Hudson Finch-Batista
Biostatistics

Final Project - Is the Doubling Time of Covid-19 Cases Related to Democracy or Healthcare?

Introduction

My primary research topic is the difference in the doubling time of Covid-19 cases across various countries and whether it is associated with different types of economies or governments. This topic came about as a consequence of examining the difference in the rate of growth of confirmed cases of Covid-19 between different countries, some of which appeared to reach a decreasing rate of growth well before others. For example, in just two months, the rate of growth of confirmed cases of Covid-19 in Mainland China began to not only decrease but nearly level off while in the US and many western European countries the rate of growth of confirmed cases was still increasing or marginally decreasing. Could this be due to the fact that China is an authoritarian government with a greater ability to control society and impose stringent measures that are critical to impeding the transmission of the virus? Is the high rate of growth of cases in the US due to the predominance of the private healthcare system and the high cost of healthcare? These are the basic questions that guided my primary observational study. Along with these questions, a secondary question that guided my secondary study was whether or not age and sex are, in fact, associated with the probability of death.

There are four datasets used in this study. The first dataset used in this study, sourced from Kaggle, contains data on individual cases in South Korea. The data contained whether someone had been released from a hospital (i.e. recovered), isolated, or deceased on an individual basis. This was crucial to determine whether the odds of survival cold be modeled as a function of age and sex since cumulative data only aggregates deaths and recoveries. Cases that were categorized as isolated were not included in the study as they represented an intermediate status of a case that has not yet been determined as a death or a recovery. In other words, the study only looks at whether a case of Covid-19 led to a death or a recovery. The second dataset used in the study, from the World Health Organization (WHO), contains data describing the healthcare systems of different countries over many years. The variable of interest in this dataset is the current predominant healthcare financing scheme of a given country. Since almost all countries have a mix of private, public, and out-of-pocket health care financing schemes, the primary healthcare financing scheme was determined by calculating the health care financing scheme that accounted for the highest dollar value in a given country in the year 2017 (since the latest year in the dataset was the year 2017, it was used as the reference year). The second dataset used in this study, from the Economist Intelligence Unit (EIU), contains data describing the democracy index of different countries, which is an average score based on five variables that countries are scored by. The variables are electoral process and pluralism, functioning of government, political participation, political culture, and civil liberties. Each country, moreover, is categorized based on its score as either a full democracy, flawed democracy, hybrid regime, or authoritarian regime. The variable of interest in this case is the regime type of a given country. Unlike the prior dataset, the reference year for this dataset was 2019.

The third dataset contains data describing the confirmed cases, deaths, and recoveries due to Covid-19 on a country by country basis and daily basis from January 22nd to April 7th of 2020. Since the initial onset of Covid-19 cases differed across the countries in this dataset, and since

most countries had at least a single case by February 1st, the cases among the countries in the sample were analyzed from February 1st to April 7th. The primary variable of interest here was the doubling time of confirmed cases of Covid-19 across different countries, which was calculated by taking the base 2 logarithm of confirmed cases and modeling it as a linear function of time.

I tried to approximate a stratified random sampling method by choosing countries that were representative of the seven continents of the world, the different types of national economies across the world (welfare state vs. unfettered market economy), and the different types of national governments (authoritarian vs. democratic) among the world, however, there was inevitably some convenience bias introduced. For example, the democratic index measures democracy by a scale and I wanted to choose countries that were either democratic or authoritarian, rather than hybrid regimes. As another example, countries with primarily household out-of-pocket payment healthcare financing schemes were classified as private healthcare financing schemes mostly out of convenience. However, in order to correct for these binary re-categorizations and to uphold the integrity of the government type and healthcare type data, the Covid-19 case data was modeled according to both the original and derived (i.e. binary) categorizations of healthcare types and government types. And with the Korean data on individual cases, there was undoubtedly convenience bias since even provisionally reliable data on individual Covid-19 cases is extremely difficult to verify. Additionally, large portions of the dataset lacked birthdate data, so they could not be part of the sample, which inevitably influenced the results of the logistic regression model.

Primary Study

The Covid-19 case doubling time was determined as the reciprocal of the slope coefficient of the linear model of $\log_2$ Covid-19 cases as a function of time. The logic, using the following output for Iran, is as follows:

	Estimate	Std. Error
Intercept	-4509	311.9
Observation Date Slope Coefficient	0.2466	0.01701
Linear Model	$\log_2 y = -4509 + 0.2466t$	

If we exponentiate both sides of the linear function, we can convert the function to exponential form. Then, if we set y = 2 and take the base 2 logarithm of both sides, we can solve for t and find the time it takes for the initial amount to double as follows:

$\log_2 y = -4509 + 0.2466t \rightarrow y = 2^{-4509} * 2^{0.2466t} \rightarrow y = 1 * 2^{0.2446t} \rightarrow 2 = 2^{0.2466t} \rightarrow \log_2(2) = 0.2446t \rightarrow 1 = 0.2446t \rightarrow t = 1/0.2466 = 4.055$

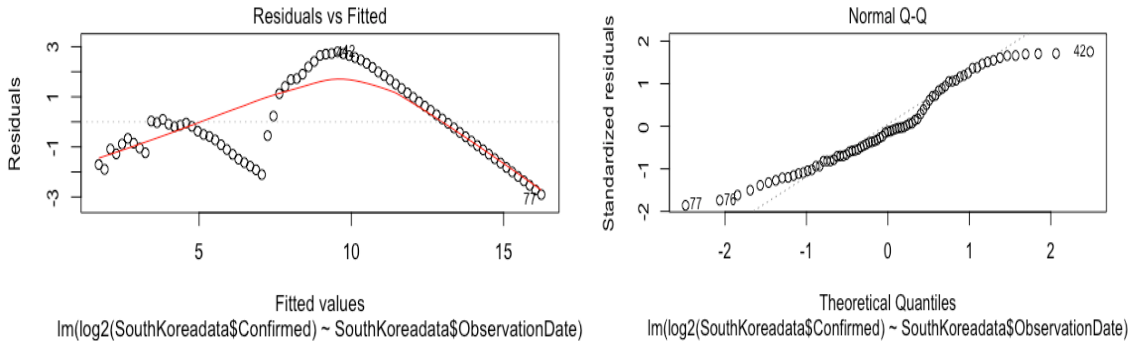
Thus, in the case of Iran, the doubling time of Covid-19 cases is approximately 4 days. Below is a table that lists the doubling times of each country in the sample as well as the corresponding predominant healthcare type and regime type.

