Infant mortality in the world

I. Introduction

Infant mortality has been for decades one of the largest challenges our society had to face. Nowadays, despite the rapid progress of modern medicine and the constant improvement of the living conditions in most of the inhabited world, it still remains a widespread issue, that concerns millions of families. The total number of deaths of children under five years of age decreased from the almost 10 million in 2000 to the 5.6 million in 2016.

Our goal is to analyze the main factors that seem to lead to higher infant mortality rates across the world and to explore the principal causes of improvement in this field, in the mentioned period of time. This analysis will focus on medical, social, environmental, and economic factors, trying to find the best model to describe the evolution in time of the infant mortality rate.

II. Data

• Hypothesis formulation and data collection

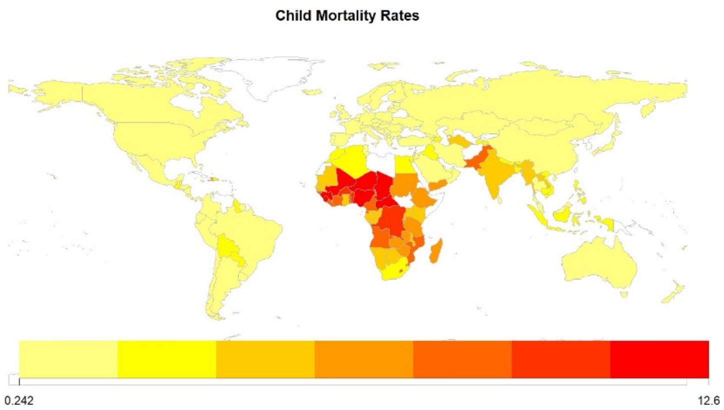
In order to investigate the topic we chose, we tried to find the most compelling factors at our disposal. We identified five predictors that we thought could vastly impact the fluctuations in infant mortality rate: the number of maternal deaths, the percentage of malaria contagions, the GDP per capita and the vaccination rates of two vaccines, MCV1 and DTP3, which respectively cure measles and diphtheria, tetanus, and pertussis. The choice of these vaccines was based on the fact that measles, tetanus and pertussis are the illnesses more frequently responsible for infant mortality, between the conditions for which vaccine. Indeed, there are several studies that show a relation between infant mortality, vaccination (Higgins et al., 2016; Omar & Muhsen, 2021) and economic factors (O'Hare et al., 2013). However, the correlation between infant mortality, malaria and maternal death remains partially explored. Furthermore, it remains to be explored if health factors, integrated with economic status, correlate with infant mortality. We collected data of these five factors and used Excel to merge the different datasets and form a coherent table, which collects data for 161 countries spanned over 17 years, from 2000 to 2016, resulting in a 2738 x 9 table. Please refer to the bibliography for the sources.

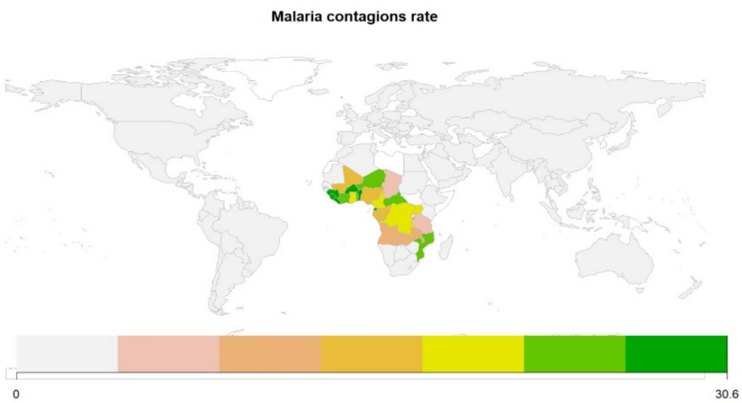
• Data visualization and manipulation

We now report some plots, charts, and maps we used to better represent and explore our data. In order to do so, we used R. Please refer to the code attached.

