Covid-19 Regression Analysis

Motivation

Over the past year, we have been bombarded with Covid-19 statistics provided by the media. With all the news about Covid-19 being thrown around, it is difficult to keep up with how regions outside of the countries we live in are affected by this virus. Although Covid-19 is a deadly virus, our group suspects that not all countries are affected equally by this pandemic. Our project aims to investigate this difference, by looking into different factors affecting both the infection numbers and death from Covid-19 within countries. Specifically, we want to investigate how the factors of GDP, population size and density, and population age affect Covid-19 the number of infection cases and deaths. Furthermore, we want to investigate if a country having a higher number of infections means that country having a higher death rate.

Data

From our data of 184 different countries, we try to explain the number of infections and Covid-19 related deaths within that country using each county's population size (measured in thousands of people), GDP (in millions of US dollars), population density (measured in amount of people per square kilometer), and age distribution (given as the ratio of youth ranging from 0-14 years of age, and the elderly, who are above 60 years of age). We use data from the UN database for the explanatory variables population size, density, and age distribution. These values were recorded in 2017, but we assume that these values would be good enough estimates for 2019, as these parameters would not differ much over the course of two years (we would not want estimates for 2020, as we want values unaffected by Covid-19 to see how the virus impacted countries). Originally we wanted to use the GDP values from the same UN database spreadsheet, but the spreadsheet only provided GDP values from 2017. We want to use GDP as an indicator for a country's wealth before the pandemic started, and so we wanted to find GDP values for countries in 2019. Therefore, we used the GDP values provided by the world bank. Data for the number of infections and deaths were from the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU). We use recorded values for the number of infections and death from January 23, 2020 to November 12, 2020. Below are links to where we got our data:

- Data about covid deaths and infections per day from kaggle: https://www.kaggle.com/antgoldbloom/covid19-data-from-john-hopkins-university
- GDP of countries from the world bank:

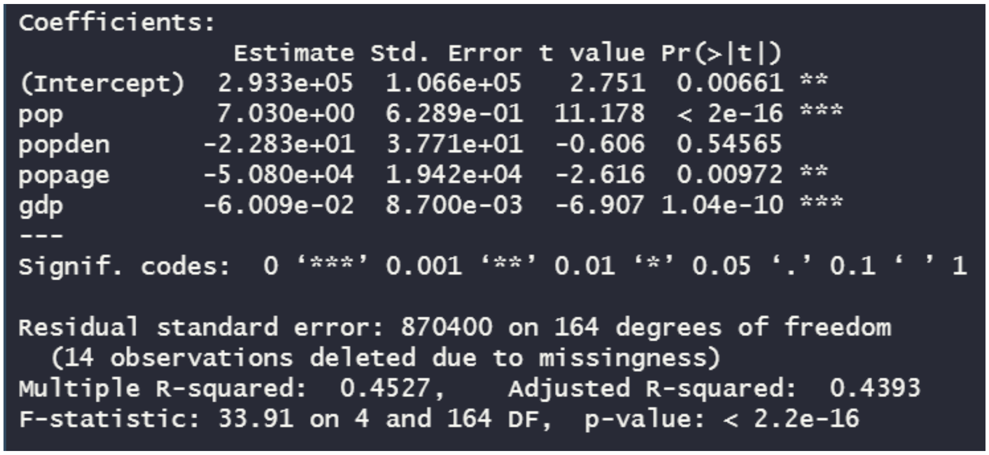
https://databank.worldbank.org/data/download/GDP.pdf

- Population size, density, age distribution
  https://www.kaggle.com/sudalairajkumar/undata-country-profiles

Process

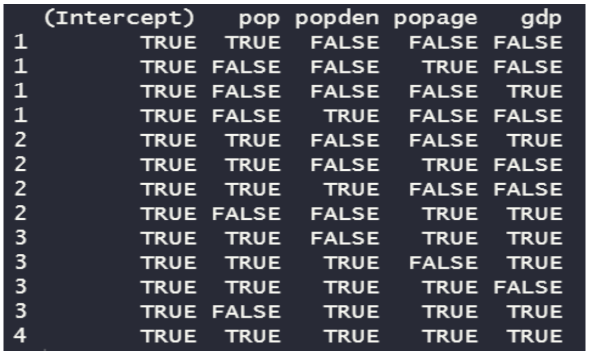
1) Fitting a model for number of Infections

