# Key Determinants of Confirmed COVID-19 Cases

Kelsey Bell

INTAF 803: Multi-Sector and Quantitative Analysis

Dr. Johannes Fedderke

May 6, 2020

## Introduction

The current coronavirus pandemic has shaken the world at its very core, caring serious implications with it as countries confirm more and more cases of COVID-19. It's particularly put an strain on countries' health systems, as well as their economies and societies at large as various consequences have occurred since the outbreak began. But disease outbreaks are nothing new. Since globalization has made cross-border travel and other forms of international exchange easier than ever before, the world has been faced with its fair share of disease outbreaks. Just in recent years, countries have seen diseases like Ebola, MERS, and many others spread across borders at an alarming rate. However, the coronavirus epidemic has largely outweighed these previous disease outbreaks, as confirmed case reach into the millions of people who test positive for COVID-19. Why is this the case? To gain some insight into this matter, this paper aims to test what aspects prove to be determinants of confirmed cases in this ongoing coronavirus pandemic.

## Key Determinants

Based on a literature review and existing theories, this section discusses what cross-sectional data is believed to be indicators of country's confirmed cases of COVID-19.

*Geographic Indicators:* As infectious diseases spread from person to person, the geography of a country is expected to have considerable contributions to how fast that spread occurs, impacting the overall number of confirmed cases that will appear. Nicogossian (2014) argues that "important contributors to the spread of infections include" factors such as "urban disparities, poor sanitation, and crowding," which "all act as amplifiers for infection."[1] This is because as people live in densely populated areas, they come into contact with more people and can face a greater risk of being infected, especially from strangers. In addition, the overall territorial size of a country might impact the number of confirmed cases and how fast it spreads domestically.

H1: Geographic indicators have a positive relationship with confirmed coronavirus cases.

*Demographic Indicators:* According to the Centers for Disease Control and Prevention, certain groups of the population are at a greater risk of contracting zoonotic diseases because of the nature of their immune systems. This is considerately important to any study on the determinants of the COVID-19 pandemic because it is a zoonotic disease. Though anyone can get infected, "people are more likely than others to

[1] Nicogossian, A., E. J. Septimus, O. Kloiber, B. Stabile, and T. Zimmerman (2014). 'Spread of Infections and Global Health Security', *World Medical and Health Policy 6*, 329-330.

get really sick, and even die, from infection with certain diseases," including those that are zoonotic.[2] One of these parties includes people over the age of 65. "Early research shows that older people are twice as likely to have serious complications if they get COVID-19," meaning they are at a higher risk of developing severe symptoms.[3] Severer symptoms increase the likelihood that these individuals will go to the hospital to be treated, and thus, tested for coronavirus. Therefore, countries with a higher percent of their population that aged 65 or older, as well as a bigger total population, may have a greater number of confirmed coronavirus cases.

H2: Demographic indicators have a positive relationship with confirmed coronavirus cases.

*Health Infrastructure Indicators:* Analyses on many infectious diseases have helped health policy researchers gain a better understanding on how a country's health system impacts the number of cases, as well as the spread, of infectious diseases. When looking at "Ebola in Africa and the Entrevirus (EV-D68) respiratory infection among U.S. children," Nicogossian (2014) found that "in both cases a major concern is the imposed strain on the existing medical infrastructure, and the ability to mount a rapid response" during the outbreaks.[4] As the coronavirus continues to spread, it is very clear that it's put considerable strain on health systems throughout the world. The sheer volume of cases at the peak of this crisis made it impossible for health professionals to test everyone claiming to be showing symptoms. Kieny (2014) noticed just how the quality of health systems impacted each country's response capabilities during the 2014 Ebola outbreak in Western Africa. "At the time the outbreak began, the capacity of the health systems in Guinea, Liberia and Sierra Leone was limited" so that "several health-system functions that are generally considered essential were not performing well."[5] These variables are expected to have a positive relationship with total confirmed coronavirus cases because a higher expenditure and service coverage score means a country has invested a considerable amount of money into its health system, increasing its ability to test for COVID-19.

