****

**Applied Regression Analysis (BUS 41100) - Autumn 2020**

**Final Project**

**Team 2**

**Maria Gambi Romero**

**Allyson Feitler**

**Guillermo Menendez Kirn**

**November 30, 2020**

**INDEX**

1. **Context**
2. **Data preliminary analysis**
3. **Relationship between variables**
   1. **Individual linear regression**
   2. **Research on COVID-19**
   3. **Multicollinearity**
4. **Model**
   1. **Selecting our Base model**
   2. **Choosing additional models**
   3. **Fitting models with train data and making predictions with test data**
   4. **Considerations for the selected model**
   5. **Application**
5. **Conclusion**

**Appendix**

**References**

1. **Context**

The aim of the project is to increase the understanding of the factors that determine deaths caused by COVID-19 in several countries. Specifically, it has been concluded that GDP per capita, seniors per capita, obesity rate and gender rate are important determinants of fatality rate among countries. Additionally, the model defined -that predicts fatality rate per country- may be used to evaluate the success in the fight against COVID-19 and/ or inform decision making in vaccine prioritization.

The work is based on the collection of COVID-19 data deposited in the Harvard Database repository. The data collection includes information from 77 different countries and 13 independent variables, such as obesity and hospital beds. The dependent variable for our study will be the average case fatality rate between August and September 2020.

As a starting point, the information available has been understood and cleaned. Regarding the analysis, first, it has been studied which variables are correlated and which ones present high multicollinearity. Understanding which variables are relevant is key in preparing the model for prediction.

Second, after identifying correlation between variables and taking into consideration different medical publications, a base model that aims to explain fatality rate among countries has been built. The existence of causality has been discussed and covered, however, as addressing this issue with statistical models is complex, the study is based in current public available findings.

Lastly, the base model has been used as starting point to construct the optimal prediction for fatality rate using a step forward regression to select extra variables and interactions. The data sample has been divided into two pieces -train and test. The train dataset has been used to determine the estimated β of each model. The test sample has been employed to calculate the out of sample error. That model with the lowest error has been selected as the best prediction model.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

1. **Data analysis**

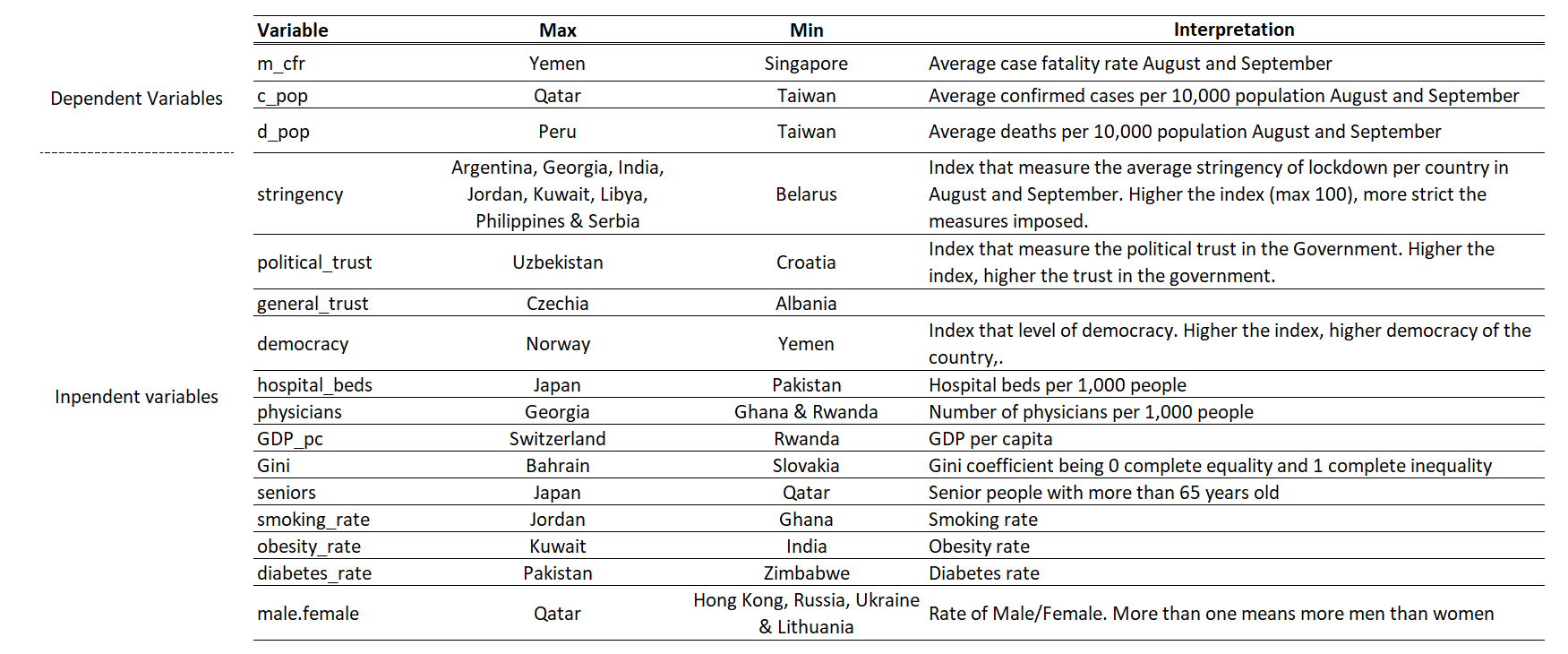
The analysis is based on COVID information available from 77 different countries collected by Chen, Dan, Li, Yong and Wu, Jiebing, 2020. The data set includes the COVID deaths and cases in September and August 2020 per country. Additionally, for each country extra tabulated information is provided, which includes the political trust, general trust, democracy, number of hospital beds, number of physicians per 1,000 habitants, GDP, Gini index, number of seniors, the smoking, obesity rate and diabetes rate as well as the male/female ratio. It should be highlighted that there is risk of selection bias in the data given only 77 out of the 195 countries are covered in the dataset.