Using R, we first fit a linear model for the number of infections, using all the explanatory variables with no interaction. Looking at the summary, we get:



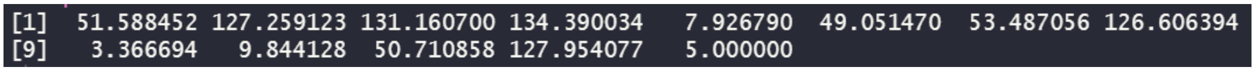
```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.933e+05  1.066e+05   2.751  0.00661 **
pop          7.030e+00  6.289e-01  11.178  < 2e-16 ***
popden      -2.283e+01  3.771e+01  -0.606  0.54565
popage      -5.080e+04  1.942e+04  -2.616  0.00972 **
gdp         -6.009e-02  8.700e-03  -6.907 1.04e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 870400 on 164 degrees of freedom
  (14 observations deleted due to missingness)
Multiple R-squared:  0.4527,    Adjusted R-squared:  0.4393
F-statistic: 33.91 on 4 and 164 DF,  p-value: < 2.2e-16
```

Looking at the estimate for the coefficients, it seems that 3 out of the 4 coefficients are significant at an alpha level of 0.001 or less. Contrary to our expectations, it seems that population density does not affect infection numbers much in our model. As expected, the population parameter is positive, whereas the GDP coefficient is negative. So with a larger population, the number of infections will increase and as the GDP of a country increases, the number of infections decrease. The coefficient for population age is negative, indicating that the more young people in a country relative to older people, the lower the number of infections in a country. We notice that the residual standard error seems really high, but that may be because the units for GDP are so large. Looking at the R squared and the adjusted R squared value, we see that it is fairly low, suggesting that this model doesn't fit the data very well. From here, we try to fit a model with less parameters. Using the regsubsets function, we get:

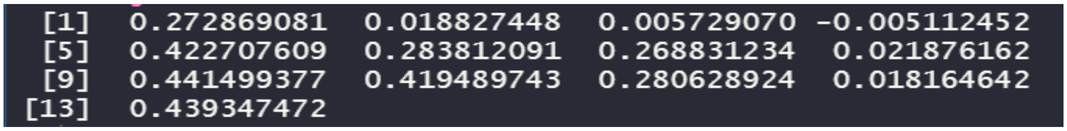


```
  (Intercept)   pop popden popage   gdp
1        TRUE  TRUE  FALSE  FALSE FALSE
1        TRUE FALSE  FALSE   TRUE FALSE
1        TRUE FALSE  FALSE  FALSE  TRUE
1        TRUE FALSE   TRUE  FALSE FALSE
2        TRUE  TRUE  FALSE  FALSE  TRUE
2        TRUE  TRUE  FALSE   TRUE FALSE
2        TRUE  TRUE   TRUE  FALSE FALSE
2        TRUE FALSE  FALSE   TRUE  TRUE
3        TRUE  TRUE  FALSE   TRUE  TRUE
3        TRUE  TRUE   TRUE  FALSE  TRUE
3        TRUE  TRUE   TRUE   TRUE FALSE
3        TRUE FALSE   TRUE   TRUE  TRUE
4        TRUE  TRUE   TRUE   TRUE  TRUE
```

We also take a look at the Cp statistics (top) for each model, along with the adjusted R-squared (bottom):