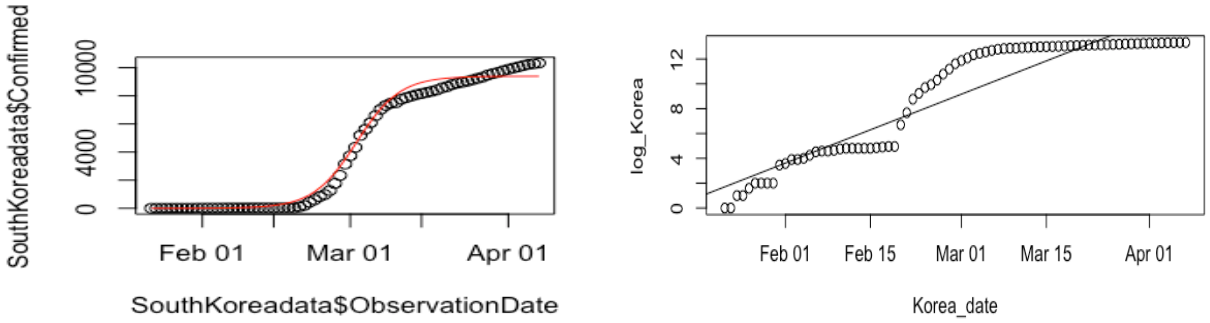
Country	Covid-19 Doubling Time	Predominant Type of Healthcare	Government Type
US	4.011	Private	Democracy
Spain	3.036	Public	Democracy
Italy	3.248	Public	Democracy
China	17.343	Public	Authoritarian
Iran	4.055	Private	Authoritarian
Venezuela	5.783	Private	Authoritarian
Ethiopia	4.970	Public	Authoritarian
France	4.199	Public	Democracy
Vietnam	11.148	Private	Authoritarian
Russia	5.09	Private	Authoritarian
South Korea	5.23	Public	South Korea

As can be observed, the doubling time of Covid-19 cases in China, along with Vietnam, stands out from the rest. That is, it takes much longer for the number of cases to double in China and Vietnam compared to the rest of the countries in the sample. The question is, is this difference due to random sampling error or can it be attributed to the type of government and/or the type of healthcare in Vietnam and China?

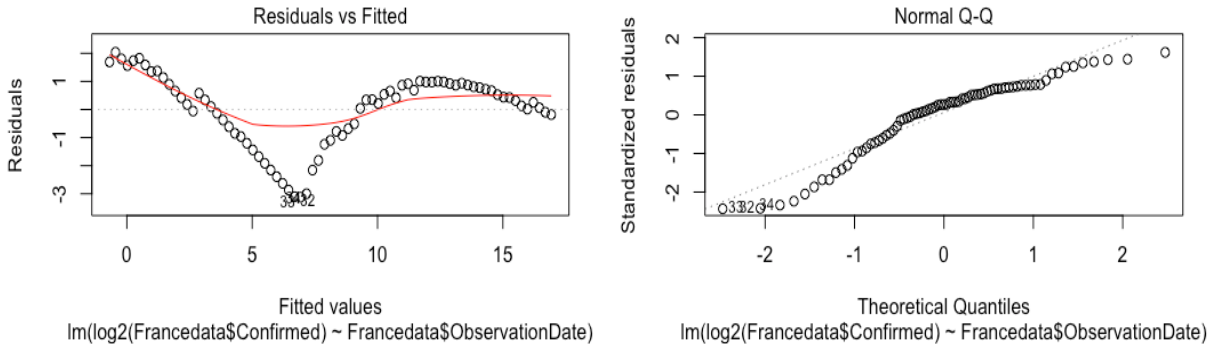
If we look at the log base 2 transformed data for South Korea and France, we see that a linear model is not a very good fit for the data in either case. The quantile-quantile plot of the residuals makes an S shape, with long tails on either side, which implies non-normality. Moreover, the plot of the fitted values against the residuals does not display constant variation but rather displays non-linear variation, which would mean that the data would be better modeled using a non-linear model of some sort. This was roughly the same for every country in the sample despite the difference in their Covid-19 case growth rates, such as France and South Korea.

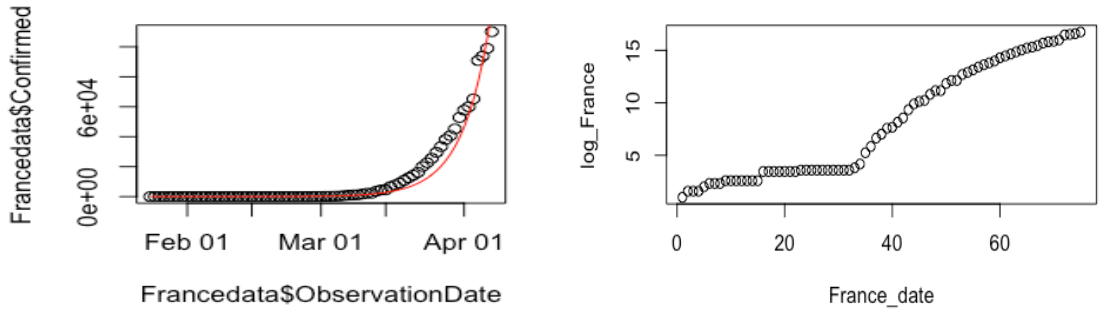
Log Base 2 Transformed Data for South Korea