In addition to these two general health infrastructure factors, it is also important for a country to have the mechanisms in place that can specifically detect the kinds of disease that the coronavirus falls under. As a zoonotic disease, the coronavirus was initially spread from animal to human before humans could share it among their own species.[6] With the growing presence of zoonotic disease, it important to

---

[2] Centers for Disease Control and Prevention (2017). 'One Health: Zoonotic Diseases.' www.cdc.gov/onehealth/basics/zoonotic-diseases.html
[3] Blocker, K. (2020). 'Older Adults Advised to 'Stay Home as Much as Possible' During Coronavirus Outbreak', *UCHealth Today*. www.uchealth.org/today/older-adults-coronavirus-can-be-more-serious/
[4] See Nicogossian et. al. (2014).
[5] Kieny, M-P., D. B. Evans, G. Schmets, and S. Kadandale (2014). 'Health-system Resilience: Reflections on the Ebola Crisis In Western Africa', *Bull World Health Organ* 92, 850.
[6] See Centers for Disease Control and Prevention (2017).

consider how a country is prepared for such events. After studying health systems' responses to the "2009 swine-origin H1N1 influenza A epidemic," Scotch (2012) noted that "integration of human and animal disease surveillance has been recommended" so countries can "better predict and prepare for future epidemics."[7]

H3: Health infrastructure variables have a positive relationship with confirmed coronavirus cases.

*Economic Indicators:* It is believed that "regions, countries and groups that are already in disadvantage economically, politically, and socially are often the most vulnerable" in disease outbreaks.[8] Therefore, including economic indicators in this model was necessary to account for this belief. Contrary to his original hypothesis, Zanakis (2007) found that GDP output actually had a positive relationship with confirmed HIV/AIDS cases. This suggests that economic development and openness may have a positive relationship with confirmed coronavirus cases as well. In addition to this, a country with a higher number of annual international tourism arrivals will have a greater chance of a disease spreading across borders.

H4: Economic development & openness indicators have a positive relationship with confirmed cases.

## Methodology

*Dependent variable*

The dependent variable used was confirmed coronavirus case, data which was collected from Johns Hopkins University's Coronavirus Resource Center.[9] A country had to have at least one confirmed case of coronavirus in their territory as of April 20, 2020 at 12:38 PM to be included in this model. In total, the model's number of observations was 178 countries. Once the final model was run, the sample size fell to 139 countries because of missing explanatory values in the dataset. It's important to note that the number of cases have since increased because the pandemic is still occurring, which might impact overall data quality and success of the models.

After running an exploratory regression, it became clear that the final model had to include the log-transform of confirmed coronavirus cases as the dependent variable to compress the scale and the presence of outliers in the data. Without this function form transformation, the regression was rejecting the Jarque-Bera Normality Test, White's Heteroscedasticy Test, and the RESET23 Test's null hypothesis,

[7] Scotch, M., J. S. Brownstein, S. Vegso, D. Galusha, and P. Rabinowitz (2012). 'Human vs. Animal Outbreaks of the 2009 swine-origin H1N1 influenza A Epidemi', *Ecohealth* 8, 376-380.
[8] Zanakis, S. H., C. Alvarez, and V. Li (2007). 'Socio-economic Determinants of HIV/AIDS Pandemic and Nations Efficiencies', *European Journal of Operational Research* 176, 1811-1838.
[9] Johns Hopkins University (2020). 'COVID-19 Dashboard by the Center for Systems Science and Engineering', *Johns Hopkins Coronavirus Resource Center*.

which violated a number of Gaussian assumptions, at a 5% level of significance. As the normality test result showed the residuals were not normally distributed, the other two tests showed me that heteroscedasticity and misspecification was present. This was largely because the incorrect functional form of confirmed coronavirus cases was being taken. As soon as the dependent variable was changed to the log of confirmed coronavirus cases, the model was passing all tests it previously failed.

