Stevens Institute of Technology

What affect the odds of Covid-19 mortality?

Team member:

Zhao, Xin

Wang, Ruixi

Course: BIA 652

Professor: Mai, Feng

Date: December 17, 2020

Table of Contents

1.Introduction

1.1 Research background

The worldwide Covid-19 pandemic has destroyed people’s livelihoods and their health. Some people lost their life in the disease and some are alive with sequelae. The economic and social disruption caused by the pandemic is devastating: without the means to earn an income during lockdowns, many people are unable to feed themselves and their families. The farmers especially those in low-income countries could not sell their foods as normal due to border closures, trade restrictions and confinement measures and new food security requirements. Therefore, many people are at risk of falling into extreme poverty.

It also has caused millions of enterprises face an existential threat. The global transportation systems are suffering from the Covid-19: Airline companies and import and export companies either bear significant loss and wait for transaction and traveling recover or are bankrupted and closed. The Covid-19 pandemic inevitably deteriorates international relation between countries.

Every country is eager for recovering Global economy and transaction market and decreasing the mortality rate.

1.2 Research motivation and methodology

This report uses logistic regression (LR) and various related-indicators as independent variable to investigate mortal possibilities of people who have taken the test of carrying Covid-19 virus during the epidemic period. The study initially identifies and examines nine indicators (represented by eighteen predicter variables) that can classify people whether is “alive” or “dead”. This study unavoidably biased the Covid-19 mortality rate because there are neglected or under-served segments of the population who are less likely to access healthcare or testing. Under-detection of cases may be exacerbated during an epidemic, when testing capacity may be limited and restricted to people with severe cases and priority risk groups (such as frontline healthcare workers, elderly people and people with comorbidities). Cases may also be misdiagnosed and attributed to other diseases with similar clinical presentation, such as influenza [1]. Another possible reason is that a substantial proportion of people with the infection are undetected either because they are asymptomatic or have only mild symptoms and thus typically fail to present at healthcare facilities.[2] But the study is still meaningful for practice.

One of the preliminary goals is to dig out which characteristics that is inherently unchanged such as sex, race and ethnicity and age plays the most essential part in causing or increasing the mortality of Covid19. The other goal is to pinpoint which characteristics that is inherently medical resource and can be adjusted such as hospitalization status and ICU status is able to decrease the mortality sharply. This paper asserts that the model development is able to not only enhance a hospital's resource arrangement ability, which may decrease the mortality in the epidemic, but also provide the government an insight to adjust the basic healthcare system to guarantee higher survival rates (that is, lower mortality) under a limited budget. Concurrent but not underlying comorbidity of Covid-19 disease, which also can influence the mortality, were not taken into account, however. This paper discusses the practical implications of using the LR method to predict the probability of Covid-19 mortality. We believe that the model can be used by hospitals, researchers, and health departments to enhance their ability to save life under the certain medical resources.

1.3 Dataset Description and Model Selection

The data is provided by CDC Case Surveillance Task Force [3]. The study sample consists of over five million country-level deidentified patient cases from beginning of the outbreak to Nov. 24, 2020 in US only. The dataset includes 11 data element public as columns and each row is a deidentified patient and is updated monthly. The columns cdc_report_dt which represents for initial case report date to CDC and pos_spec_dt which represents for date of first positive specimen collection are disregarded in our regression model. The reason is that at the start of the outbreak, detected cases are more likely to be severe or fatal. Patients with severe illness are more likely to present at health facilities and to be confirmed by laboratory test. Therefore, at beginning of pandemic the mortality rate is much higher. If we put the columns of initial time report date and date of first positive specimen collection into the regression model, the result is always biased because there is potential bias in detections of cases and deaths.

Then we use the remaining 8 columns as initial input of the regression model. After we remove missing and unknown values for sex, hospitalization status, ICU admission status, Death status, and Presence of underlying comorbidity for, the remaining 411 thousand of samples will be used for this research propose. Moreover, if symptom onset date is later than the CDC report date, this deidentified patient is considered asymptomatic carrier. So, our team created a new column called asymptomatic, yes if the patient is an asymptomatic carrier, no if symptomatic.

In addition to data cleansing and to create initial hypothesis of our model, our team conducts some initial statistics analysis of the data. The data consist of 395399 laboratory-confirmed case and 16463 probable case; the data collects the number of 216852 females and the number of 195010 males. There are 29543 cases use ICU, which is 7.17% of total cases, and 217607 cases presence of underlying comorbidity or disease, which is 52.83% of total cases. The majority ethnicity of the data is white, the second ethnicity group is Hispanic and Latino, then black, and last Asian.