Log Base 2 Transformed Data for France



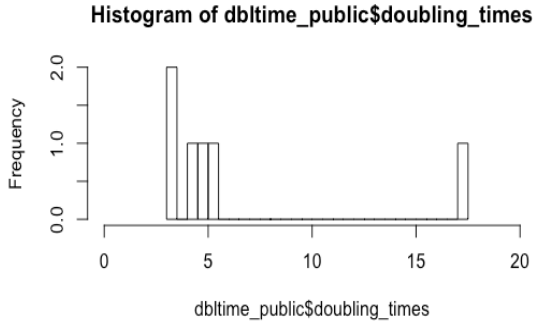


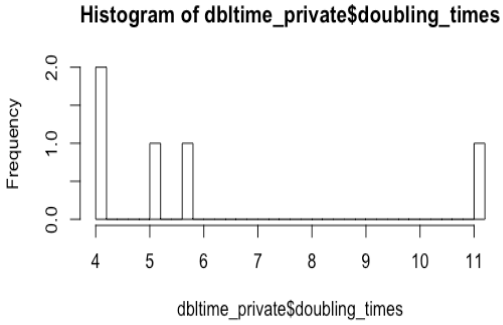
The R^2 value of 0.88 and a correlation coefficient value of 0.94, along with the p-value of 2.2 E-16, for the South Korean data on the other hand, more than satisfy the conditions of a good linear model. The same is true of France, which had an R^2 value of 0.94 and a correlation coefficient value of 0.97. Despite the seemingly good fit implied by the summary statistics of the linear models, however, based on the residual plots it's clear that a linear model is not applicable. Since the doubling time is the reciprocal of the growth constant in an exponential model, however, it could be argued that goodness of the fit is irrelevant because the purpose of linearizing the data is essentially to find the growth constant. The scatterplot of data for France, for example, demonstrates exponential growth and, as can be observed, if an exponential line of best fit is laid over the plot using the slope coefficient of the linear model as a growth constant, it

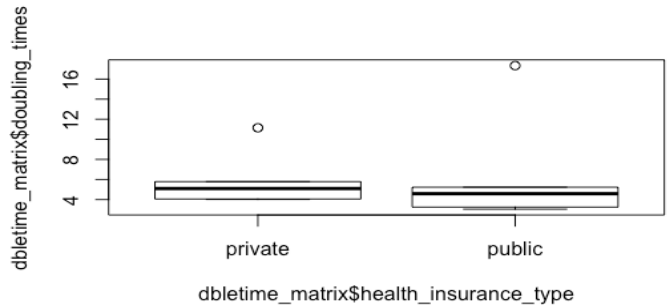
seems to predict the data quite well. That is, despite the fact that a linear model is not a good fit for log base 2 transformed confirmed case values, the doubling time should be viewed as valid since the raw data can, for the most part, be modeled by exponential growth. The scatterplot of the untransformed confirmed cases of Covid-19 in South Korea as a function of time, on the other hand, demonstrates logistic growth, so the doubling time is perhaps not as valid. This is partially reflected by the lower R^2 and R values for the linear model of South Korean cases compared to the linear model of French cases. The only other country that also demonstrated logistic growth was China, so the doubling time is for the most part fairly reliable across the sample of countries.

Doubling Time by Type of Healthcare

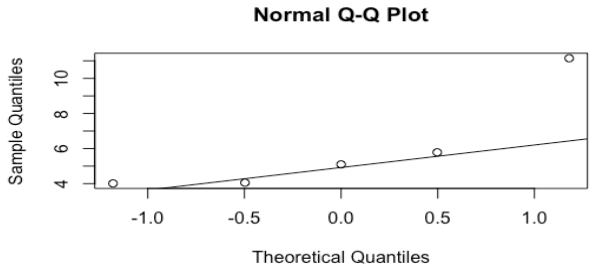
As can be inferred from the histograms below, the distribution of doubling times for countries with either type of healthcare system is highly right skewed or even bimodal, which is due to the doubling times for China and Vietnam respectively. The quantile-quantile plots below also demonstrate that the data are highly non-normal due to the two aforementioned outliers. The Shapiro-Wilks test yielded a p-value of 0.0357 for the sample of countries with predominantly private healthcare systems and a p-value of 0.001 for the sample of countries with predominantly public healthcare systems, which indicates that both samples likely come from non-normal populations. Therefore, only a randomization test or Wilcox-Mann-Whitney test would make sense in this context. Ignoring the high non-normality of the data, one can see from the side by side boxplot below that countries with public health care systems tended to have lower doubling times compared to countries with private health care systems. The median doubling time for countries with private health care was 5.097 while the median doubling time for countries with public health care was 4.585.



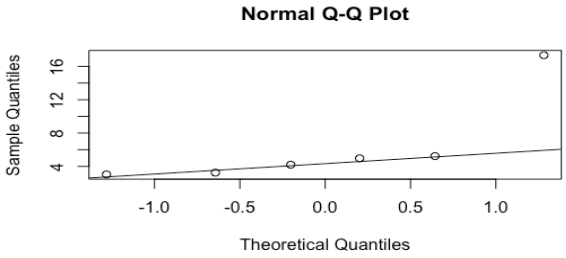




Private Healthcare Doubling Time



Public Healthcare Doubling Time QQ Plot



Summary Statistics for Countries with Public Healthcare

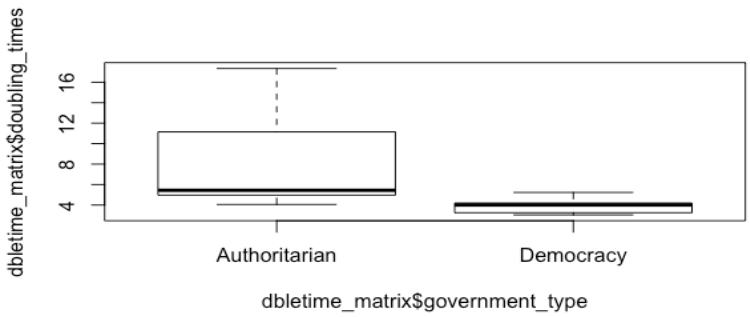
Mean Doubling Time	6.338
Standard Deviation of Doubling Time	5.463
Median Doubling Time	4.585
IQR	1.678
n	6
Shapiro-Wilks Test (P-value)	0.001