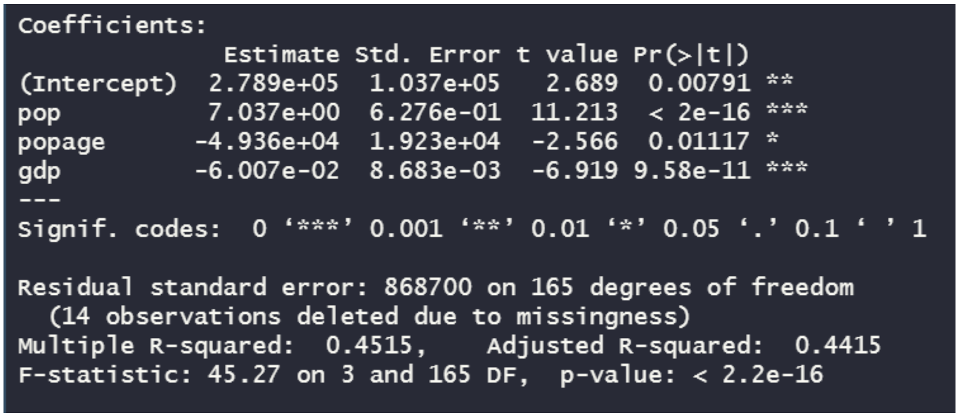


```
[1]  51.588452 127.259123 131.160700 134.390034   7.926790  49.051470  53.487056 126.606394
[9]   3.366694   9.844128  50.710858 127.954077   5.000000
```



```
  [1]  0.272869081  0.018827448  0.005729070 -0.005112452
  [5]  0.422707609  0.283812091  0.268831234  0.021876162
  [9]  0.441499377  0.419489743  0.280628924  0.018164642
 [13]  0.439347472
```

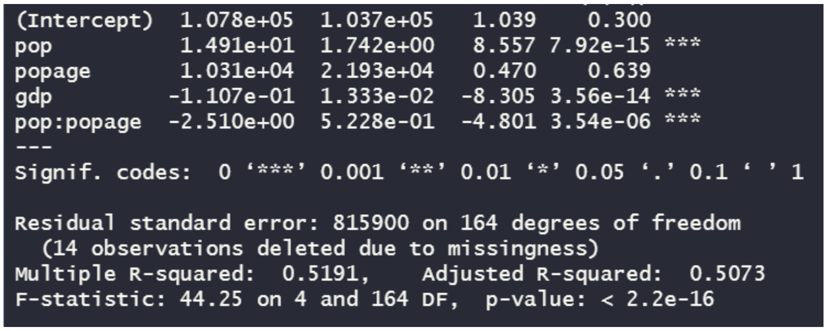
From the vector of Cp statistics, we see that the 9th model fitted has the lowest amount of unexplained error. Furthermore, it is the closest number to the number of parameters the model has (Cp is close to p). Looking at the adjusted R-squared values, we see that it is the highest, suggesting that it fits the model the best. Models with similar R-squared values, like model 13 and 10, have Cp values farther away than the number of parameters the respective model has. Therefore, in the case with no interaction terms, we would choose to fit the 9th model, which has 3 parameters (excluding intercept term). Fitting that model, which just omits the explanatory variable for population density, we get the following summary:



```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.789e+05  1.037e+05   2.689  0.00791 ** 
pop          7.037e+00  6.276e-01  11.213  < 2e-16 ***
popage      -4.936e+04  1.923e+04  -2.566  0.01117 *  
gdp         -6.007e-02  8.683e-03  -6.919 9.58e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 868700 on 165 degrees of freedom
  (14 observations deleted due to missingness)
Multiple R-squared:  0.4515,	Adjusted R-squared:  0.4415 
F-statistic: 45.27 on 3 and 165 DF,  p-value: < 2.2e-16
```