Some adjustments have been made in order to prepare the data for our model. First, number of deaths and cases two months of data was available, so in those cases it has been decided to consider the average between those numbers as the variable for the study. In addition, after evaluating political and general trust, it has been detected that these two measurements were not very consistent with the rest of the data and there was not enough information to understand them. For these reasons, it has been decided to keep both variables out of the model. Finally, some transformations to have per capita numbers have been added.

Three dependent variables that help to better understand the COVID problem have been adopted. The first variable is the fatality rate (m\_cfr), which is the number of deaths divided by the total cases of the country. The second variable is cases per capita (c\_pop), and the last variable is deaths per capita (d\_pop). Between these three variables, deaths and cases per capita would not be a good measure if countries are in a different part of their pandemic cycle. A country that is starting to fight with the pandemic would have a small number of deaths and cases of COVID-19. On the other hand, a country that started to fight the pandemic a long time ago will have many deaths and cases. It will not be correct to compare these two in our model. Fatality rate solves this problem as, once you have a certain number of cases, the rate maintains its value in relationship with characteristics shown by independent variables. Therefore, it has been decided to employ the fatality rate as the dependent variable.

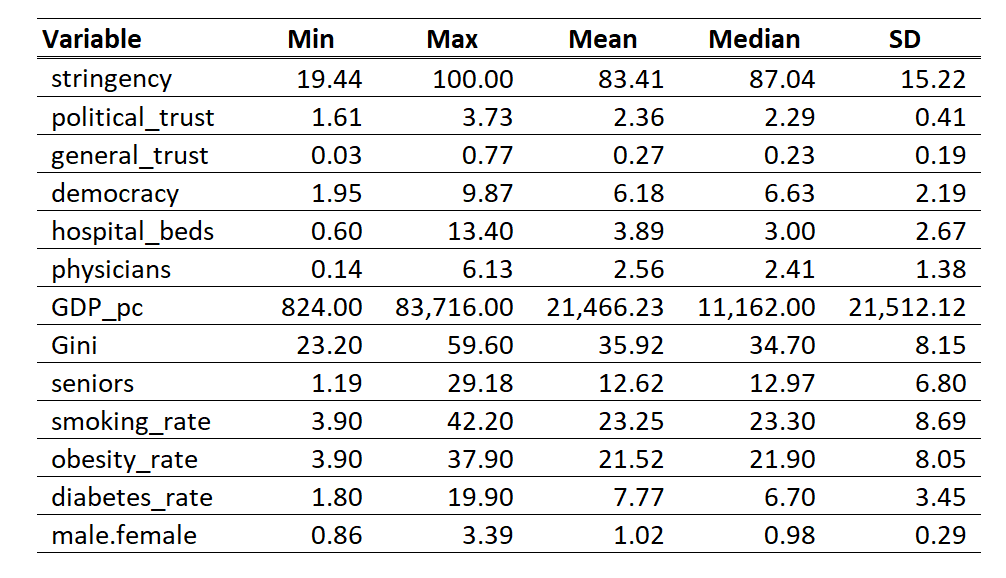
Tables 1 and 2 below illustrate the characteristics of the data set we had analyzed. In addition, boxplots in the Appendix helps visualize and get a sense of the magnitude and distribution of the variables.

*Table1. Variables used for analysis*



*Table 2. Distribution of values per variable. Includes Maximum, minimum, mean, median and*

*standard deviation*

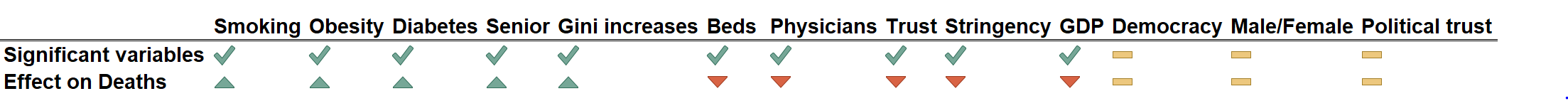
**

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

1. **Relationship between variables**

All the variables have been analyzed in order to determine which ones were relevant to evaluate the fatality rate in each country. Intuitively, a positive relationship between smoking rate, obesity rate, diabetes rate, the number of seniors, Gini coefficient and the deaths per country was expected. In a similar manner, a negative relationship between hospital beds, physicians, trust, stringency, GDP and deaths per country was also expected. For the rest of the variables, it is difficult to predict the relationship, as it is the case with democracy, male/female rate and political trust. Table 4, summarizes our preliminary hypothesis of the relationship between variables.

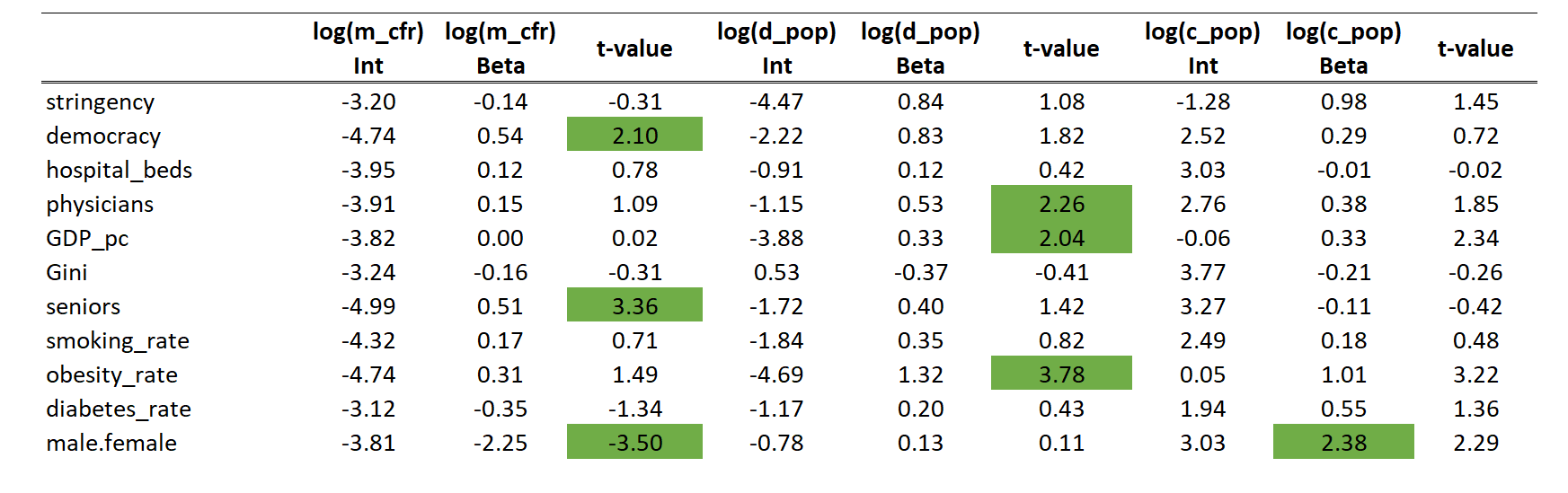
*Table 4. Expected correlation between the independent variables and the average rate fatality rate and the effect on deaths.*