*Explanatory variables*

Based on the hypotheses laid out in the previous section, eleven explanatory variables were included in this model. Data for all but one of these variables was collected from the World Bank's World Development Indicators database. Since the most recent data has not been recorded yet, the fullest set of data currently available was from 2017 for these variables. Geographic indicators in this model included UrbanPop%TotalPop, which measures what percentage of a country's population lived in urban areas as of 2017, and land area (sq. km). Demographic indicators included Ages65Plus%ofTotalPop, Female%ofTotalPop, and each country's total population per million as of 2017. Health infrastructure indicators include the 2017 Current Health Expenditure (% of GDP), which estimates "healthcare goods and services consumed during" 2017, and UHC services coverage index score, which measures access to health services and service capacity.[10]The 2018 Zoonotic Events and Human-animal Interface score collected from the World Health Organization is also included as a health infrastructure indicator because it measures whether "mechanisms for detecting and responding to zoonoses and potential zoonoses are established and functioning" in a state.[11] Economic development and openness indicators included the 2017 GDP per capita (annual %), the Ease of doing business score, and International Tourism Arrivals variables collected from the World Bank.

## Results

After realizing I needed to use the log-form of confirmed coronavirus cases as the dependent variable, the first multivariate cross-sectional model I ran is recorded in Table 1 as a log-linear model. Based on its output, model 1's statistical strength was mediocre. It passes the Jarque-Bera Normality Test and the White's Heteroscedasticity Tests, showing that the residuals are normally distributed and homoscedasticity is present. The R^2 values showed that the model accounts for 67.629% of the total

[10] World Bank (2020). 'Data Catalog.' datacatalog.worldbank.org/search?search_api_views_fulltext_op=AND&query=UHC+Service+Coverage&nid=&sort_by=search_api_relevance&sort_by=search_api_relevance
[11] World Health Organization (2020). 'The Global Health Observatory: Zoonotic Events and the Human-animal Interface.' www.who.int/data/gho/data/indicators/indicator-details/GHO/zoonotic-events-and-the-human-animal-interface

variation in log-confirmed coronavirus cases is explained by the eleven explanatory variables. This is necessarily weak, but it could definitely be stronger. The F-test showed that the variables have joint significance since the F-test's p-value=0.000, implying that the model accepts the alternative hypothesis that joint significance is present. In addition to this, six of the explanatory variables were statistically significant at the 5% level of significance. However, this model rejects the null hypothesis of the RESET23 test at a 5% significance level because the RESET23 test's p-value=0.002. This meant that misspecification was still present, which violates the ninth Gaussian assumption. This can occur for a host of ideas from including irrelevant variables to having data in the incorrect function form or measurement errors. The PcGives output also suggested I rescale the data in a warning that popped up with the output, as seen below.

