Final Project: Public Safety Measures and Public Health for Covid-19

Data 603 - Statistical Modelling with Data

Paul Croome, Rodrigo Rosales Alvarez, Ann Siddiqui and Kane Smith

2022-12-14

Professor: Dr. Paul Galpern
University of Calgary
Calgary, Alberta

Contents

Introduction 3

Research Questions . . . 3

Data Set Definition 4

Methodology 5

Libraries . . . 5

Data Import . . . 5

Data Cleaning . . . 5

Variable Definition. . . . 5

Data Preparation . . . 6

Multiple Linear Regression Models for Research Question 1 . . . 6

New Cases Model . . . 7

New Deaths Model . . . 10

Results 11

Discussion 12

References 13

Introduction

The domain of our project covers healthcare-related indicators of the wellbeing of countries during the coronavirus disease 2019 (COVID-19) pandemic. In particular, we will be examining data related to the prevalence and severity of the COVID-19 pandemic and the governmental and societal measures taken to reduce the spread of the disease. These data were all daily reported between January 2020 and October 2022.

This is an interesting and important topic of study because, in our increasingly interconnected world, contagious diseases can be transmitted over vast distances remarkably easily. Even small, remote outbreaks of diseases anywhere in the world can swiftly turn into a global pandemic, which can then cause devastation on personal, societal, and worldwide scales.

Research Questions

1) What population-related metrics of countries around the world are most strongly related to the prevalence and severity of COVID-19 experienced in a country between February 2020 and October 2022 (as measured by average daily COVID-19 cases and deaths)?

    a) What is the best model that can be built from these data for predicting the average daily new COVID-19 cases experienced in a country?

    b) What is the best model that can be built from these data for predicting the average daily new COVID-19 deaths experienced in a country?

2) Among countries with reliably reported data relating to cases, positive test rates, vaccinations, and boosters, what societal and governmental responses to the COVID-19 pandemic are most strongly related to the prevalence and severity of COVID-19 experienced in a country between February 2020 and October 2022 (as measured by average daily COVID-19 cases and deaths)?

    a) What is the best model that can be built from these daily-reported data for predicting the daily new COVID-19 cases in a country?

    b) What is the best model that can be built from these daily-reported data for predicting the daily new COVID-19 deaths in a country?

Data Set Definition

The dataset we will use consists of diverse information related to the COVID-19 pandemic, including a country's daily rates of COVID-19 diagnoses, hospitalizations, deaths, vaccinations, and booster shots. We will use features of these data to determine the prevalence and severity of the COVID-19 pandemic for each country . The dataset consists of daily information from January 1, 2020 to October 26, 2022 for more than 220 countries; each row corresponds to the Covid-19 information reported by an specific country in a certain date.

This dataset is in tabular form contained in a CSV file and is licensed for open access under the Creative Commons BY license. The dataset was put together by Our World in Data; more importantly, the data set is being updated daily by the same organization, for more information about the data pipeline and how the data set is being maintained click here.

Methodology

Libraries

Data Import

As our data is in CSV format, we simple use the function **read_csv()** to import our file into a data frame.

```
covid_raw <- read_csv('data.csv', show_col_types = FALSE)
```

Data Cleaning

Many countries and facilities are under reporting Covid-19 statistics like cases and deaths, according to Claire Klobucista from Council of Foreign Relations. This is could be catastrophically as government could not respond accordingly to the real situation. For our research this is very important as well, if we put bad data into our model, we would create a bad model. To solve this problem we will remove from our dataset the countries that are present in the bottom 5% of number of new cases smoothed per million and number of new deaths smoothed per million.

```
# Getting countries with bottom 5% of new_cases_smoothed_per_million
tenth_percentile_cases <- quantile(covid_raw$new_cases_smoothed_per_million, probs = 0.05, na.rm = TRUE
bad_country_cases <- covid_raw[covid_raw$new_cases_smoothed_per_million < tenth_percentile_cases,]
bad_country_list_cases <- unique(bad_country_cases$location)

# Getting countries with bottom 5% of new_deaths_smoothed_per_million
tenth_percentile_deaths <- quantile(covid_raw$new_deaths_smoothed_per_million, probs = 0.05, na.rm = TRU
bad_country_deaths <- covid_raw[covid_raw$new_deaths_smoothed_per_million < tenth_percentile_deaths,]
bad_country_list_deaths <- unique(bad_country_deaths$location)

# Remove countries that appear in either above lists
bad_country_list <- append(bad_country_list_deaths, bad_country_list_cases)
good_countries <-   covid_raw[!covid_raw$location %in% bad_country_list, ]

covid_data <- good_countries
```

More cleaning tasks were missing, for starters we generated a new column called *smokers* that was the average of *female_smokers* and *male_smokers*, after that we replaced all the null values for 0s, as we can´t assign a number to a factor, we decided to drop continent and iso_code, columns, test_unitsand date as are not important for our analysis.

```
# Creating the column "smokers"
covid_data$smokers <- (covid_data[['male_smokers']] + covid_data[['female_smokers']]) / 2

# Drop column continent
covid = subset(covid_data, select = -c(iso_code, continent, tests_units, date) )

# Changing Null Values to 0s
covid[is.na(covid)] = 0
```

Variable Definition.