****

**3.a.Individual linear regressions**

The first approach to understand the relationship between variables was to run an individual linear regression for all independent variables, with all of the dependent variables mentioned above: deaths per capita (d\_pop), cases per capita (c\_pop) and fatality rate (m\_cfr). As shown in table 5, with this approach the first variables that appear to be relevant for our study are democracy, number of seniors per capita, gender ratio, physicians, GDP per capita and obesity rate.

*Table 5. Linear regression, intercept, beta and t-value*

****

*Note: M\_cfr (fatality rate), d\_pop (# deaths) and c\_pop (# of cases)*

It is important to mention that there could be causality between some of these variables. We are aware that it is very hard to address causality with a statistical model. Therefore, we decided to explore different research that could help us determine which variables are strongly related with COVID-19 fatalities and cases.

**3.b. Research on COVID-19**

Researchers, Rabail Chaudhrya , George Dranitsarisb , Talha Mubashirc , Justyna Bartoszkoa and Sheila Riazia have found that increased COVID caseloads were associated with countries with higher obesity, higher median country age and longer time to border closures. They found COVID-19 mortality was associated with increased obesity, advanced age and per capita GDP. However, they recognized that the relationship with per capita GDP might be due to access to testing. Moreover, they found variables that were negatively associated with increased COVID-19 mortality were reduced income dispersion within a country, smoking prevalence and the number of nurses per capita. It should be noted smoking rates are associated with lower median ages and that COVID-19 health outcomes also relied on national preparedness and scale of testing.

On the other hand, other studies have found that the older population is more vulnerable to COVID-19. Specifically, Clara Bonanad has found the highest mortality rate among individuals over 80, while Ueyama, Hiroki MD has found men are more likely to develop severe COVID-19 symptoms. To sum up, the research that we analyzed suggests that GDP per capita, the age of the population, smoking rates, obesity rate and gender are all related with deaths and cases of COVID 19 in different countries.1

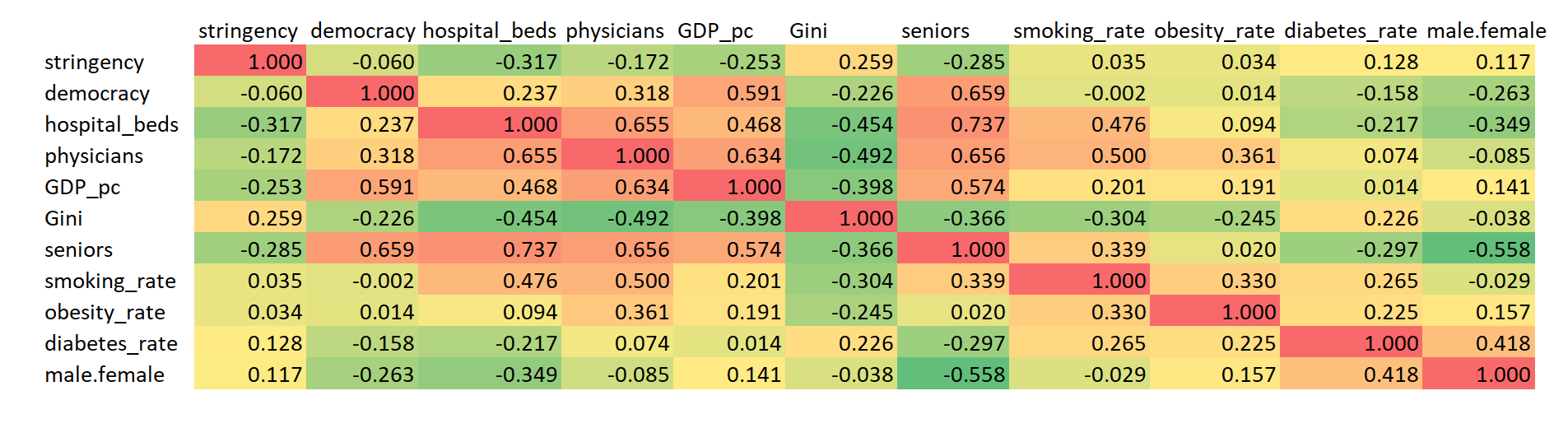
1 Source: See reference section at the end of the document

**3.c Multicollinearity.**

Finally, within the data analysis, multicollinearity in the variables has been studied. The team acknowledge that some variables were closely related to each other. As an example, it has been detected that the number of senior people is highly correlated with hospital beds and the number of physicians. This makes sense, as it is expected that countries with older people to have a bigger health system. This pattern has been identified across different variables.