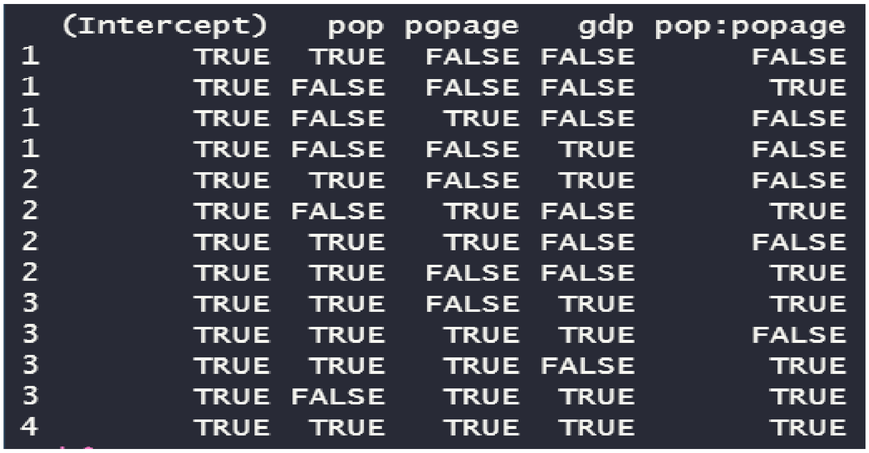
With this reduced model, we see that now every explanatory variable is significant at an alpha level of 0.01. The residual standard error is now a bit lower, and the adjusted R squared value is now a bit higher. However, the adjusted R squared value is still low, so our model does not fit the data too well.

We now consider the possibility that there is interaction between the variables, namely between the population of a country and the age distribution of its citizens. We choose to fit a model with the same explanatory variables present in model 7 (population, population age, and GDP) along with an interaction term between population and age. The summary output from R gives us



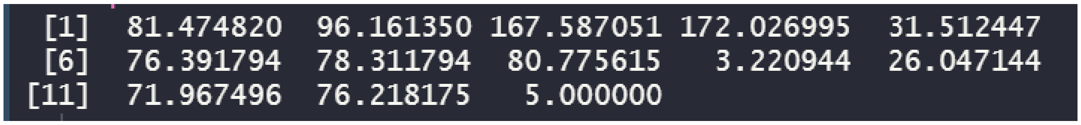
```
(Intercept)  1.078e+05  1.037e+05   1.039    0.300    
pop          1.491e+01  1.742e+00   8.557 7.92e-15 ***
popage       1.031e+04  2.193e+04   0.470    0.639    
gdp         -1.107e-01  1.333e-02  -8.305 3.56e-14 ***
pop:popage  -2.510e+00  5.228e-01  -4.801 3.54e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 815900 on 164 degrees of freedom
  (14 observations deleted due to missingness)
Multiple R-squared:  0.5191,    Adjusted R-squared:  0.5073 
F-statistic: 44.25 on 4 and 164 DF,  p-value: < 2.2e-16
```

We see that this new model, which includes interaction, has a higher adjusted R-squared value than the previous model we fitted, suggesting that this new model fits the data better. The interaction term is highly significant, which suggests the term has significant impact on the response and is worth keeping. Once again, we use the regsubsets command to look at alternative models:

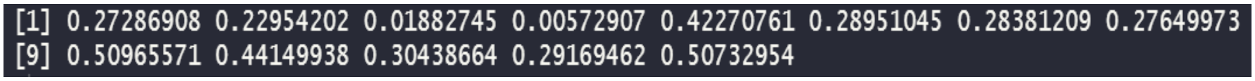


```
  (Intercept)   pop popage   gdp pop:popage
1        TRUE  TRUE  FALSE FALSE      FALSE
1        TRUE FALSE  FALSE FALSE       TRUE
1        TRUE FALSE   TRUE FALSE      FALSE
1        TRUE FALSE  FALSE  TRUE      FALSE
2        TRUE  TRUE  FALSE  TRUE      FALSE
2        TRUE FALSE   TRUE FALSE       TRUE
2        TRUE  TRUE   TRUE FALSE      FALSE
2        TRUE  TRUE  FALSE FALSE       TRUE
3        TRUE  TRUE  FALSE  TRUE       TRUE
3        TRUE  TRUE   TRUE  TRUE      FALSE
3        TRUE  TRUE   TRUE FALSE       TRUE
3        TRUE FALSE   TRUE  TRUE       TRUE
4        TRUE  TRUE   TRUE  TRUE       TRUE
```

We also explore the Cp values (top image) and the adjusted R-squared values (bottom):