Independent Variable

- *new_cases*: new confirmed cases of COVID-19. Continuous Variable.
- *new_deaths*: new deaths attributed to COVID-19. Continuous Variable.

Dependent Variables

1) Population metrics

   - *extreme_poverty*: The number of the population per million that is considered to be in extreme poverty.
   - *gdp_per_capita*: GDP per capita of the country.
   - *median_age*: Median age of the population.
   - *population*: Number of people in the country.
   - *human_development_index*: Human development index as calculated by Human Development Reports.
   - *population_density*: Population density of the country.
   - *aged_65_older*: The number of the population per million that is aged over 65 years old.

2) Health metrics

   - *cardiovasc_death_rate*: The death rate caused by cardiovascular disease.
   - *diabetes_prevalence*: The number of the population per million that is diagnosed with diabetes.
   - *life_expectancy*: The life expectancy of the population of a country.
   - *reproduction_rate*: The rate of reproduction of the population of a country.
   - *smokers*: The number of the population per million that smokes cigarettes.

3) COVID metrics

   - *stringency_index*: A measure of how stringent the policies related to controlling the spread of COVID is.
   - *hosp_patients*: The number of people hospitalized due to COVID.
   - *new_tests*: The number of new COVID tests conducted in a day.

Data Preparation

To create the finala dataset that will be used to create the *New Cases Model* and *New Deaths Model* we performed an aggregation function (mean) across the variable country as we are only interested in having one data point per country. The new data point will be the mean of every other variable present in the table; we decided to use the mean as that is the best way to aggregate the variables that we are interested on using like new cases, new deaths, population, stringency index and more.

```
covid_agg <- covid %>% group_by(location) %>% summarise(across(everything(), mean), .groups = 'drop') %>
```

Multiple Linear Regression Models for Research Question 1

Our research question number 1 is:

1) What population-related metrics of countries around the world are most strongly related to the prevalence and severity of COVID-19 experienced in a country between February 2020 and October 2022 (as measured by average daily COVID-19 cases and deaths)?

   a) What is the best model that can be built from these data for predicting the average daily new COVID-19 cases experienced in a country?

b) What is the best model that can be built from these data for predicting the average daily new COVID-19 deaths experienced in a country?

For this reason we are going to build two Multiple Linear Regressions Models, one for New Cases and one for New Deaths, using the tools we learnt during class.

New Cases Model

We started by defining our full model, including all the variables that make senses to predict New Covid-19 Cases for a specific country. Using the summary() function we are able to see the most important information about our model.

```
model_cases_full = lm(new_cases ~ aged_65_older + smokers + cardiovasc_death_rate + diabetes_prevalence

summary(model_cases_full)
```

```
##
## Call:
## lm(formula = new_cases ~ aged_65_older + smokers + cardiovasc_death_rate +
##     diabetes_prevalence + extreme_poverty + gdp_per_capita +
##     median_age + life_expectancy + population + stringency_index +
##     human_development_index + reproduction_rate + hosp_patients +
##     new_tests + population_density, data = covid_agg)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -135070   -3339    -436    2256  291891
##
## Coefficients:
##                             Estimate      Std. Error t value
## (Intercept)             17429.787503478  6955.001085235   2.506
## aged_65_older              74.657491834   729.011284481   0.102
## smokers                    41.781739261   204.451708918   0.204
## cardiovasc_death_rate     -10.166601889    18.038661647  -0.564
## diabetes_prevalence       -69.070879646   425.366481995  -0.162
## extreme_poverty           -39.075079934   135.376174301  -0.289
## gdp_per_capita              0.040382206     0.139989919   0.288
## median_age                 15.705160162   364.108739242   0.043
## life_expectancy          -208.614794627   118.703736147  -1.757
## population                  0.000063233     0.000003123  20.248
## stringency_index         -173.926503142   137.954144230  -1.261
## human_development_index  5300.552157460 13117.322191405   0.404
## reproduction_rate        3377.515261862 10219.756466058   0.330
## hosp_patients               1.858607913     0.905939370   2.052
## new_tests                  -0.017359408     0.031050257  -0.559
## population_density          0.091345018     1.119234406   0.082
##                         Pr(>|t|)
## (Intercept)               0.0129 *
## aged_65_older             0.9185
## smokers                   0.8383
## cardiovasc_death_rate     0.5736
## diabetes_prevalence       0.8711
## extreme_poverty           0.7731
```

```
## gdp_per_capita                         0.7732
## median_age                             0.9656
## life_expectancy                        0.0802 .
## population              <0.0000000000000002 ***
## stringency_index                       0.2087
## human_development_index                0.6865
## reproduction_rate                      0.7413
## hosp_patients                          0.0413 *
## new_tests                              0.5767
## population_density                     0.9350
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 30250 on 232 degrees of freedom
## Multiple R-squared:  0.6932, Adjusted R-squared:  0.6734
## F-statistic: 34.95 on 15 and 232 DF,  p-value: < 0.00000000000000022
```