For instance, GDP per capita is highly correlated with democracy, hospital beds, physicians, and seniors. Therefore, to reduce multicollinearity in the model, some of the variables have been discarded. It has been decided to keep GDP per capita and seniors in the model, subtracting hospital beds and physicians. At the end, controlling for GDP per capita and seniors would be also controlling for hospital beds, physicians, and democracy as they all move together. Multiple transformations of the data have been evaluated in order to simplify the scope. Eliminating multicollinearity completely it is not possible, and this has been taken into consideration when evaluating the results.

*Table 6. Correlation between independent variables*

****

Finally, the selection of variables for the base model has been informed by the teams’ expectations, guidance from research, and the adjustment for multicollinearity. It has been decided to include GDP per capita, Gini coefficient, seniors per capita, smoking rate, diabetes rate, obesity rate and gender rate. These independent variables helped to understand fatality rate and to control for key characteristics of each country.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

1. **Model**

After analyzing the relationship between variables, the next step was developing a model that allowed to predict possible fatality rates given certain characteristics for each country. In order to get the best prediction model given the data, a 6 step methodology was followed:

1. Create a base model
2. Use the step forward regression to select different models
3. Select the models that we want to compare
4. Divide randomly data into train and test sections
5. Compute out-of-sample predictions and compare them with test data
6. Take the square errors from each model, running the process 10,000 times with resample
7. Rank the mean of errors for each model
8. Choose the best model for predictions

**4.a. Selecting our Base model**

As discussed, the base model takes into consideration variables that previous research suggests are relevant for the study of COVID 19 controlling with key variables that allow to measure differences between countries. The independent variables include number of seniors per capita, smoking rate, obesity rate and diabetes rate; controlling for GDP per capita, Gini, and the gender rate. It has been decided to use a log-log model, as all variables were positive. It was very helpful to control for a non-constant variance, and has generated a good transformation of the variables used, spreading them out.

Base model

Estimate Std. Error t value Pr(>|t|)

(Intercept) -5.33422 2.64168 -2.019 0.0473 \*

GDP\_pc -0.26918 0.14574 -1.847 0.0690 .

Gini 0.37037 0.57158 0.648 0.5192

seniors 0.83105 0.32407 2.564 0.0125 \*

smoking\_rate -0.25343 0.27479 -0.922 0.3596

obesity\_rate 0.53875 0.21505 2.505 0.0146 \*

diabetes\_rate -0.01482 0.30675 -0.048 0.9616

male.female -0.24209 1.09200 -0.222 0.8252

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8815 on 69 degrees of freedom

Multiple R-squared: 0.2658

Adjusted R-squared: 0.1913

F-statistic: 3.569 on 7 and 69 DF, p-value: 0.002475

There are some important considerations for the Base model. Between all possible variables, the only three that are significant with 90% confidence are GDP per capita, number of seniors per capita and obesity rate. The model suggests that higher seniors per capita and higher obesity rate, keeping everything else fixed, are related with a higher fatality rate. On the other hand, an increase in GDP per capita is related to a decrease in fatality rate.

Another relevant issue to discuss is the reason some of the variables kept are not statistically different from zero. Even though it cannot be argued that these variables are different from zero, many papers and previous research suggest that they could be related to fatality rates. Therefore, maintaining them in the model is not affecting the results and they could add value as we consider interactions in the following steps. It is also important to note that the F test for the model rejects the null hypothesis that the model is not able to explain the fatality rate. Therefore, we are confident that we can use this model as a base for the next section of the study.

Finally, as previously discussed, there is some multicollinearity in the model. To correct this problem, the variables have been transformed to maintain the model as simple as possible. After analyzing possible variables, the ones that were highly correlated with each other have been eliminated. However, it must be highlighted that eliminating multicollinearity completely is not something achievable. As a result, the coefficients in the model are not true marginal effects, and higher standard errors on the coefficients due to multicollinearity are expected.

**4.b Choosing additional models**

After constructing the Base model, a forward stepwise regression has been conducted, utilizing BIC and AIC information criteria to select for the best interactions between variables in order to get the best prediction model. In addition to these processes, a second stepwise regression utilizing BIC and AIC has been made, but this time it does not start with the Base model. We let the algorithm choose the best variables from scratch. Table 7, summarized the 5 preliminary models selected to generate predictions, which will be compared.

*Table 7. Preliminary models*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Base** | **Base BIC** | **Base AIC** | **Pred AIC** | **Pred BIC** |
| The model uses the variables selected by us in the first part of our analysis | The model obtained after running a forward stepwise regression, starting with the Base model and allowing for interactions using the BIC to select the best model | The model obtained after running a forward stepwise regression, starting with the Base model and  allowing for interactions using the AIC to select the best model | The model obtained after running a forward stepwise regression, starting from scratch and  allowing for interactions using the AIC to select the best model | The model obtained after running a forward stepwise regression, starting from scratch and  allowing for interactions using the BIC to select the best model |
| Dependent variable:  *m\_cfr*  Independent variables:  *GDP\_pc, Gini, seniors, smoking\_rate, obesity\_rate, diabetes\_rate, male.female* | Dependent variable:  *m\_cfr*  Independent variables:  GDP\_pc, Gini, seniors, smoking\_rate,  obesity\_rate, diabetes\_rate, male.female, GDP\_pc:seniors,  smoking\_rate:obesity\_rate, Gini:smoking\_rate, obesity\_rate:male.female, Gini:seniors, Gini:diabetes\_rate | Dependent variable:  *m\_cfr*  Independent variables:  GDP\_pc, Gini, seniors, smoking\_rate,  obesity\_rate, diabetes\_rate, male.female, hospital\_beds, GDP\_pc:seniors, smoking\_rate:obesity\_rate,  Gini:smoking\_rate, obesity\_rate:male.female, Gini:seniors, Gini:diabetes\_rate, seniors:obesity\_rate, seniors:smoking\_rate, seniors:hospital\_beds, GDP\_pc:obesity\_rate, diabetes\_rate:hospital\_beds, seniors:diabetes\_rate | Dependent variable:  *m\_cfr*  Independent variables:  Male.female, obesity\_rate,  seniors,  Hospital\_beds,  GDP\_pc,  Gini, male.female:obesity\_rate, seniors:hospital\_beds, obesity\_rate:hospital\_beds, seniors:GDP\_pc, obesity\_rate:Gini, male.female:seniors | Dependent variable:  *M\_cfr*  Independent variables:  Male.female, obesity\_rate |

**4.c. Fitting models with train data and making predictions with test data**

After narrowing the number of models, in order to define the estimated β and select the best model all 5 have been trained and tested. For this, the data has been divided in two categories: train and test data. 90% of the data has been allocated into the train category and 10% of the data has been allocated into the test category. All models have been fitted with the train data, and predictions have been made with the test data. Finally the errors on every model have been computed. This process was repeated 10,000 times with resample and we took the average error for every model as shown in Table 8.