Moreover, among 411862 rows of data, the number of 80810 cases received hospitalization service, which roughly equal to 19.62% of total cases. By given the number of death cases, there are the number of 26381 death cases received hospitalization service, and 3798 death cases never receive hospitalization. Intuitively, it seems that hospitalization would increase the odds of death. However, the correlation between hospitalization and death cannot simply demonstrate the causal effect. There are some error terms that do not include in the model may have effect on both hospitalization and death. For example, if a patient identified as severe cases, it might cause he or she receives hospitalization service and death. Accordingly, our team assumes that there are 3798 cases that died because that they never receive hospitalization service.

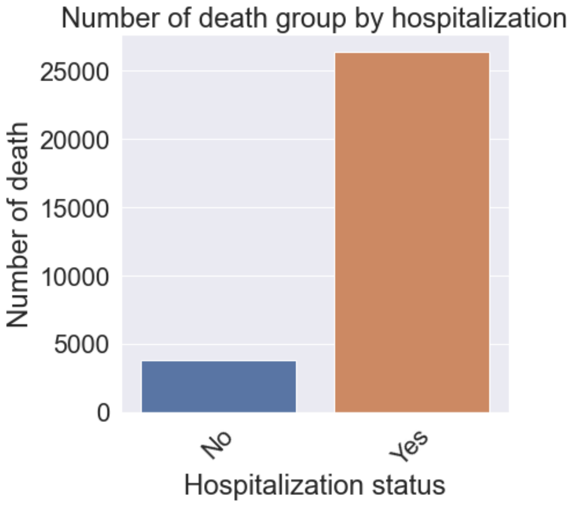

Figure 1. bar plot of hospitalization and death

2.Regression model

2.1 Model approach

Based on the intuition, our initial hypothesis is that there is a logistic relationship between the predictor variables and the log-odds of the event that response variable Y=1, which means being "dead". The ultimate goal is to not only observe the effect of X on Y but also find the most important factors for Covid-19 mortality. For example, it is possible to decrease the country-level related Covid-19 mortality by arranging hospitals' bed and ICU resources. Since some predictor variables are categorical variable, we need to create dummy variable to replace them. Then the logistic relationship can still be written in the following mathematical form:

$$\ell = log_b \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_{41} x_{41} + \beta_{42} x_{42} + \beta_{43} x_{43} + \beta_{44} x_{44} + \beta_{45} x_{45} + \beta_{46} x_{46} + \beta_{47} x_{47} + \beta_{48} x_{48} + \beta_{51} x_{51} + \beta_{52} x_{52} + \beta_{53} x_{53} + \beta_6 x_6 + \beta_7 x_7$$

where ℓ is the log-odds, b is the base of the logarithm, β_{ij} are parameters of the model, p=P(Y=1), and eighteen predictors:

$x_1, x_2 , x_3, x_{41}, x_{42}, x_{43}, x_{44}, x_{45}, x_{46}, x_{47}, x_{48}, x_{51}, x_{52}, x_{53}, x_6, x_7$.

Response variable(Y): the binary variable, death_yn, indicates that the death status.

Predictor variables(X):

X_1. binary variable x_2: 0 means probable case; 1 stand for laboratory-confirmed case.

X_2. binary variable: sex, 0 means female, 1 is male.

X_3. categorical variable: age_group, which divides all cases into 9 age groups. Prepressing this variable by transferring nine groups as eight binary variables X_{31}-X_{38} by leaving 0-9 years old group out.

X_4. categorical variable: race and ethnicity(combined), which includes Hispanic/Latino; American Indian/Alaska Native, Non-Hispanic; Asian, Non-Hispanic. Prepressing it by transferring four groups as three binary variables X_{51}-X_{53} by leaving Unknown out.

X_5. binary variable: hosp_yn, that indicates hospitalization status. 0 for no, 1 for yes.

X_6. binary variable: icu_yn, that indicates ICU admission status. 0 for no, 1 for yes.

X_7. binary variable: medcond_yn, which indicates the presence of underlying comorbidity or disease.