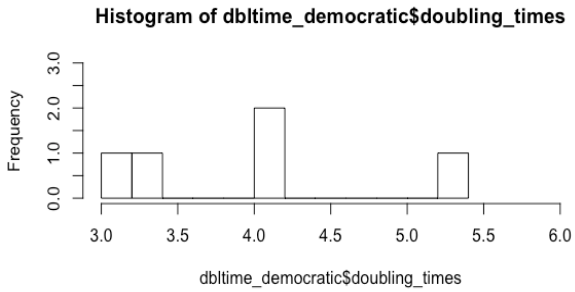
Summary Statistics for Countries with Private Healthcare

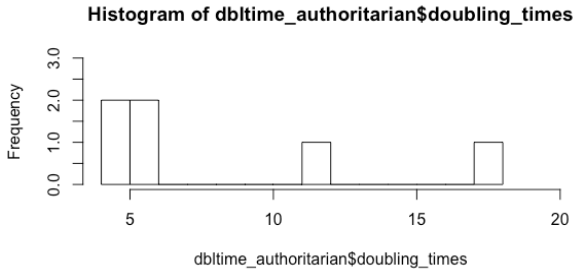
Mean Doubling Time	6.019
Standard Deviation of Doubling Time	2.962
Median Doubling Time	5.097
IQR	1.729
n	5
Shapiro-Wilks Test (P-value)	0.035

While the mean and IQR for both types of healthcare were similar, the standard deviation for countries with public healthcare was much larger at 5.463 compared to the standard deviation of 2.692 for countries with private healthcare. While there is a fair degree of spread in both samples, the sample of countries with predominantly public healthcare had a greater amount of spread, which indicates that the doubling rate across countries in this sample is highly variable. Removing China and Vietnam would no doubt lessen this high degree of variability, however, I felt it was important to include them to uphold the integrity of the study.

Doubling Time by Type of Government

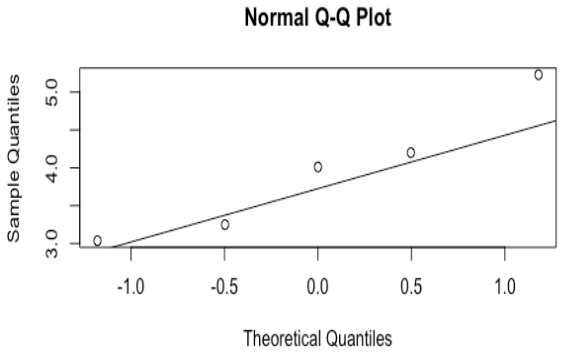


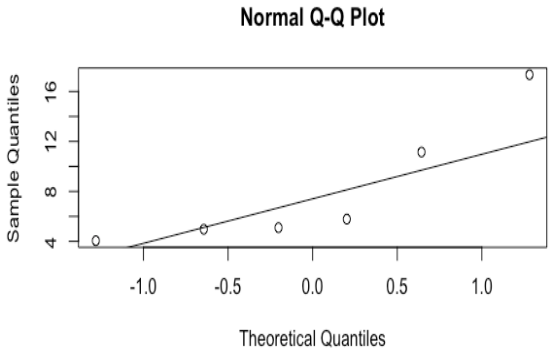




Democratic Country Doubling Time

Authoritarian Country Doubling Time





The situation is similar for a comparison of doubling times by regime type or government type, however, on this occasion it appears that one of the samples is normal while the other is non-normal. The histogram on the left, which displays the doubling times for democratic countries, appears fairly normal as the majority of observations occur in the center. If anything, there is some slight right skew but that seems to be the extent of the non-normality. The histogram on the right, which displays the doubling time data for authoritarian countries, by contrast, looks very right skewed and this can also be inferred in the quantile-quantile plot, which looks very non-linear whereas the quantile-quantile plot for democratic countries appears far more linear. Not surprisingly, the Shapiro-Wilks test yielded a p-value of 0.66 for the sample of democratic countries and a p-value of 0.047 for the sample of authoritarian countries. This of course implies that only a randomization test and/or Wilcox-Mann-Whitney test would be appropriate in this case. Once again, ignoring this non-normality, one can observe from the side by side boxplot that the two distributions do appear to truly differ. In this case, it appears that countries run by authoritarian regimes have a much longer Covid-19 case doubling time than democratic regimes. The mean doubling time for countries with authoritarian regimes was 8.066 compared to 3.945 for countries with democratic regimes. The two outliers of Vietnam and China are, again, likely to blame for this large difference in mean doubling times. However, if we look at the median for both regime types, authoritarian regimes still have a higher median than democratic regimes by almost one and a half days. The question now is whether or not this is a significant difference.

Summary Statistics for Countries with Authoritarian Regimes

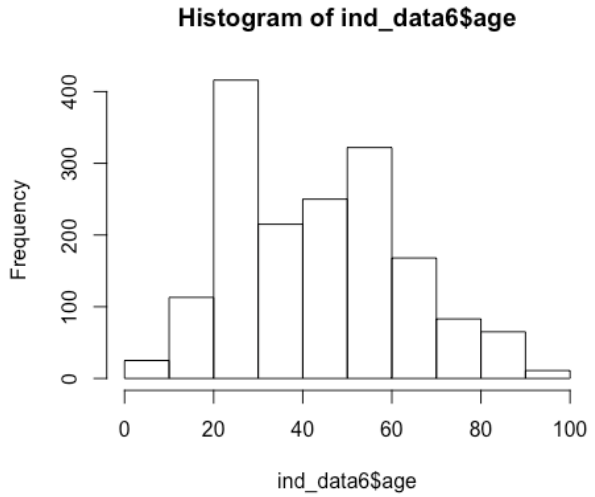
Mean Doubling Time	8.066
Standard Deviation of Doubling Time	5.201
Median Doubling Time	5.440
IQR	4.805
n	6
Shapiro-Wilks (P-value)	0.048