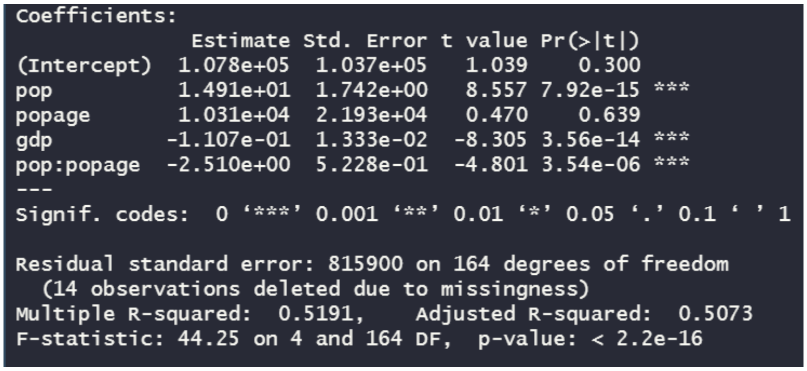
```
 [1]  81.474820  96.161350 167.587051 172.026995  31.512447
 [6]  76.391794  78.311794  80.775615   3.220944  26.047144
[11]  71.967496  76.218175   5.000000
```



```
[1] 0.27286908 0.22954202 0.01882745 0.00572907 0.42270761 0.28951045 0.28381209 0.27649973
[9] 0.50965571 0.44149938 0.30438664 0.29169462 0.50732954
```

From the adjusted R-squared values, we see that model 9 and 13 have the highest values, so we assume these models fit the data the best. Looking at the Cp statistics, model 9 has a Cp value of 3.22 and model 13 has a Cp value of 5. As model 9 has 3 parameters and model 13 has 4, we would choose model 9 over model 13 as its Cp value is closer to the number of parameters than model 13. However, model 9 includes the interaction parameter without one of the main variables, population age. We want to avoid this, so we choose the next best model, which is model 13.

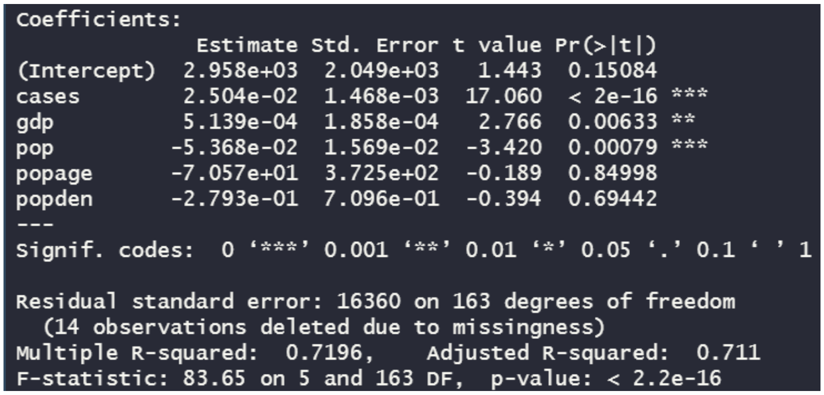
Based on these observations, we choose to explain the number of infections within a country with the model with interactions,using model 9 from above. The summary for this model is:



```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.078e+05  1.037e+05   1.039    0.300
pop          1.491e+01  1.742e+00   8.557 7.92e-15 ***
popage       1.031e+04  2.193e+04   0.470    0.639
gdp         -1.107e-01  1.333e-02  -8.305 3.56e-14 ***
pop:popage  -2.510e+00  5.228e-01  -4.801 3.54e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 815900 on 164 degrees of freedom
  (14 observations deleted due to missingness)
Multiple R-squared:  0.5191,    Adjusted R-squared:  0.5073
F-statistic: 44.25 on 4 and 164 DF,  p-value: < 2.2e-16
```