All the charts in this section represents data from 2016 only. The first map (on the left) represents the infant mortality rates in the 161 countries we considered in our analysis. The second (on the right) represents the rate of people infected by malaria in those countries. Already from this early visual representation, we can see that higher rates are more common in the Global South of the world: this suggests that economic factors are effectively relevant in the analysis of infant mortality rates.

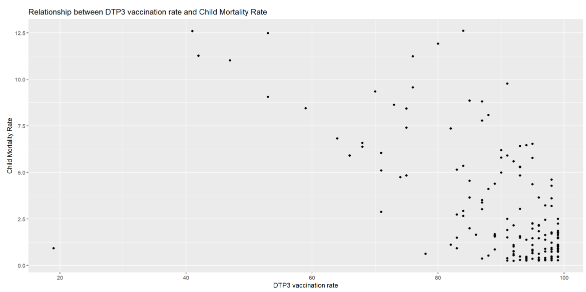
Furthermore, it also hints at a relation between infant mortality rates and malaria contagions percentage, as it is shown in the comparison between the two maps.

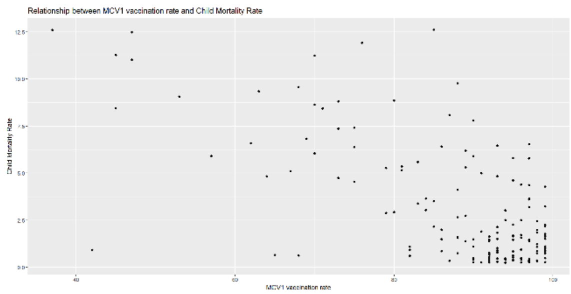


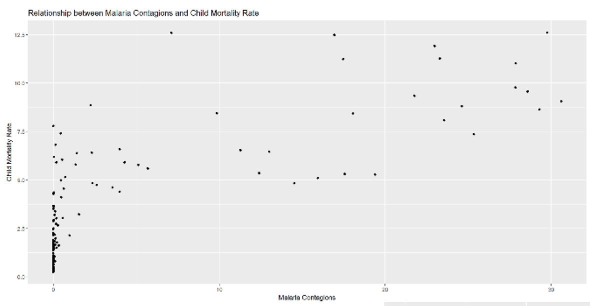


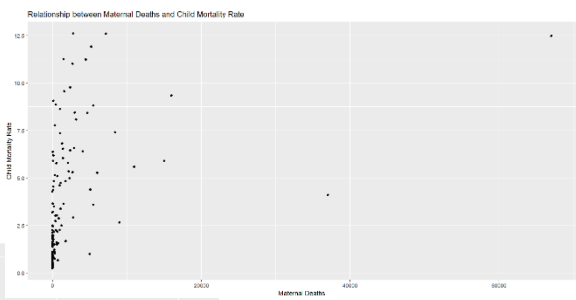
We proceed by reporting some scatter plots that compare the independent variables of our research to the dependent one. As expected, the plots show some kind of correlation between each individual factor and infant mortality rate. Still, some outliers are visible in the plots: we will further investigate them in the study limitations section.

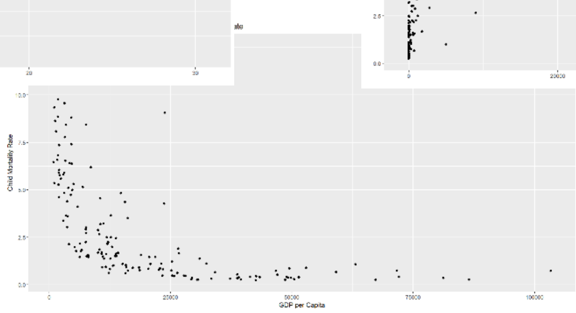
All plots are fairly satisfying and confirm our initial beliefs: GDP per capita and vaccination rates have a negative correlation with infant mortality, while maternal deaths and malaria contagions have a positive correlation. Thus, we understand that countries with higher GDP and vaccination rates report a lower infant mortality rate; countries with higher maternal deaths count and malaria contagions rate, instead, present a more dramatic situation concerning infant mortality.