Summary Statistics for Countries with Democratic Regimes

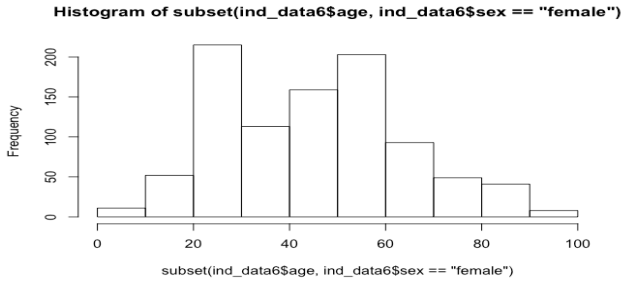
Mean Doubling Time	3.945
Standard Deviation of Doubling Time	0.870
Median Doubling Time	4.011
IQR	0.951
n	5
Shapiro-Wilks (P-value)	0.663

Secondary Study

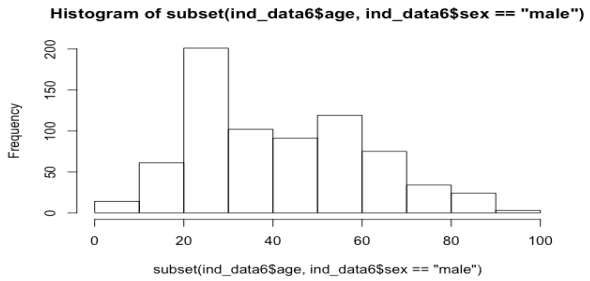
The histogram below shows that age is approximately normally distributed among Korean individuals that contracted Covid-19 in this dataset. Moreover, the adjacent histograms and side by side boxplots of age by sex show that the distribution of the age of males and females who have contracted Covid-19 are very similar. Males appear to be marginally younger on average. However, if we look at the adjacent histograms and the side by side boxplots of age by status (dead or recovered), it can be observed that the distributions are very different. That is, it appears that older individuals are more likely to die from Covid-19.



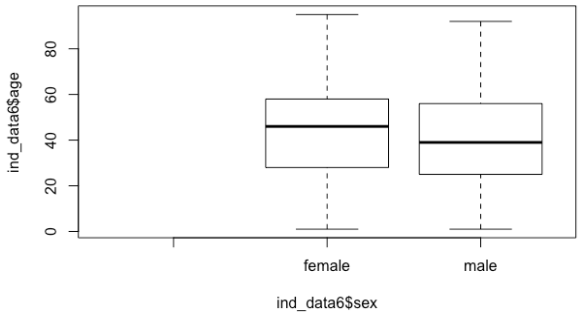
Histogram of Age by Female Sex



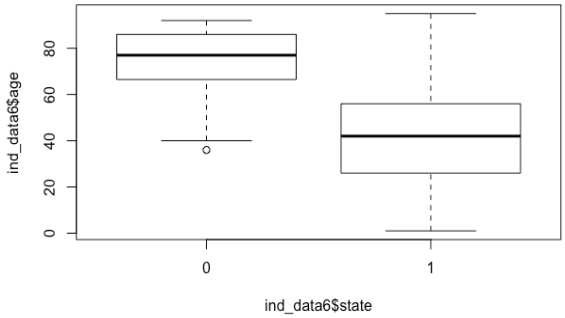
Histogram of Age by Male Sex



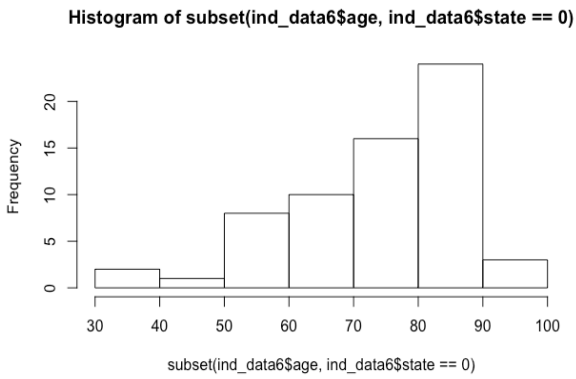
Side by Side Boxplot of Age by Sex



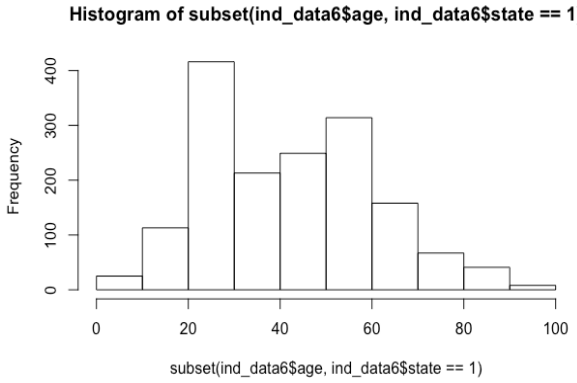
Side by Side Boxplot of Age by Status



Histogram of Age by Death



Histogram of Age by Survival



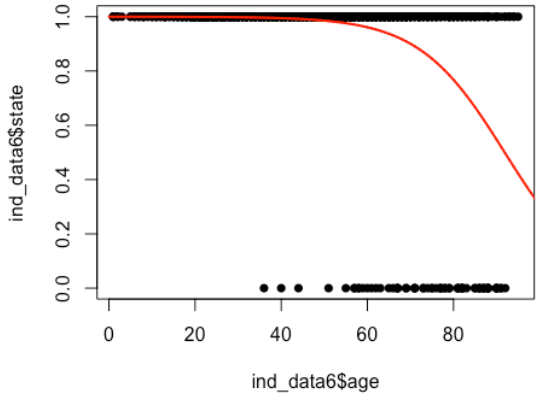
Summary of Age by Deceased

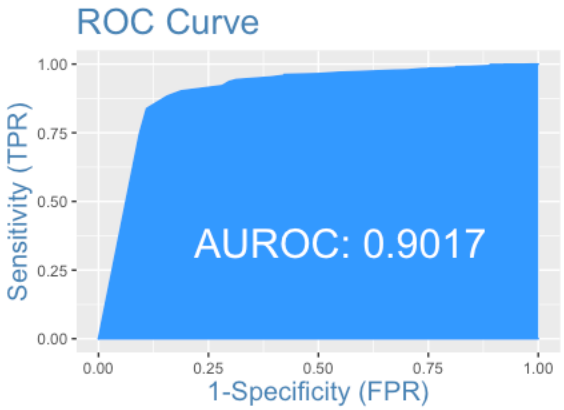
Mean Age	74.44
Age Standard Deviation	13.32
Median Age	77.00
IQR	19.25
n	65