Given the large amount of data, our first step in model building is to explore the correlation between each independent variable. This initial analysis was designed for eliminating these variables that are high correlate with each other. A variable that is highly correlate with both some other independent variables and dependent variable could medicate the effect of some other independent variables on death. Thus, it would create challenge for analyzing the result of our final model. Accordingly, we conduct a heat map for exploring the correlation between each independent variable. Surprisingly, we find out that each independent variable is independent with each other. There is no such pair of variables that has correlation above 0.5 or below -0.5. In addition, we also use PCA method to test our result again. We will discuss PCA method in next paragraph.

After exploring the correlation between each independent variable, we proceed to build our final model use PCA to dimension reduction and use logistic regression to fit our data. To start with, our team split the data into training set and testing set for testing the result of our model. We decide to take 20% of the data as testing set and set random state at 42. Next, we build a pipeline to contain PCA and logistic regression to search for the best model for the data. The method we used is gridsearch CV. We use 5 folds cross validation method to each of 90 candidates, totaling 450 fits. The parameters we look for are the number of PCA components and parameter C in logistic regression. After the grid search, the best model we can obtain as our final model is logistic regression with C = 0.1 and 17 to 19 components.

2.2 Model evaluation

To evaluate the performance of our model, we analyze the classification ability by calculating precision, recall, and f1-score, and the overall area under the curve (AUC) of the Receiver Operating characteristic (ROC) curve.

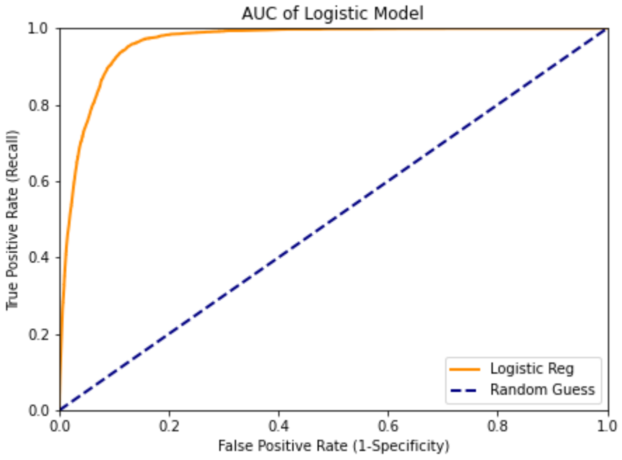

Figure 2, AUC of logistic model

The AUC gives the probability that a randomly selected pair of prediction (one death, one alive) would have their predicted probabilities correctly ordered. The AUC score of our model is 0.96. The high AUC score supposed to tell that our model is accurate to predict the odds of death. However, there is a question draws from the AUC score. For example, if a model supposed to predict the earthquake of a given area. The model says that the odds of the earthquake is 0. It is useless prediction, but the AUC score would be very high because earthquake is a rare event. Same as our model. COVID-19 has high infection rate but relatively low death rate. The death rate of our dataset is roughly equal to 7.33%. Consequently, the percent of true positives correctly classified would be the key factor to evaluate our model. Hence, we conduct classification report to best evaluate our model.

Table 3, Classification report:

	Precision	*Recall*	*F1-score*
0	0.97	0.98	0.97
1	0.66	0.47	0.55
accuracy			0.95
Macro avg	0.81	0.73	0.76
Weighted avg	0.95	0.95	0.95

Precision is the percent of true positives over all predicted positives. Recall is the percent of true positives over true positives plus false negatives. F1-score is 2 times precision times recall over precision plus recall. The table above shows that all score for predicted alive is super high, but scores for predicted death is around 0.55 to 0.66. Since death rate is relatively low, it is reasonable to predict the odds of alive correct, but it is hard to predict the odds of death accurate.

3.Regression analysis

3.1 Race and ethnicity analysis

If keep other independent variables fixed, we found that among different race and ethnicity, Asian American is associated with 13.93% increases in odds of COVID-19 mortality; Hispanic and Latino is associated with 10.64% increases in odds of COVID-19 mortality; African American is associated with 4.74% increases in odds of COVID-19 mortality. However, Native Hawaiian, other, and white is associated with decrease in odds of COVID-19 mortality. Native Hawaiian and other races decrease in odds of COVID-19 mortality probably because they have relatively small population and away from cities. The findings of white draw our attention because surprisingly white is associated with 18.55% decreases in odds of COVID-19 mortality. One possible explanation is that white represent a majority group of wealth holders. Wealthy individuals are more likely to receive better health care and support. Thus, it causes the decreases in odds of mortality of white.