*Table 8. Out of sample error per model*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Base** | **Base BIC** | **Base AIC** | **Pred AIC** | **Pred BIC** |
| Errors on prediction (out of sample) | 0.963 | 0.756 | 0.828 | 0.828 | 0.828 |

The results suggest that the Base BIC model is the best to make predictions of the fatality rate. It is not surprising to notice that the Base model is the worst in terms of prediction. As discussed earlier, this model was built selecting the variables that are closely related to fatality rate in many discussions around the world. This was not a prediction model and therefore, it was not expected to attain relevant predictions from it.

Another interesting point to note, is that the error on predictions is the same for the Base AIC, Pred AIC, and Pred BIC. This suggests that the three of them have the same prediction value and therefore, extra variables in the Base AIC and Pred AIC model are not helping to create a more powerful tool. The Pred BIC is simpler than the rest of the models as it only uses the rate between women and men with the obesity rate to make predictions. This speaks of the importance of these two variables in order to predict the fatality rate for each country.

**4.d Considerations for the selected model**

The model chosen explains the fatality rate per country with several variables including many interactions. It is important to mention that as we built the model for prediction, it is not possible to make inference from it. Therefore, trying to understand the variables and coefficients in the selected model is not recommended. Additionally, the interactions included are not chosen to explain something from the reality, they were selected to make a better prediction for our data.

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 2.4602 20.1028 0.122 0.903045

GDP\_pc -2.6592 0.3755 -7.082 2.77e-09 \*\*\*

Gini 0.1786 5.0936 0.035 0.972149

seniors -21.4741 4.0380 -5.318 1.97e-06 \*\*\*

smoking\_rate 22.6417 4.8919 4.628 2.29e-05 \*\*\*

obesity\_rate 4.9194 1.2289 4.003 0.000189 \*\*\*

diabetes\_rate -15.7615 5.9625 -2.643 0.010670 \*

male.female 22.7443 7.1306 3.190 0.002353 \*\*

obesity\_rate:male.female -5.5082 2.0562 -2.679 0.009724 \*\*

GDP\_pc:seniors 1.0294 0.1457 7.065 2.95e-09 \*\*\*

smoking\_rate:obesity\_rate -1.3582 0.4102 -3.311 0.001645 \*\*

Gini:smoking\_rate -5.1389 1.2218 -4.206 9.66e-05 \*\*\*

Gini:seniors 3.6977 0.9191 4.023 0.000177 \*\*\*

Gini:diabetes\_rate 4.2815 1.6395 2.611 0.011600\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6402 on 55 degrees of freedom

Multiple R-squared: 0.6672,

Adjusted R-squared: 0.5885

F-statistic: 8.48 on 13 and 55 DF, p-value: 4.557e-09

**4.e. Application**

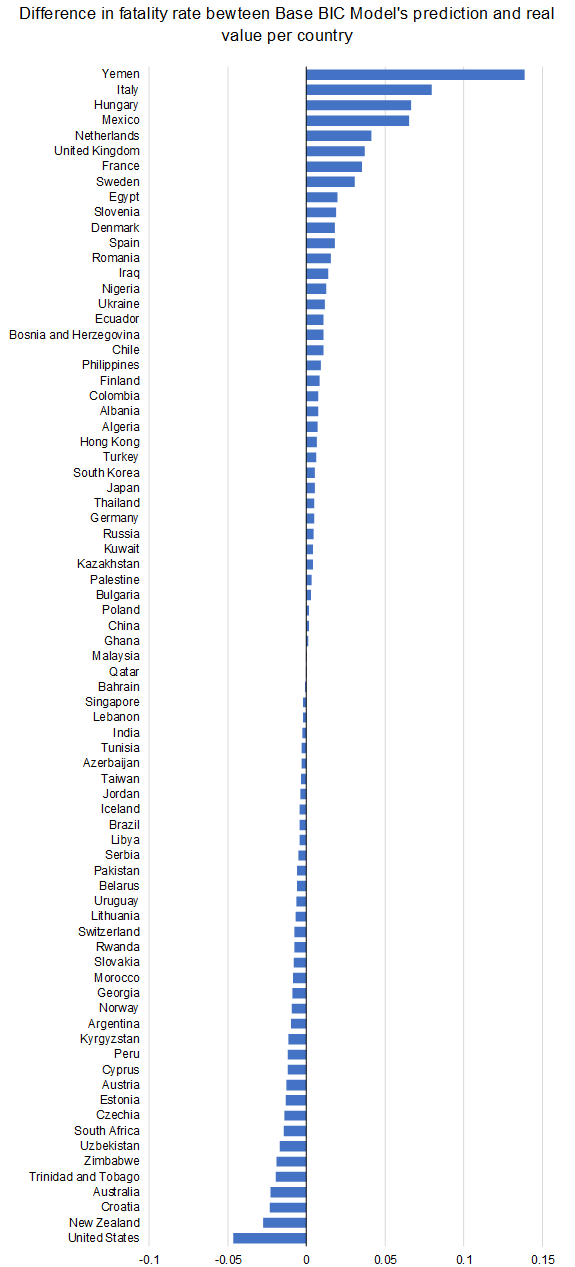
The model constructed could have an application for countries fighting against the pandemic. The model estimates the expected fatality rate given characteristics from 77 different countries. Therefore, for any given country not included in the data set, the model may help us to make predictions for fatality rate and inform decision making in vaccine prioritization.