Summary of Age by Recovered

Mean Age	42.65
Age Standard Deviation	18.51
Median Age	42.00
IQR	30
n	1604

Model 1

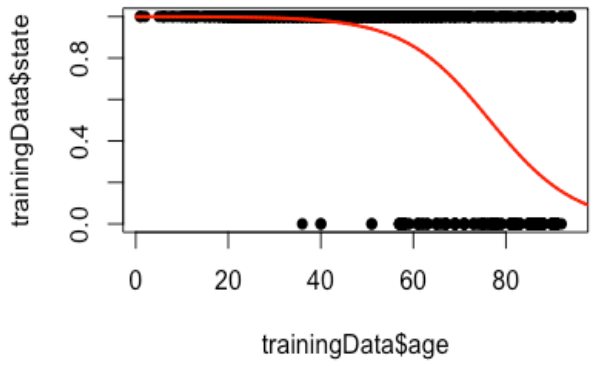


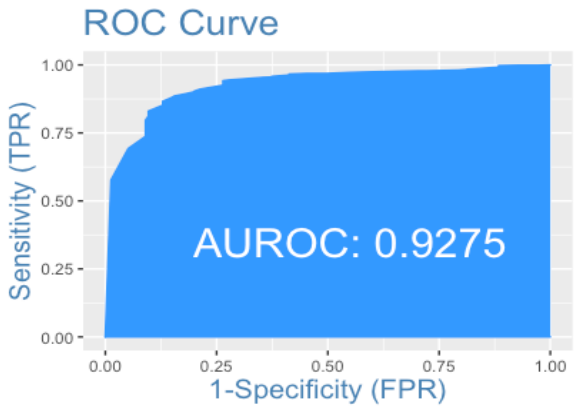


Confusion Matrix

	Actual 0	Actual 1
Predicted 0	7	8
Predicted 1	57	1596

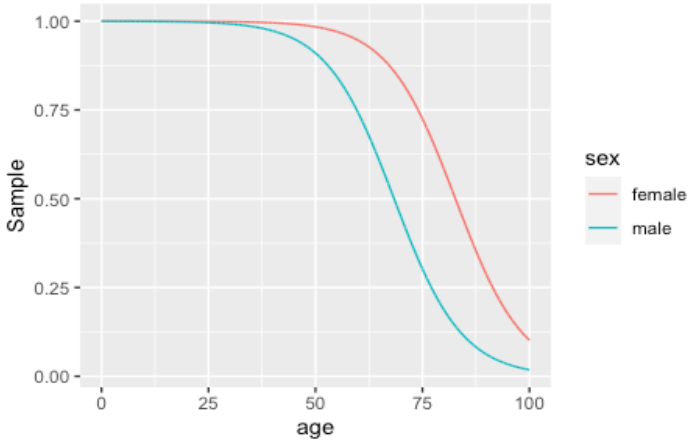
Model 2





Confusion Matrix

	Actual 0	Actual 1
Predicted 0	110	44
Predicted 1	45	786



Research Questions

The parameter of interest in the primary study is the difference in the mean doubling time of Covid-19 cases between democratic countries and authoritarian countries and the difference in mean doubling time of Covid-19 cases between countries with public healthcare systems and private healthcare systems. These two parameters can be expressed in hypotheses tests as follows.

Is the type of healthcare system of a nation associated with the doubling time of the confirmed cases of Covid-19? Specifically, are nations with public healthcare systems more likely to have lower doubling times of confirmed cases of Covid-19 than nations with private healthcare systems?

H_0: $\mu_{\text{doubling time for countries with private healthcare}} - \mu_{\text{doubling time for countries with public healthcare}} = 0$
H_a: $\mu_{\text{doubling time for countries with private healthcare}} - \mu_{\text{doubling time for countries with public healthcare}} \neq 0$

Is the type of government of a nation associated with the doubling time of the confirmed cases of Covid-19? Specifically, are authoritarian or non-democratic nations more likely to have lower doubling times of confirmed cases of Covid-19 than democratic nations?

H_0: $\mu_{\text{doubling time for countries with authoritarian governments}} - \mu_{\text{doubling time for countries with democratic governments}} = 0$
H_a: $\mu_{\text{doubling time for countries with authoritarian governments}} - \mu_{\text{doubling time for countries with democratic governments}} \neq 0$

The last question with respect to this study, which is not necessary to express in terms of hypotheses, is essentially tripartite. Is there any correlation between a nation's healthcare system and a nation's Covid-19 doubling time? If not, is there any correlation between a nation's regime type and a nation's Covid-19 doubling time? And, if not, is there any correlation between a nation's regime type and healthcare system and a nation's Covid-19 doubling time?

The parameter of interest in the second study is the probability of survival for Korean individuals that have contracted Covid-19. This parameter is also not necessary to express in terms of hypotheses. Rather, this parameter can be studied by asking the question of whether the odds of surviving Covid-19 are different for females than males and older people than younger people.