Furthermore, one article posted in Health Affairs states that "Asian Americans experience a four times higher case fatality rate than that of the overall population[4]" in San Francisco. Due to lack of geographic control variables in our model, we cannot confirm that our findings are connected to the Asian COVID-19 death rate in San Francisco, but we can confirm that there is a problem associate with race and COVID-19 mortality. To address the issue, one possible explanation is that Asian American needs more health care resources and support in U.S. society. It reflects that not only Asian Americans but also other minorities such as Latino and African Americans need more health care support to survive in this COVID-19 pandemic. Accordingly, future studies between race and health care system would help reduce the odds of COVID-19 mortality.

3.2 Age group analysis

Based on our model, we also found that among different age groups, 20 to 29 years group has lowest odds of COVID-19 mortality, and above 80 years old group is associated with 2837.69% increase in odds of COVID-19 mortality, which is the maximum coefficient in our model.

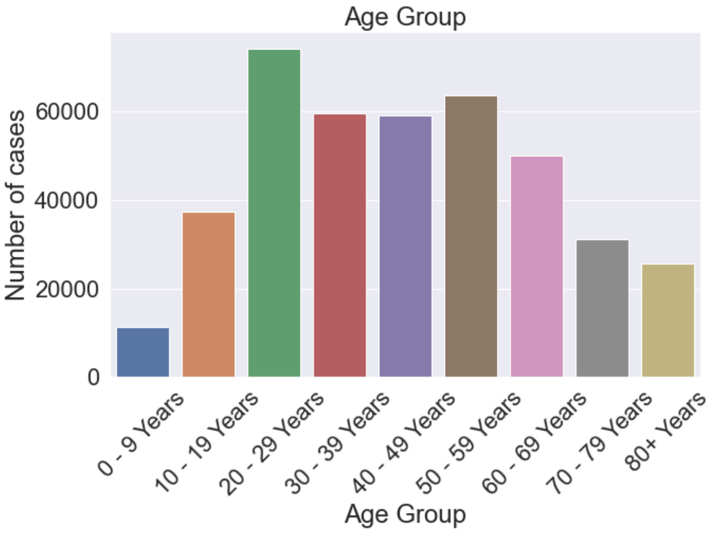

Figure 4, bar plot of age group with number of confirmed cases

The figure 4 demonstrates that the sample size is distributed like a bell-shape curve. 20 to 29 years age group has lowest odds of mortality but has the most cases. Seniors has highest death rate but has the lower cases. However, it cannot explain that children or seniors are less infected by the virus. The distribution of our sample might correlate to the total population of U.S. citizens.

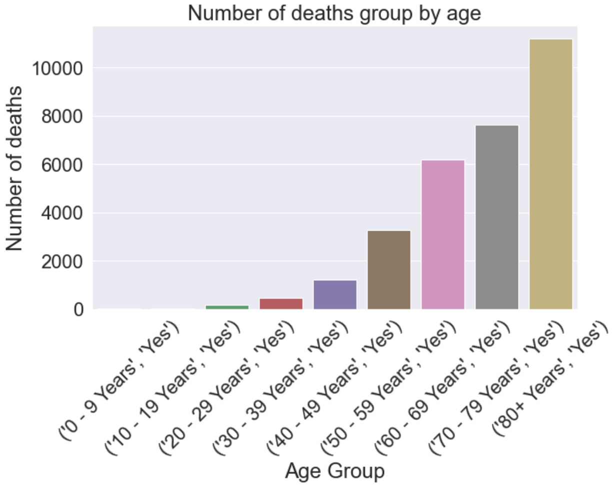

Figure 5, bar plot of age group with number of deaths

Nevertheless, if we plot the chart using the number of deaths rather than the total sample size, the chart will tell a different story. Though the figure 2 displays that 80+ years group has fewer positive cases than other groups, the figure 3 shows that patients in the 80+ age group has the greatest number of deaths than other age groups. Thus, the data clearly demonstrates that even though death rate for 20 to 29 years group is relatively low, this group of people are the major carriers of the coronavirus. Noticing that there are 52.83% cases presence of underlying comorbidity or disease, and therefore, 20 to 29 years group should try to social distancing and other methods to prevent the spread of the virus to protect the old people and themselves.