It is possible to take the actual fatality rate of the country and compare it with the expected fatality rate from the model. If the difference is positive, it may be said that the country has a higher fatality rate than our best estimation. If the difference is negative, it may be inferred that the country has a lower fatality rate than our best estimation. As we got the estimation from different important characteristics in 77 different countries, this value may be interpreted as the success or failure for the strategy against COVID 19 in each country compared with the rest of the regions in our analysis.

It is important to be prudent when doing this interpretation, as the team is aware there are many other factors not included in this analysis. For example, a country that makes a small number of tests but measures all deaths is going to have a big fatality rate. As a result its value would be very positive, and it is not completely accurate to say that its strategy is a failure against the pandemic. However, taking additional considerations to this application could be helpful to understand and compare the advance in COVID 19 between countries.

In the next table it is reflected that Yemen, Italy and Hungary are the countries that have higher fatality rate in comparison with our expected value. On the other hand, United States, New Zealand and Croatia have lower fatality rates than the model prediction. It is interesting to see the case of Sweden, which maintained a very lax policy against COVID 19. Considering all the control variables in our analysis, its fatality rate is higher than the expected.

*Table 9. Difference in fatality rate between Base BIC model’s prediction and real value per country*



1. **Conclusions**

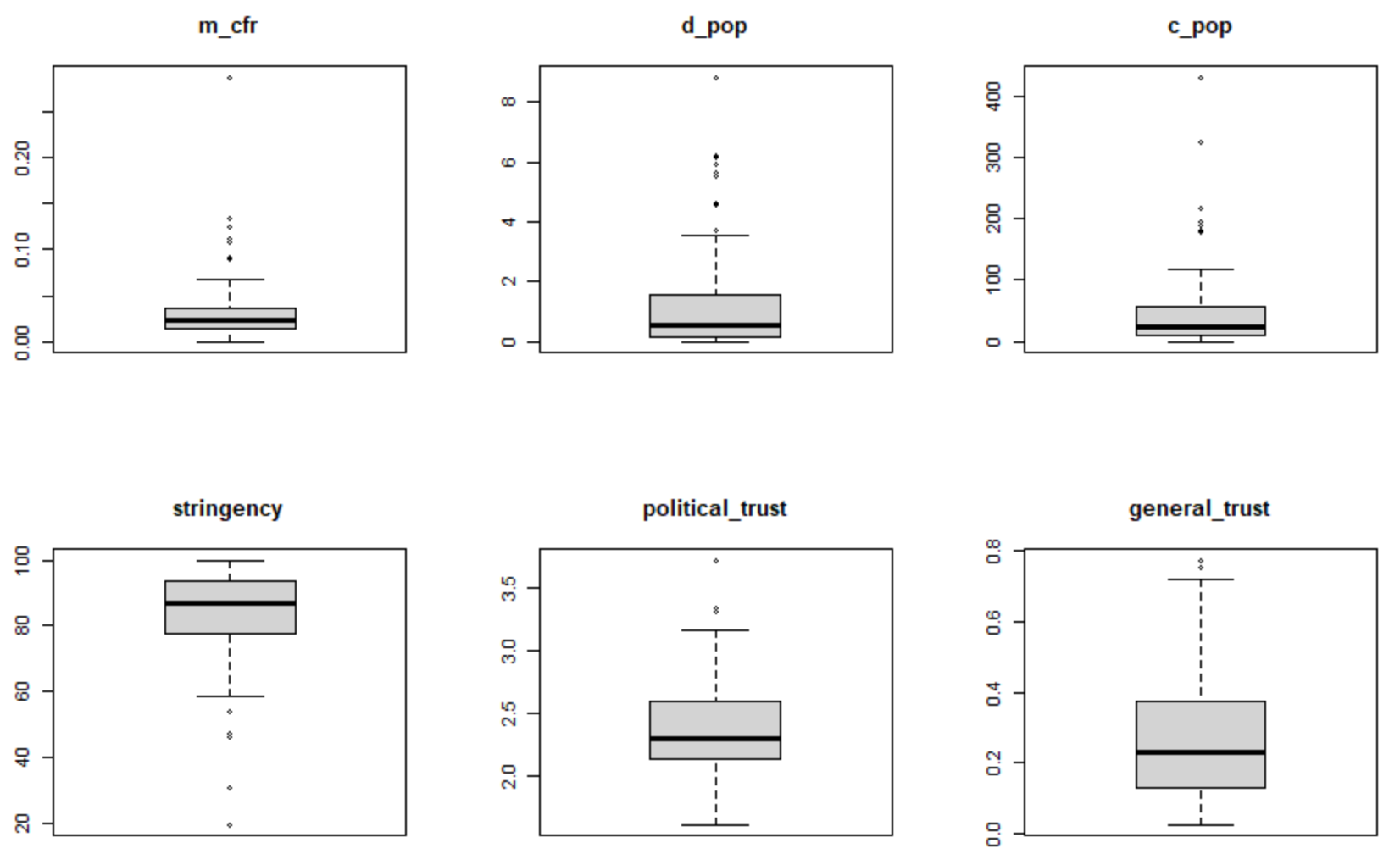
The project covered COVID-19 cases and deaths in August and September 2020 in 77 countries, of which we had its demographic information available. The main objective was to identify which variables are determinant in the fatality rate and to help to identify which countries not included in the data set present higher risk of COVID- 19 fatality rate. In order to provide an answer, a prediction model has been built based on the variables that have the strongest relation with the fatality rate globally.