Table 1: Model 1

```
*** Warning: diagonal elements of the second moment matrix are very small or very different.
Numerical accuracy is endangered, try rescaling the data.

EQ( 1) Modelling Log(Confirmed Cases)) by OLS (static model)
       The dataset is: C:\Users\97kbe\Desktop\Bell_INTAF 803 Final Paper Data.xlsx
       The estimation sample is: 1...178
       Dropped 20 observation(s) with missing values from the sample

                                                        Coefficient  Std.Error  t-value  t-prob Part.R^2
Constant                                                    3.78655      2.538     1.49  0.1379   0.0150
Urban%ofTotalPop                                            3.15270     0.8878     3.55  0.0005   0.0795
Land area per 1000 sq. km.                             0.000130414  7.551e-05     1.73  0.0863   0.0200
Ages65Plus%ofTotalPop                                       7.78877      3.818     2.04  0.0432   0.0277
Female%ofTotalPop                                          -7.02749      4.502    -1.56  0.1207   0.0164
Total Population                                       2.52557e-09  9.962e-10     2.54  0.0123   0.0422
Current health expenditure (% of GDP)                    0.0951946    0.05669     1.68  0.0953   0.0189
UHC service coverage index                             -0.00779039    0.01546   -0.504  0.6150   0.0017
2018 Zoonotic Events Interface                           0.0122426   0.005554     2.20  0.0291   0.0322
GDP per capita growth (annual %)                         0.0406475    0.03997     1.02  0.3109   0.0070
Ease of doing business score                             0.0347158    0.01290     2.69  0.0080   0.0473
International tourism, number of arrivals              4.95752e-08  1.143e-08     4.34  0.0000   0.1141

sigma                1.61037  RSS                378.622537
R^2                  0.67629  F(11,146) =   27.73 [0.000]**
Adj.R^2             0.651901  log-likelihood       -293.234
no. of observations       158  no. of parameters          12
mean(Y)              6.47777  se(Y)                2.72945

Normality test:   Chi^2(2)  =  0.52025 [0.7710]
Hetero test:      F(22,135) =   1.4133 [0.1187]
Hetero-X test:    F(77,80)  =   1.3099 [0.1165]
RESET23 test:     F(2,144)  =   9.0459 [0.0002]**
```

To fix these problems, I ran a second log-linear model with the log-form of confirmed coronavirus cases. In model 2, I changed the functional form of the land area, international tourism arrivals, and total population variables by putting them in a log-transform because I believed they were highly skewed variables that could benefit the most from a logarithmic transformation because countries like the United States and China, which have extremely big populations, land area, and tourism numbers, were included in my data set with much smaller countries in my data set like Antigua and Barbuda. The other explanatory variables were left in their original linear form.

Table 2: Model 2

```
EQ( 2) Modelling Log(Confirmed Cases) by OLS (static model)
       The dataset is: C:\Users\97kbe\Desktop\Bell_INTAF 803 Final Paper Data.xlsx
       The estimation sample is: 3...178
       Dropped 39 observation(s) with missing values from the sample

                                                          Coefficient  Std.Error  t-value  t-prob Part.R^2
Constant                                                     -7.69462      2.142    -3.59  0.0005   0.0922
Urban%ofTotalPop                                              2.74216     0.7077     3.87  0.0002   0.1057
Log(Land area (sq. km))                                     -0.101693    0.07599    -1.34  0.1832   0.0139
Ages65Plus%ofTotalPop                                         15.2183      2.888     5.27  0.0000   0.1794
Female%ofTotalPop                                            -14.1258      3.472    -4.07  0.0001   0.1153
Log(Total Population)                                         1.04778     0.1208     8.67  0.0000   0.3719
Current health expenditure (% of GDP)                        0.0503731    0.04667     1.08  0.2825   0.0091
UHC service coverage index                                   0.0320709    0.01406     2.28  0.0242   0.0394
2018 Zoonotic Events Interface                             -0.000637101   0.004495   -0.142  0.8875   0.0002
GDP per capita growth (annual %)                             0.0203571    0.04145    0.491  0.6242   0.0019
Ease of doing business score                                 0.0333182    0.01046     3.19  0.0018   0.0740
Log(International tourism, number of arrivals)               -0.123354     0.1221    -1.01  0.3143   0.0080

sigma                 1.14333  RSS                166.016221
R^2                  0.838193  F(11,127) =   59.81 [0.000]**
Adj.R^2              0.824178  log-likelihood       -209.576
no. of observations       139  no. of parameters          12
mean(Y)               6.69876  se(Y)                2.7267

Normality test:   Chi^2(2)  =  0.21818 [0.8966]
Hetero test:      F(22,116) =   1.3759 [0.1411]
Hetero-X test:    F(77,61)  =   1.2989 [0.1448]
RESET23 test:     F(2,125)  =  0.40513 [0.6678]
```