III. Multivariate linear regression

• Applying the regression

In this section, we are going to explore the results of the multivariate linear regression we used to analyze the data and to further examine the relation between the factors and the dependent variable. Please note that here the results refer to the complete study of the dataset, i.e., we will consider all years between 2000 and 2016. The analysis of the whole dataset of 161 countries is useful in order to consider all possible patterns. For the output, please check the R code.

```
Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         1.509e+01  3.142e-01  48.019  < 2e-16 ***
Maternal_Deaths     1.686e-05  5.296e-06   3.183  0.00147 **
Malaria_Contagions  2.021e-01  4.694e-03  43.064  < 2e-16 ***
MCV1               -6.119e-02  7.371e-03  -8.301  < 2e-16 ***
DTP3               -6.493e-02  7.361e-03  -8.820  < 2e-16 ***
gdp_per_capita     -4.946e-05  2.050e-06 -24.123  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.015 on 2731 degrees of freedom
Multiple R-squared:  0.7939,    Adjusted R-squared:  0.7935
F-statistic:  2104 on 5 and 2731 DF,  p-value: < 2.2e-16
```

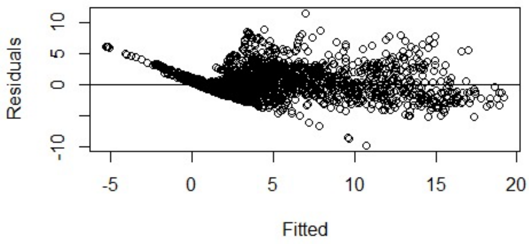
The results of the linear regression show first of all the negative correlation that infant mortality rate has with the vaccination rates and the GDP, and the positive one with maternal deaths and malaria contagions, as already shown by the scatter plots.

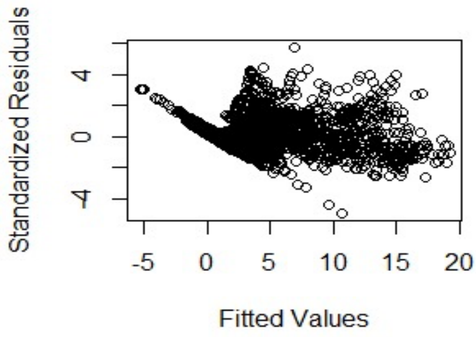
Furthermore, the significantly low results of the p-value for each predictor, together with the output of the t-test which is different from 0 for all the independent variables, indicates that the relations we highlighted are statistically significant: we can thus reject the null hypothesis, that claimed that each predictor was not relevant in the analysis. Only the maternal deaths variable has a slightly higher p-value, but it is still well below the 0.05 threshold.

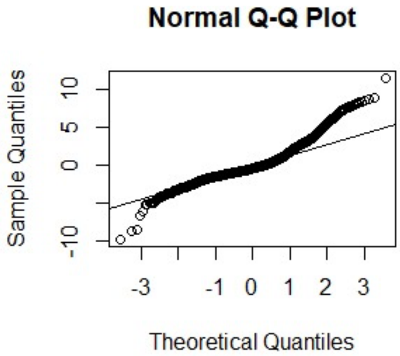
As for the model, the high F-statistic contribute to strengthen our hypothesis, and so does the low p-value. Moreover, the multiple and adjusted R-squared, respectively of 0.7939 and 0.7935, are fairly high and indicate that almost 80% of the variance of infant mortality rate is explained by the predictors. It is important to note that the adjusted R-squared is a reliable parameter, since it performs well even with a high number of covariates. The RSE, instead, is low enough not to give any concerns, especially if considering the high values of the adjusted R-squared.