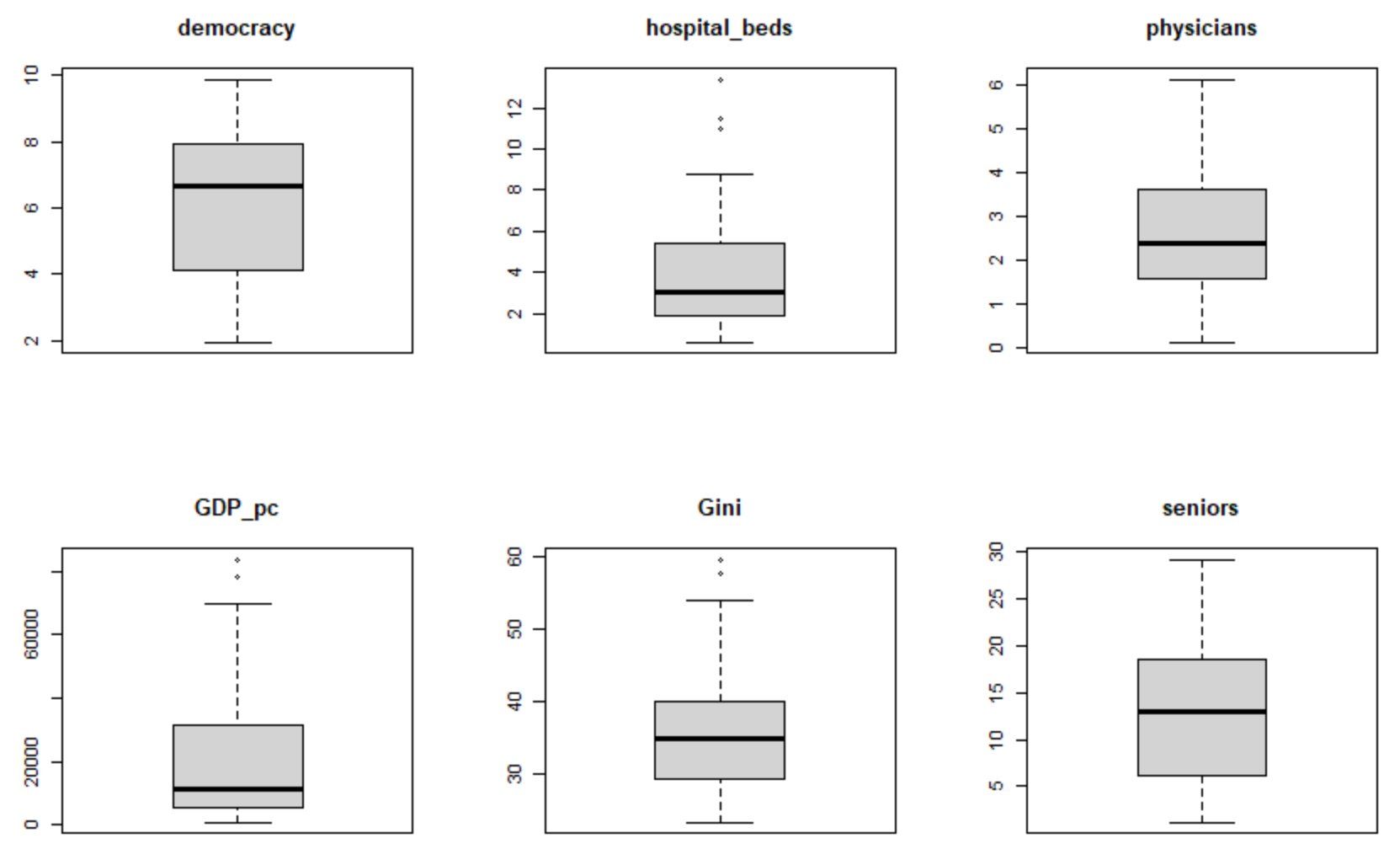
The selected best model predicts fatality rate per country. This could be used as tool to measure the effectiveness to fight COVID-19 in countries not included in the data set and/or inform decision making in vaccine prioritization, with additional considerations. It is relevant to note that interactions do not allow to read the proposed model straight forward, as variables cannot be interpreted by itself.

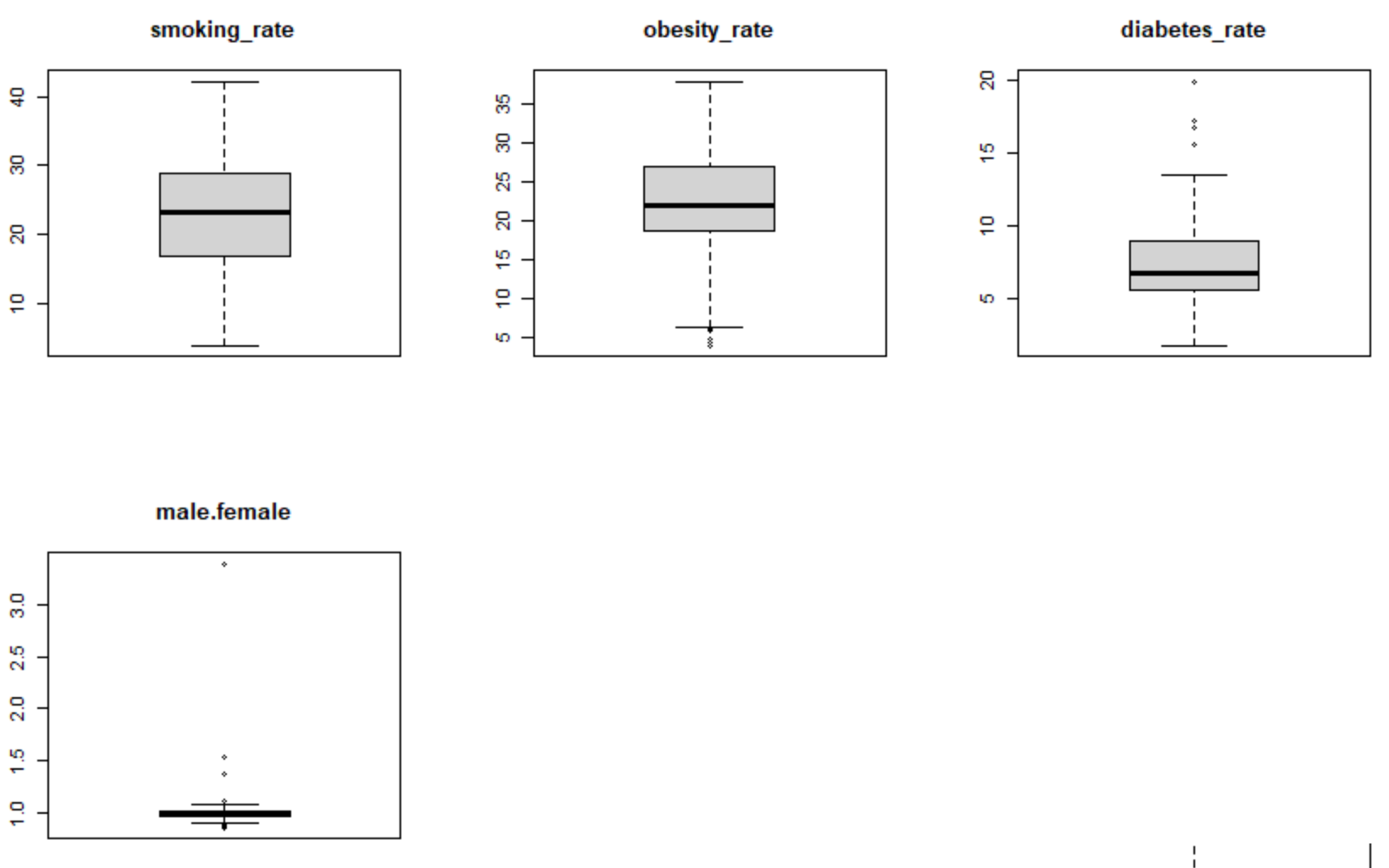
Regarding vaccine prioritization, it has been determined that GDP per capita, seniors per capita, obesity rate and gender rate are the most important determinants of fatality rate among countries. These could help in identifying those countries that could benefit from a sooner immunization in and out of the dataset.