3.3 Hospitalization and ICU status analysis

The result shows that the indicators of resource allocation Hospitalization status and ICU admission status both have high positive coefficient: 1.6313 for hospitalization and 2.0487 for ICU. And the odds of Hospitalization and ICU is 5.1103 and 7.7575 individually, which means that if a patient is in hospital, he or she is 411.03% more likely to die rather than to alive; and if a patient is in ICU, he or she is 675.75% more likely to die rather than to alive. The conclusion sounds a bit awkward but is reasonable. That is because medical staffs will allocate the limited hospital beds and ICU rooms to patients with severe cases. Those patients are far more likely than common patients to die. The hospitalization and ICU assist them to capture the opportunity to be alive. Even though when they get hospital bed or ICU ward, they still face with high death possibility, the hospital will still arrange the hospital bed or ICU room to them as long as other common patients have a certain alive possibility without hospitalization and ICU.

3.4 Presence of underlying comorbidity analysis

The indicator medcond_yn (presence of underlying comorbidity or disease) also has high positive coefficient: 1.2, which indicates that a patient with underlying comorbidity is more likely to die than a patient without underlying comorbidity. The odds of medcond_yn is 3.3210, which means that the patient with underlying comorbidity is 232.10% more likely to die than to be alive. The result seems convincing. It is also worth to mention that even patients with underlying comorbidity are alive after treatment, they will inevitably suffer ambiguous sequelae, and their organs are no longer as healthy and complete as original.

Although the mortality of Covid-19 is not too high to regard Covid-19 as notorious disease, the influence of its comorbidity and ambiguous sequelae may cause far more serious problems in future. Therefore, anti-epidemic actions such as wearing facemask, washing hands, avoiding touch public facilities and staying at home as possible are needed. It’s one of the best ways to prevent people from getting infected with Covid-19 virus. The government can call or even force citizens to take certain actions.

3.5 Gender analysis

The indicator sex (1 represents for male and 0 represents for female) has a relatively small positive efficient, which indicates that sex has less influence on mortality. The odd of sex is 1.3053, which means that a male patient is 30.53% more likely to die than a female patient. That is because female have lower mortality rates at every age. For 10 leading causes of death in the united states in 2000, the odd of sex is 1.4113[5], which means that a male is 41.13% more likely to die than a female. The sex itself is biased for female. A possible explanation is that male may be less careful in life and have more bed habits (alcoholic, fight-abused, drug-taking). Therefore, the result that male patient is more likely to die is not meaningful in this case.

4. Limitation

The limitation of our model is that we cannot improve the model performance to accurately predict true positives, which is true deaths in our case. The best precision we get is around 66%. However, considering low death rate of COVID-19 dataset, we believe 66% is relatively good enough to perform analysis. Another limitation is that the dataset does not include geographic control variable. It limits us to perform analysis precisely. Moreover, a hospital must make a trade-off to save life as much as possible. The certain percentile arrangement cannot be inferred from this model because the dataset itself does not provide sufficient information. If the dataset can further identify patients by illness level (either determined by the apparatus status or by the time from getting infected), the model will be more precise and practical.

If we were going to continue work on the project, we would study the relation between race and public health care to find out what cause the increases in odds of COVID-19 mortality associate with different ethnicity.

5.Conclusion

In this model we try to figure out if certain characteristic of a patient will cause a higher mortality and then the result is able to assist hospitals make medical resources arrangement and facilitate the US government to designate the ‘biased’ healthcare policy for saving patients who are likely to be survive if they can get corresponding and on-time treatment. For instance, the regression shows that Asian American is associated with 13.93% increases in odds of COVID-19 mortality. If we put the income conditions of different ethnicity into considerations, it is obvious that ethnicity with low-income are more likely to die. A possible explanation is that some poor Asian American cannot afford health insurance. The government can buy the special health insurance designed for epidemic for those citizens to save life.

Reference

[1] “Estimating Mortality from COVID-19.” Edited by Scientific Brief, *World Health Organization*, World Health Organization, 4 Aug. 2020.

[2] Kim G-U, Kim M-J, Ra SH, Lee J, Bae S, Jung J, et al. Clinical characteristics of asymptomatic and symptomatic patients with mild COVID-19. Clin Microbiol Infect. 2020;26: 948.e1–948.e3.

[3] “COVID-19 Case Surveillance Public Use Data.” Edited by CDC Case Surveillance Task Force, *Data.CDC.org*, 5 Dec. 2020, data.cdc.gov

[4] Yan Fiona Ng Janet Chu Janice Tsoh Tung Nguyen, Brandon W, et al. “Asian Americans Facing High COVID-19 Case Fatality: Health Affairs Blog.” *Health Affairs*, 13 July 2020.

[5] POPULATION REFERENCE BUREAU. “The Gender Gap in U.S. Mortality.” *Population Reference Bureau*, 1 Dec. 2002.