Test Results

Hypothesis Tests

	Difference in Doubling Time by Healthcare Type	Difference in Doubling Time by Government Type
t-Test	0.905	0.111
Randomization Test	0.454	0.068
Wilcox-Mann-Whitney Test	0.662	0.052

* Note: All values in this table are p-values

Multiple Linear Regression

	R^2	Adjusted R^2	P-value
Healthcare Type (Binary)	0.002	-0.109	0.910
Government Type (Binary)	0.251	0.168	0.117
Government Type & Healthcare Type (Binary)	0.346	0.182	0.183
Healthcare Type (Non-Binary)	0.246	-0.077	0.550
Government Type (Non-Binary)	0.110	-0.113	0.628
Government Type & Healthcare Type (Non-Binary)	0.344	-0.312	0.752

Logistic Regression

	Log Equation	Logit Equation	Death Prediction Accuracy	Survival Prediction Accuracy	ROC	AIC	Residual Deviance
Original Data	$\ln[p/(1-p)] = 10.734 - 0.110age - 1.563male$	$(p/1-p) = e^{10.73 - 0.11age - 1.56male}$	11%	99.5%	0.902	349.81	343.81
Oversampled Training Data	$\ln[p/(1-p)] = 10.418 - 0.126age$	$(p/1-p) = e^{10.42 - 0.13age - 1.81male}$	71%	94.6%	0.928	449.72	443.72

	– 1.806male						

Odds Ratios

	Odds Ratio (Age)	Odds Ratio (Male Sex)
Original Data	0.895	0.210
Oversampled Training Data	0.882	0.164

Even though the t-test was not valid due to the non-normality identified in each set of samples, I conducted a t-test to see whether the results would be significant or not in the unlikely but possible event that both population distributions were normal and the sample distributions were non-normal due to sampling error. In the case of the analysis of country Covid-19 case doubling time by type of healthcare, the t-test resulted in a p-value of 0.90, which was well above any reasonable level of required statistical significance. Not surprisingly, the confidence interval included 0 as well. I then conducted a Wilcox-Mann-Whitney test and came up with a high p-value of 0.66, indicating a high degree of overlap between the two population distributions. Finally, I conducted a randomization test with 10,000 trials of random samples of groups of 5 and 6 observations and found that the probability of finding a difference in sample means as large or larger than the difference between the original sample means due to sampling error or random chance is about 0.45. Thus, while the finding of the t-test was not valid, it was confirmed by both the Wilcox-Mann-Whitney test and the Randomization test.

However, in the case of the analysis of country Covid-19 case doubling time by type of government, the t-test resulted in a p-value of 0.1109 and a confidence interval between -9.57 and 1.33. While the confidence interval included 0 and while this p-value is still above 0.10, which is generally the lowest level of required significance for a p-value to be considered significant, this seemed to indicate that there may be a statistically significant difference between the two populations. Upon further testing, this proved to be the case, as the Wilcox-Mann-Whitney test yielded a p-value of 0.0519, which is significant at the 0.10 level of significance, indicating that the population distribution for the doubling time of cases of Covid-19 in authoritarian countries differs from that of democratic countries. Moreover, after conducting a randomization test with 10,000 trials of random samples of groups of 5 and 6 observations, I found that the probability of finding a difference in sample means as large or larger than the difference between the original sample means due to sampling error or random chance is about 0.0682. Like the Wilcox-Mann-Whitney test, this p-value is also significant at the 0.10 level.

I decided to test whether a multiple linear regression model would work with the original data categories in addition to the binary recategorizations I applied. Firstly, I tried finding a multiple linear model of the Covid-19 case doubling time of countries as a function of binary healthcare type (i.e. public or private). I came up with an adjusted R^2 value of -0.1094 and a p-value of 0.9099. The R^2 value can never be negative because it is simply the square of the correlation coefficient. However, the adjusted R^2 value can be negative when the R^2 value is very small and the ratio of observations to explanatory variables is high (i.e. few observations and many variables) because it is calculated as $1 – (1 – R^2)* [(n – 1)/(n – p – 1)]$ where n denotes the number of observations and p denotes the number of explanatory variables.[i] This was precisely the case here because the R^2 value was 0.001504, the number of explanatory variables was 2, and the number of observations here was 11. In effect, from the large p-value and very low R^2 value,

one can infer there is absolutely no correlation between healthcare type and doubling time of Covid-19 cases. That is, healthcare type does not at all predict the doubling time of Covid-19 cases.

I then did a multiple linear regression analysis of binary regime types (i.e. as either democratic or authoritarian) and came up with a much lower p-value of 0.11 and a better R^2 value of 0.251 and a better adjusted R^2 value of 0.167. However, this R^2 value was still very low, implying a correlation coefficient of about 0.50. Thus, government type is a very weak predictor of Covid-19 case doubling time. This confirms the results of the t-test, Wilcox-Mann-Whitney test, randomization test, and effect-size test of the Covid-19 case doubling time by government type, which showed significance at the 0.10 level but no importance. When I tried to model the doubling time of Covid-19 cases as a function of both dichotomous healthcare type and government type, I ended up with a higher R^2 value of 0.346 and a slightly higher adjusted R^2 value of 0.1823, but a higher p-value of 0.183, which is likely due to the fact that healthcare type is such a poor predictor of the doubling time of Covid-19 cases. Since the p-value is higher than the required level of significance of 0.10, we can essentially reject the linear model of Covid-19 doubling times as a function of healthcare type and government type. In sum, it appears that the doubling time of Covid-19 cases is best predicted by government type alone, although government type is still a poor predictor of Covid-19 case doubling time.

Upon converting the categories from their dichotomous forms back to their original forms, I came up with a negative adjusted R^2 for each linear model I tested. When modeling the doubling time of Covid-19 cases as a function of multiple healthcare types, I came up with a higher R^2 value of 0.2463 compared to the model based on the dichotomous healthcare type variable, which was expected purely due to the increase in explanatory variables. However, I also came up with an adjusted R^2 value of -0.078, and a p-value of 0.55, which indicates that when the R^2 value is adjusted for the effect of the increase in explanatory variables, it is basically negligible. Likewise, modeling the doubling time of Covid-19 cases as a function of multiple government types, I came up with an R^2 value of 0.1098, an adjusted R^2 value of -0.1128, and a p-value of 0.628, which is notable because the model of Covid-19 doubling time as a function of the dichotomous government type variable was statistically significant. And, finally, modeling the doubling time of Covid-19 cases as a function of both multiple government types and multiple healthcare types, I came up with an R^2 value of 0.344, a p-value of 0.752 and a multiple R^2 value of -0.312. Thus, in all three cases in which the doubling time of Covid-19 cases was modeled as a linear function of multiple categories, the adjusted R^2 value was negative and the p-value well above 0.10. This was probably due to the sample size being so small relative to the number of explanatory variables.