• Validation analysis

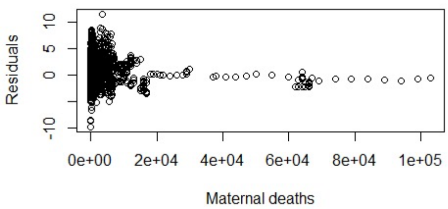
Now let us check for normality and homoscedasticity. First of all, we will consider the Normal Q-Q plot and the plot of the fitted values against the residuals and the standardized residuals. The Normal Q-Q plot is fairly adherent to the 45° line, with a slight curvature on both ends, which can be attributed to outliers. The plot of the fitted values against the residuals and the standardized residuals don’t present any particular pattern, although they show different points far from the 0 line, which again indicates the presence of outliers. As said before, we will expand the topic of outliers in the limitation of study section.

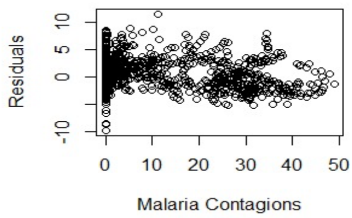


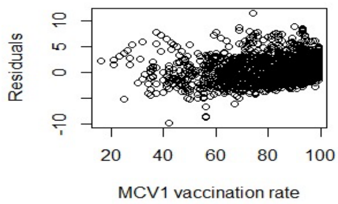


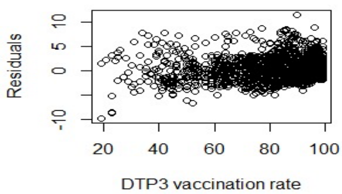


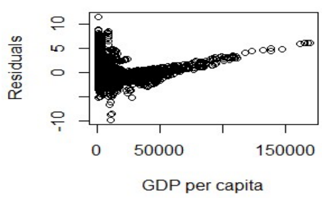
We also plotted each predictor against the residuals. The only one presenting a slightly problematic behavior is the plot of maternal deaths, while the other show no patterns and fairly constant spread.











• Model testing

Now let us check if there exist other models that would be more performing. In order to do so, we did a “Step-both” test, that indicates that the model including all the covariates is more performing than any model with less covariates, since every step without any one of the predictors reports a higher AIC than the model with all the predictors. Please refer to the R code to check the result shown below.

```
Start:  AIC=3841
Child_Mortality ~ Maternal_Deaths + Malaria_Contagions + MCV1 +
    DTP3 + gdp_per_capita

                     Df Sum of Sq   RSS    AIC
<none>                            11088 3841.0
- Maternal_Deaths     1      41.1 11129 3849.1
- MCV1                1     279.7 11368 3907.2
- DTP3                1     315.8 11404 3915.9
- gdp_per_capita      1    2362.6 13450 4367.7
- Malaria_Contagions  1    7529.2 18617 5257.4
```

IV. Limitations of study and conclusion

The first limitation encountered in this analysis was merging different datasets, which had information about different countries and years, which forced us to reduce the number of countries to 161 and the period of time between 2000 and 2016, although it is still fairly significative. We should also consider the reliability of the data collected by each country: clearly, countries more focused on research (and with more resources to support it) have more sound data than others.

Concerning the outliers, we should first of all consider that we analyzed the problem from a worldwide perspective, which can lead to find more patterns, but also negates the possibility to conduct the analysis on a more specific level. Moreover, it is also important to note that, while our predictors have a significant statistical impact on the issue of infant mortality, there are other factors to consider: as an example, hygiene, transport systems (to easily reach healthcare facilities) and genetic conditions. This last factor could also explain why the maternal death rate seemed to be the most problematic covariate with respect to the plot against residuals.

However, our research shows a good correlation between the factors we analyzed and infant mortality using a multivariate linear regression. Although not sufficient to prove causality, the evidence provides useful suggestions on the topic thereby opening the possibility for future in-depth analyses.

V. Bibliography

- Child mortality rate: Our world in data - UN Inter-agency Group for Child Mortality Estimation (2023)
- Vaccination rates: Our world in data – WHO/Unicef (2022)
- Maternal deaths: Our world in data – WHO via World Bank (2023)
- Malaria contagions rate: Our world in data – IHME, Global Burden of Deases (2019)
- GDP per capita: Our world in data - Feenstra et al. (2015), Penn World Table (2021)