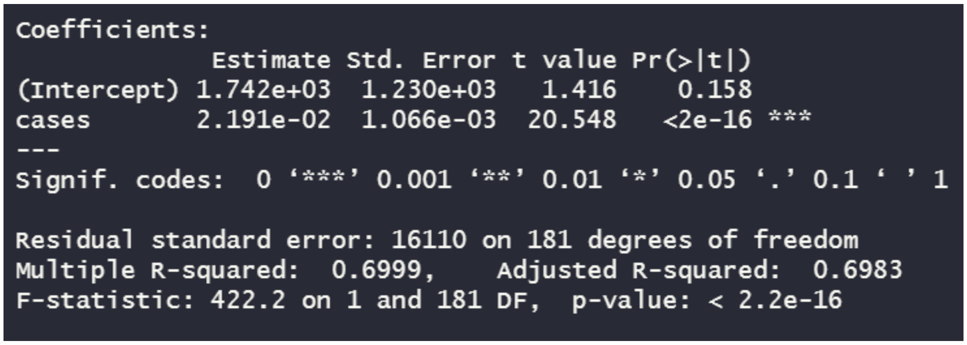
2) Fitting a model for number of Covid-related deaths

We follow a similar procedure in fitting a model as above. However, for this section we also add the number of infections as one of the explanatory variables, as we are interested if the number of covid-related infections affects the number of deaths. First, we fit a model with no interaction:



```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.958e+03  2.049e+03   1.443  0.15084
cases        2.504e-02  1.468e-03  17.060  < 2e-16 ***
gdp          5.139e-04  1.858e-04   2.766  0.00633 **
pop         -5.368e-02  1.569e-02  -3.420  0.00079 ***
popage      -7.057e+01  3.725e+02  -0.189  0.84998
popden      -2.793e-01  7.096e-01  -0.394  0.69442
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 16360 on 163 degrees of freedom
  (14 observations deleted due to missingness)
Multiple R-squared:  0.7196,    Adjusted R-squared:  0.711
F-statistic: 83.65 on 5 and 163 DF,  p-value: < 2.2e-16
```

Looking at the R-squared and adjusted R-squared values, both are around 0.71, so we assume this simple model fits the data decently. Our residual standard error is also pretty high, and we continue to suspect that it is due to GDP values being large. From the summary, we see that only the explanatory variables for number of covid infections, GDP, and population size are significant. This is surprising, as our group would have thought that population age parameters, representing the ratio between young and old people in a country, would contribute more to explaining deaths due to covid. Another interesting observation is that the parameter for population is negative, implying that according to our model more people in a country results in less deaths. Also, the parameter for GDP is positive, implying that countries with higher GDP have higher death rates due to covid. This contradicts our intuition, as our group expected that a country with higher GDP would have more means to prevent deaths. We suspect that this result is due to differing policies in how each country deals with covid cases. From this viewpoint, there seems to be a difference in how wealthier countries deal with covid versus how less wealthy countries deal with covid (we use GDP as an indicator for wealth). We infer this difference from the difference in deaths from covid between wealthy and poor countries, which we see from the negative GDP parameter above for our model (which is highly significant under the null hypothesis that this parameter is non-zero). There also seems to be a difference in how countries with larger populations deal with covid versus countries with smaller populations, as seen from the population parameter in our model above (this parameter is also highly significant under the null hypothesis of this parameter being non-zero). As we suspect that there seems to be some confounding between population size, GDP, and a country’s covid policy, we are hesitant to continue using GDP and population as explanatory variables in modeling covid related deaths. We do not go more in-depth about how policy, GDP, and population size affect one another, as the main focus of this project is to assess how well our predetermined factors of GDP, population size and density, and age ratio explain covid infections and deaths. So in this case, it seems that none of our predetermined variables can explain the response adequately, despite the model having a decent R-squared and adjusted R-squared term. We could still fit a model to explain the number of covid deaths based solely on the number of infections:



```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.742e+03  1.230e+03   1.416    0.158
cases       2.191e-02  1.066e-03  20.548   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 16110 on 181 degrees of freedom
Multiple R-squared:  0.6999,    Adjusted R-squared:  0.6983
F-statistic: 422.2 on 1 and 181 DF,  p-value: < 2.2e-16
```

From the summary above, we see that even in this overly simplified model, the number of cases is indeed strongly associated with the number of deaths, as the cases parameter is highly significant. The R-squared and adjusted R-squared are around 0.7, which is a similar value to the previous model fitted. This suggests that in the previous model, most of the data was already explained by this variable, the number of covid infection cases. We do not adopt this model as the best at explaining the response, but we use it to verify that there is indeed a strong association between the number of covid related deaths and the number of covid related cases.