**Appendix**

1. **Boxplot distribution for each variable**

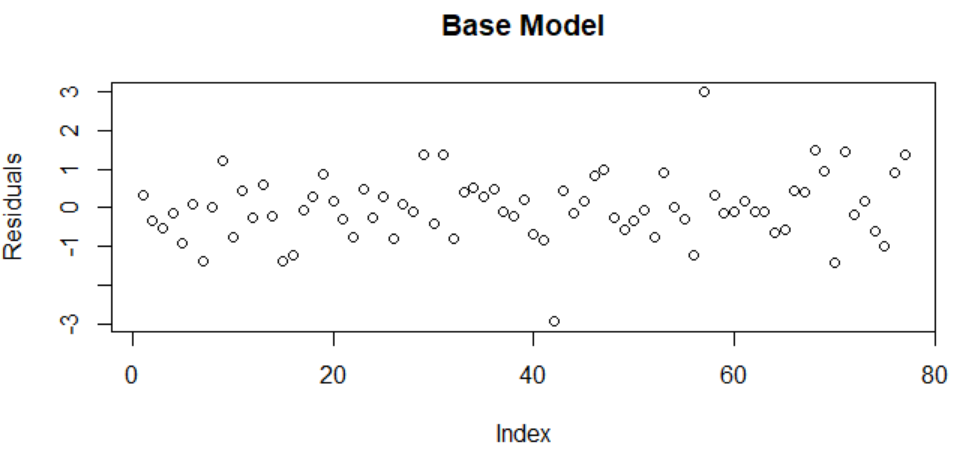
**

**

**

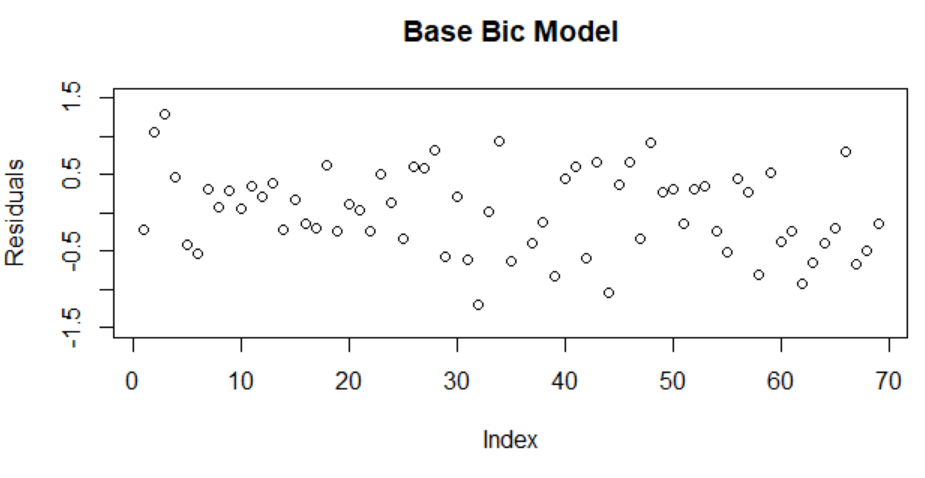
1. **Base Model constant variance assumption**

As we can see in the graph below, it is hard to argue that the residuals of our regression have a non-constant variance. It is true that two values are making some noise, however in general terms we can say that the variance of our model is constant.

****

1. **Base BIC Model constant variance assumption**

As we can see in the graph below, in general terms we can say that the variance for the Base Bic Model is constant. Nonetheless, it is possible to argue that our model presents some heteroscedasticity. It is possible to run a model for non-constant variance, however this would only change the standard errors and p-values for our regression. As the only objective of this model is prediction, it does not make sense to make this adjustment.

****

**References**

Journals

<https://www.sciencedirect.com/science/article/abs/pii/S1525861020304412>

<https://www.thelancet.com/journals/eclinm/article/PIIS2589-5370(20)30208-X/fulltext>

<https://journals.lww.com/ccejournal/fulltext/2020/06000/gender_difference_is_associated_with_severity_of.26.aspx>