Logistic Regression for Covid-19 Mortality

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 identifies and examines nine indicators (represented by eighteen predicter variables) that can classify people whether is “alive” or “dead”. This paper asserts that the model development is able to enhance a hospital’s resource arrangement ability, which may decrease the mortality in the epidemic. 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.

The data is provided by CDC Case Surveillance Task Force. 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. After removing missing and unknown values for sex, hospitalization status, ICU admission status, Death status, and Presence of underlying comorbidity, the remaining 411 thousand of samples will be used for this research propose. In addition, 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.

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. Therefore, 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:

$$\ell = log_b \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_{41} x_{41} + \beta_{42} x_{42} + \beta_{43} x_{43} + \beta_{44} x_{44} +$$

$$\beta_{45} x_{45} + \beta_{46} x_{46} + \beta_{47} x_{47} + \beta_{48} x_{48} + \beta_{51} x_{51} + \beta_{52} x_{52} + \beta_{53} x_{53} + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8 + \beta_9 x_9$$

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:

$x_1, x_2, x_3, x_{41}, x_{42}, x_{43}, x_{44}, x_{45}, x_{46}, x_{47}, x_{48}, x_{51}, x_{52}, x_{53}, x_6, x_7, x_8, x_9$.

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

Predictor variables(X):

X_1. discrete variable x_1: onset_dt, symptom onset date, if symptomatic

X_2. binary variable x_2: 0 means probable case; 1 stand for laboratory-confirmed case.

X_3. binary variable: sex, 0 means female, 1 is male.

X_4. categorical variable: age_group, which divides all cases into 9 age groups. Prepressing this variable by transferring nine groups as eight binary variables X_{41}-X_{48} by leaving 0-9 years old group out.

X_5. 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_6. binary variable: hosp_yn, that indicates hospitalization status. 0 for no, 1 for yes.

X_7. binary variable: icu_yn, that indicates ICU admission status. 0 for no, 1 for yes.

X_8. binary variable: medcond_yn, which indicates the presence of underlying comorbidity or disease.

X_9. binary variable: asymptomatic, indicates symptomatic or asymptomatic on initial case report date.

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. Our team would try to prove this casual effect assumption in our final model.

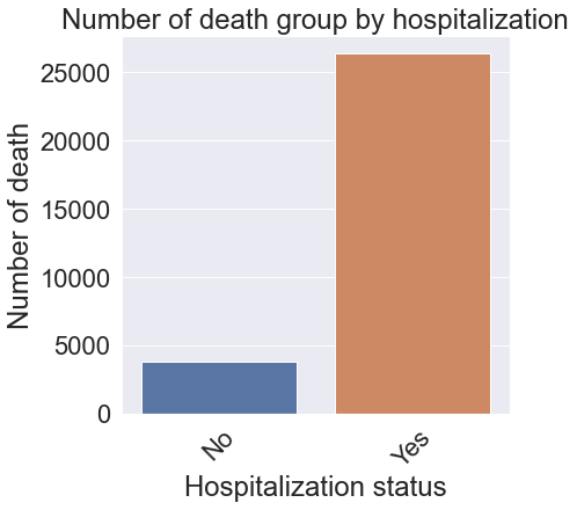

Figure 1. bar plot of hospitalization and death

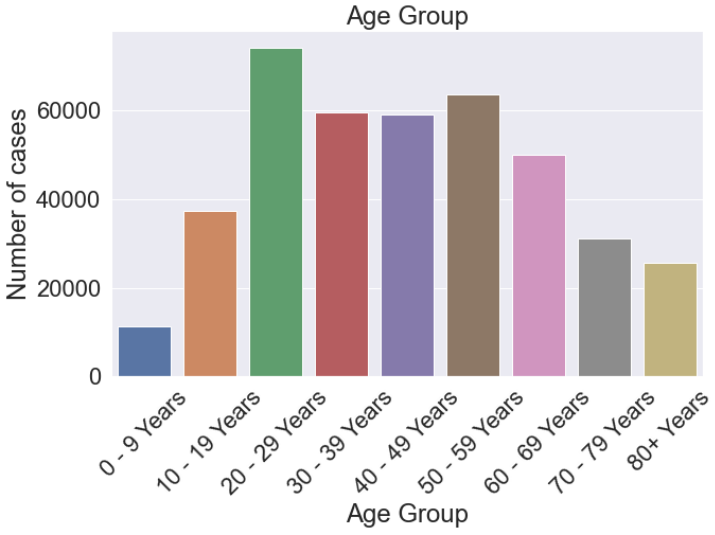

Figure 2. bar plot of age group and number of cases

The figure 2 demonstrates that the sample size is distributed like a bell-shape

curve. Children and seniors have less people infected, and 20 to 29 years age group has the most 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.

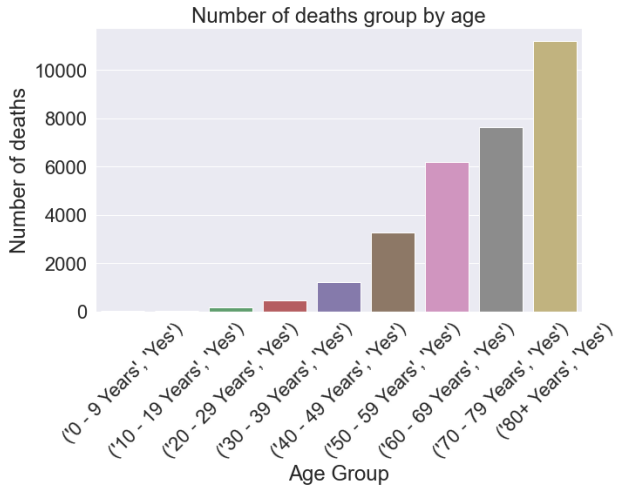

Figure 3. bar plot of age group and number of death

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.