Based on the PcGive output, this function form transformation of three explanatory variables solved the misspecification problem. Model 2, as shown in Table 2, accounts for 83.8193% of the total variation in log-confirmed coronavirus cases is explained by the eleven explanatory variables. Zanakis (2007) argues that "reasonably good fit regression models" have an adjusted R^2 between 70-90%.[12] With an adjusted R^2=82.4178%, it realistic to consider Model 2 as a "reasonably good fit regression model. This model also passes all of the normality, heteroscedasticity, and misspecification tests, while also showing that there is joint significance between the variables as F(11,127)=59.81[0.000]. These elements show that Model 2's statistical power is strong, allowing me to trust the coefficients and t-prob values of the explanatory variables.

Despite this information, it's also important to note that high multicollinearity might be present between the dependent variable and at least one of the explanatory variables. The problem is inherent to real world data because of how data is collected and how humans behave. An appropriate remedy is to do nothing to your model, instead expand your level of significance to account for this. As such, I'm using a 5% level of significance as opposed to a 1% level to account for this problem and for the possibility of making a Type II error. Also, endogeneity, which "occurs when a predictor variable in a regression model is correlated with the error term", might be present because it's possible that important variables were

[12] See Zanakis et. al. (2007).

omitted from the final model in this paper since the R^2 isn't 100%.[13] Solving this problem is one way this model can improve in the future. Also, since the data used in this paper was not time-series data, I wasn't able to use the Breusch-Godfrey test to identify whether or not autocorrelation was present in the model.

## Conclusion

Using Model 2's output, there is evidence that supports this paper's four hypotheses. UrbanPop%TotalPop's coefficient, which was the only statistically significant geographic indicator at the 5% level of significance, shows there's a positive relationship between it and Log(ConfirmedCases), supporting Hypothesis 1. In other words, when UrbanPop%TotalPop increases, so does the percentage of confirmed cases in a country. The regression output also shows that health Infrastructure indicators, as well as economic development and openness indicators, support their associated hypotheses. Regarding Hypothesis 3, the only statistically significant health infrastructure indicator at a 5% significance level was UHC service coverage index. With a positive, coefficient, UHC service coverage index has a positive relationship with Log(ConfirmedCases). Regarding Hypothesis 4, the only statistically significant economic development and openness indicator is the Ease of Doing Business score. Therefore, countries with a high performance on this index, meaning there's great ease domestically in doing business, will experience an increase in the percentage of confirmed cases. This supports Hypothesis 4, as economic development and openness has a positive relationship with confirmed coronavirus cases.

Demographic indicators prove to have a more complicated relationship. When it came to testing Hypothesis 2, all the demographic indicators were statistically significant in model 2. And as hypothesized, Age65Plus%TotalPop and Log(Total Population) are positively related to confirmed coronavirus cases. Female%ofTotalPop, however, has a negative relationship with Log(ConfirmedCases). This shows that countries with a higher population size and percent of the population that's aged 65+ are expected to have a larger number of confirmed cases.

In conclusion, the statistical analysis of Model 2 shows what key determinants impact the number of confirmed cases in the ongoing COVID-19 pandemic. The output leaves me to infer that countries with a large urban population, bigger and older total population, high UHC services index, and high ease of doing business score are expected to have more confirmed cases of COVID-19 based on their statistical significance and coefficient values. This is reasonable to comprehend because it suggests countries that

[13] Lynch, S. M. and J. S. Brown (2011). 'Stratification and Inequality Over the Life Course', *Handbook of Aging and the Social Sciences* 7.

are more open to economic opportunities, as well as having large and older populations, should expect to see more cases within their countries. In addition, when they have better health systems, people have more access to health services and a bigger opportunity of being tested for the coronavirus. Inversely, countries with a larger percent of females in their total population will see less confirmed cases. The evidence presented here could be improved in the future by using World Bank data from 2019 once its recorded, as well as the final total number of confirmed cases once the pandemic ends.

**Word Count: 2495**