With respect to the secondary study, I performed a logistic regression analysis of status as a function of age and sex with status representing the probability of surviving Covid-19. The log equation is $\ln[p/(1-p)] = 10.73 - 0.11\text{age} - 1.562\text{male}$ and the logit equation is $(p/1-p) = e^{10.73 - 0.11\text{age} - 1.562\text{male}}$. Thus, the output for this model gave an intercept of 10.734, a coefficient of -0.11 for age, and a coefficient of -1.562 for males. The intercept of 10.734 represents the log odds of survival for a Korean individual with Covid-19 and an age of 0. The slope coefficient for age indicates that for a one unit increase in age, there will be a log -0.11 decrease in odds of survival. The slope coefficient for sex indicates that for a male individual, there will be a log -1.562 decrease in the odds of survival. In order to interpret the results in terms of probability, the coefficients (i.e. the logged odds units) must be exponentiated to find the odds ratios. Exponentiating the coefficients yields an intercept of 4.59, a coefficient of 0.895 for age and a

coefficient of 0.2095 for males. Therefore, the odds ratio for age is approximately 0.895, meaning the odds of survival are 0.895 times as great for every one-unit increase in age, which is to say that the odds of survival decrease with every unit increase in age. The 95% confidence interval for age, in turn, is (0.131 to -0.089). Likewise, the odds ratio for sex is 0.2095, meaning the odds of surviving are 0.2095 times as great for men as for women, which is to say the odds or survival decrease for males. Moreover, if we substitute 75 for age and male for sex, we get a probability of surviving of approximately 71.5%. If, on the other hand, we substitute 15 for age and male for sex, we get a probability of survival of approximately 100% ($p = 0.999$). If we substitute 75 for age and female for sex, we get a probability of survival of approximately 92.3%. And, finally, if we substitute 15 for age and female for sex, we get a probability of survival of approximately 100% ($p = 0.9999$).

I then looked at whether or not there was any significant class bias in the data and found that deaths accounted for only 4% of the total state data, so I oversampled from my original dataset and appended the results to my original dataset. That is, I took a random sample of my existing dataset with replacement with a 99% probability of choosing cases of deaths. I then added this random sample to my dataset, which increased the proportion of deaths to about 18%. Next, I created training and testing data for my dataset by splitting half of my dataset into training data and half into testing data. I then checked to see whether there was collinearity between the independent variables, age and sex, and came up with 1.0989, which is below 4 and, therefore, indicates no collinearity. The AIC for this model was about 449.72 with 982 degrees of freedom, which was significantly higher than the AIC for the previous model of 349.81 given 1667 degrees of freedom. The null deviance and residual deviance were also higher for the class balanced model, however, the confusion matrix for the latter model was far better at predicting deaths than the previous model. While the former model predicted survivals with a high degree of accuracy (99.5%), out of 64 actual deaths, only 7 were correctly predicted, which means that only 11% of the deaths were correctly predicted. However, with the new model, 110/155 deaths were correctly predicted, which equates to about 71% accuracy, and it still predicted survivals with a high degree of accuracy (94.6%). Moreover, as would be expected, the ROC plots demonstrate that the predictive accuracy of the latter model is superior in comparison with that of the former model, although both models seem to have fairly astute predictive capabilities.

Conclusions

The null hypothesis of part a cannot be rejected. That is, we have no evidence that there is a difference in the population mean doubling time of Covid-19 cases for countries with private healthcare systems and countries with public healthcare systems. That is, the doubling time of Covid-19 cases is unrelated to the type of healthcare system in a country. However, the null hypothesis of part b cannot be rejected. That is, we have moderate evidence that there is a difference in the population mean doubling time of Covid-19 cases for countries with authoritarian governments and countries with democratic governments at the 0.10 level of significance. However, the effect size of the result, 0.79, is less than 1 and, as a result, we cannot ascribe any importance to the difference in the population means. Thus, while there may be association between government type and the mean doubling time of Covid-19 cases, it is significant, but not important. Moreover, it was clear that the doubling rate of Covid-19 is not a linear function of either healthcare type or government type since the correlation coefficient and adjusted R^2 value were never meaningfully high and the p-value was typically well-above 0.10.

With regard to the secondary study, it is difficult to say which model is superior since they both had flaws, however, both models provided evidence that sex and age are very good predictors of the likelihood of survival from Covid-19. The intercept and two coefficients of both models yielded p-values that were statistically significant at a level of 0.01. Moreover, there was no collinearity between sex or age in the model in either model. Thus, we can interpret the logistic regression as providing significant evidence that the odds of survival among Koreans with Covid-19 are higher for younger people than older people and for females than for males.

Bibliography

1. SRK. (2020, May 13). *Novel Corona Virus 2019 Dataset.* Retrieved April 7th from https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset#covid_19_data.csv

2. Kim, Jihoo. (2020, May 1). *Data Science for COVID-19 in South Korea*. Retrieved April 20th from https://www.kaggle.com/kimjihoo/coronavirusdataset#PatientInfo.csv

3. Economist Intelligence Unit (EIU). (2020). *Democracy Index 2019: A year of democratic setbacks and popular protest.* Retrieved April 15th from https://www.eiu.com/topic/democracy-index

4. World Health Organization (WHO). (2020, May 13). *Global Health Expenditure Database.* Retrieved April 7th from https://apps.who.int/nha/database/Select/Indicators/en

Endnotes

[i] Nau, Robert. *What's a good value for R-squared?* Duke University Fuqua School of Business. June 2, 2019. https://people.duke.edu/~rnau/rsquared.htm