

[2] A. Nair, "Beginner's guide to geographical plotting with plotly," Analyticsindiamag.com, 30-Jul-2019. [Online]. Available: https://analyticsindiamag.com/beginners_guide_geographical_plotting_with_plotly/. [Accessed: 11-Dec-2021].

[3] "Plotly colours list," Plotly.com, 13-Jul-2018. [Online]. Available: https://community.plotly.com/t/plotly-colours-list/11730/3. [Accessed: 11-Dec-2021].

[4] S. L. Sarath, "PAIRPLOT VISUALIZATION - Analytics Vidhya - Medium," Analytics Vidhya, 29-Sep-2019. [Online]. Available: https://medium.com/analytics-vidhya/pairplot-visualization-16325cd725e6. [Accessed: 12-Dec-2021].

[5] Sthda.com. [Online]. Available: http://www.sthda.com/english/articles/40-regression-analysis/168-multiple-linear-regression-in-r/. [Accessed: 13-Dec-2021].

[6] J. Frost, "Standard error of the regression vs. R-squared - statistics by Jim," Statisticsbyjim.com, 25-Mar-2017. [Online]. Available: https://statisticsbyjim.com/regression/standard-error-regression-vs-r-squared/. [Accessed: 13-Dec-2021].

[7] J. Frost, "Check your residual plots to ensure trustworthy regression results! - statistics by Jim," Statisticsbyjim.com, 05-Apr-2017. [Online]. Available: https://statisticsbyjim.com/regression/check-residual-plots-regression-analysis/. [Accessed: 13-Dec-2021].