Using the *Step Wise Regression Procedure* to create the best fit additive model we found out that only 3 variables were significant; for these 3 terms we can reject reject the Null Hypothesis for the individual coefficient t test, meaning that the terms are significant and the coefficients should be different than 0. The 3 terms that passed the mentioned test are: **population**, **hosp_patients**, **life expectancy**.

Hypothesis Testing for Individual Coefficient t tests.
$H_0 : \beta_i = 0$
$H_a : \beta_i \neq 0$
(i = AGE, NUMBIDS)

```
model_cases_stepwise = ols_step_both_p(model_cases_full, pent=0.1, prem=0.3, progress=TRUE)
```

```
## Stepwise Selection Method
## ---------------------------
##
## Candidate Terms:
##
## 1. aged_65_older
## 2. smokers
## 3. cardiovasc_death_rate
## 4. diabetes_prevalence
## 5. extreme_poverty
## 6. gdp_per_capita
## 7. median_age
## 8. life_expectancy
## 9. population
## 10. stringency_index
## 11. human_development_index
## 12. reproduction_rate
## 13. hosp_patients
## 14. new_tests
## 15. population_density
##
## We are selecting variables based on p value...
##
## Variables Entered/Removed:
##
```

```
## - population added
## - hosp_patients added
## - life_expectancy added
##
## No more variables to be added/removed.
##
##
## Final Model Output
## ------------------
##
##                            Model Summary
## ----------------------------------------------------------------------
## R                        0.829       RMSE                   29759.194
## R-Squared                0.688       Coef. Var                276.282
## Adj. R-Squared           0.684       MSE                885609652.468
## Pred R-Squared           0.527       MAE                     8744.962
## ----------------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                       ANOVA
## ------------------------------------------------------------------------------------
##                        Sum of
##                       Squares         DF         Mean Square       F         Sig.
## ------------------------------------------------------------------------------------
## Regression    475827434069.052         3    158609144689.684    179.096    0.0000
## Residual      216088755202.299       244       885609652.468
## Total         691916189271.352       247
## ------------------------------------------------------------------------------------
##
##                                        Parameter Estimates
## ----------------------------------------------------------------------------------------------------
##           model          Beta    Std. Error    Std. Beta      t        Sig        lower        upper
## ----------------------------------------------------------------------------------------------------
##     (Intercept)    14508.968      6525.407                  2.223    0.027     1655.652    27362.283
##      population        0.000         0.000        0.798    21.374    0.000        0.000        0.000
##   hosp_patients        1.713         0.616        0.100     2.783    0.006        0.501        2.926
## life_expectancy     -192.669        91.168       -0.079    -2.113    0.036     -372.246      -13.092
## ----------------------------------------------------------------------------------------------------
```

The model obtained after performing the *Step Wise Regression Procedure* is:
$\widehat{y} = 14508.9677 + 0.0001x_{population} + 1.7133x_{hosp_patients} - -192.6691x_{life_expectancy}$

Adjusted R Squared of our model is: 0.6839, meaning that the proportion of the total variation that is explained by the model is 68.39%.

```r
model_cases_stepwise = lm(new_cases ~ population + hosp_patients + life_expectancy, data=covid_agg)
summary(model_cases_stepwise)
```

```
##
## Call:
## lm(formula = new_cases ~ population + hosp_patients + life_expectancy,
##     data = covid_agg)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -134366   -2921    -267    1190  293994
##
## Coefficients:
##                        Estimate      Std. Error t value            Pr(>|t|)
## (Intercept)     14508.967709997  6525.406746098   2.223             0.02710
## population          0.000063892     0.000002989  21.374 < 0.0000000000000002
## hosp_patients       1.713312947     0.615663425   2.783             0.00581
## life_expectancy  -192.669068031    91.168336956  -2.113             0.03559
##
## (Intercept)     *
## population      ***
## hosp_patients   **
## life_expectancy *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 29760 on 244 degrees of freedom
## Multiple R-squared:  0.6877, Adjusted R-squared:  0.6839
## F-statistic: 179.1 on 3 and 244 DF,  p-value: < 0.00000000000000022
```

New Deaths Model

Results

Discussion

References

Klobucista, Claire (2021, May 10). By How Much Are Countries Underreporting COVID-19 Cases and Deaths?. Council of Foreign Relations. https://www.cfr.org/in-brief/how-much-are-countries-underreporting-covid-19-cases-and-deaths