Conclusion

For the number of Covid infection cases, we were able to fit a model using the explanatory variables population, population age, GDP, and an interaction term between the population and the population age. Most of these terms were significant under the null hypothesis, besides the intercept term and population age. However, for this model the adjusted R-squared value is not too high (around 0.5), suggesting that the model does not fit the data very well.

For the number of Covid death cases, we were unable to fit a suitable model with the predetermined explanatory variables. This was due to potential confounding between GDP, population size, and Covid-19 policies for each country. The other two explanatory variables we had were not significant under the null hypothesis that the parameters are non-zero. However, we were able to discover that the number of Covid-19 cases is strongly associated with the number of deaths.

In both examples, we feel that using more explanatory variables would help explain the data more. One such variable we could include, if we were to do this project again, would be geographical location, indicated by latitude and longitude. Perhaps colder countries and warmer countries would handle the pandemic differently, affecting the number of infections and deaths. Another variable we could include would be how fast a country closed its borders since the pandemic started. Movement of people would spread disease, and we hypothesize that countries that had faster lockdown procedures would have less infection rates and deaths.

Overall, we discovered that the explanatory variables of GDP, population size, population density, and age ratio of a country are not sufficient to adequately model either the number of infections due to covid or the number of covid-related deaths. Using just these variables simplifies the situation too much, and we would investigate using more relevant explanatory variables to hopefully fit a better model if we were to redo this project.

We first investigate into the relationship between number of infections and county's population size (measured in thousands of people), GDP (in millions of US dollars), population density (measured in amount of people per square kilometer), and age distribution (given as the ratio of youth ranging from 0-14 years of age, and the elderly, who are above 60 years of age). Our fitted model shows that with a larger population, the number of infections will increase and as the GDP of a country increases, the number of infections decreases, also the larger the young-and-old ratio, the lower the number of infections in a country. The relationship indicates that with a larger population there might be more people gathered causing more infections, with a higher GDP there will be more means to deal with the Covid-19 varius, and young people are less likely to get infected than older people. However, soon we find out that the value of R squared and the adjusted R squared are relatively low, which indicates this model does not fit the data well. So we fit the data with a reduced model which has less parameters by omitting the population density variable. In this new model we find out that there might be interactions between variables. After we add an interaction term between population and age, the model fits better. By doing the regsubsets again, the model shows that the number of infections is related to population, GDP and an interaction term between population and age.

For the second model which is about Covid related number of deaths, we fitted it with no interactions as we are interested if the number of covid-related infections affects the number of deaths. In this model, only the explanatory variables for number of covid infections, GDP, and population size are significant. It is beyond our expectation that age ratio doesn't contribute much. The death rate of infection people may be similar among all age groups. Basically, in this model, the results contradict our intuition. Countries with more people resulted in less death and countries, while countries with higher GDP resulted in more death cases. Although both the R square and adjusted R square are convincing, predetermined factors of GDP, population size and density, and age ratio seems cannot correctly reflect the response adequately. Our thoughts are when referring to the death cases, it is not like the infection cases as it is linked to every country's policies and actions. It seems that the only correlation is confirmed covid cases and the death cases. We finally decided to simplify the second model, only the covid infection cases and death cases would be considered. As said, the number of cases is strongly associated with the number of deaths.

By doing the study, we found out that there is no variable that would dominate the final death cases caused by Covid-19, no matter the GDP, age ratio and people density. We appeal that the government or health agency would find an efficient and effective way to stop preventing the pandemic and cure the confirmed cases.