AMS 572 Project Group 1 Members: Yichun Hao Ben Marantz
Topic: COVID-19 Surveillance Data Paul Vespe Zhe Zhou

COVID-19 SURVEILLANCE DATA - AN EXPLORATORY ANALYSIS

Introduction:

If we had one word to describe our lives for the past two years, it would be COVID-19. This virus has affected our daily lives in countless ways: how we work, shop, study, socialize, and more. The primary reason for all of this is because COVID is deadly. We know generally what makes us more at risk of dying, but policy decisions regarding the virus should not be made on these generalities. Instead, they should be made on concrete results obtained via statistical methods. As a result, we found it prudent to study COVID data, particularly COVID patient surveillance data provided by the CDC Case Surveillance Task Force. This data is reported to the CDC by regional health departments across the country and is typically collected from healthcare providers and COVID testing labs. With this data, we aimed to analyze the accuracy of reported estimates regarding the virus's death rate, the proportion of people infected with COVID-19 that die, both in general and with regard to age, which we believed to be one of the most important factors in the death rate. We obtained estimates for the death proportions from the National Center for Health Statistics and Johns Hopkins University of Medicine. Furthermore, we aimed to assess the significance of a patient's attributes on their chance of death using logistic regression. We wanted to determine when patients and healthcare professionals should take cases more seriously (as in, consider more intensive treatment) using the resulting logistic model.

To do this, we performed the following procedures:

Section 1. Comparing Reported Statistics to the Data - Surveyed U.S. COVID Patients.

Question 1. Are the death rates distributed among the following age groups in the proportions estimated by the National Center for Health Statistics?

Question 2. Is the true surveyed U.S. COVID patient death rate equal to the overall U.S. COVID death rate as estimated by Johns Hopkins University of Medicine?

Section 2. Logistic Regression for Patient Attributes on Death Probability

Question 1. Which patient attributes are significant predictors of COVID-19 death given infection?

Question 2. At what point should one take their COVID case more seriously (as in, consider more intensive treatment)?

To answer section 1 question 1, we performed a chi-squared goodness-of-fit test and found that the distribution of death proportion by age group among COVID patients was significantly different from what the National Center of Health Statistics reported. To answer section 1 question 2, we performed a one-sample proportion test and found that the death rate was significantly different from that estimated by Johns Hopkins University of Medicine. To answer section 2 question 1, we constructed a logistic regression model using several attributes and conducted tests for their significance. We also used stepwise selection and best subsets selection to select only the most significant attributes. To answer section 2 question 2, we used the knowledge from the previous procedure to choose the attributes to include in our final model for death probability, then found the best threshold for determining whether or not patients are at a relatively high risk of death (and therefore whether their cases should be taken more seriously - treated more intensively) using an ROC plot.

Unfortunately, because of the limitations associated with the dataset, we were unable to conclude anything substantial regarding our first hypothesis beyond our sample data; that is, our results (concluding that the distribution of death rates across ages and the total death rate were inaccurate) are not necessarily applicable to samples outside of the dataset. However, we reasoned that the results obtained from our logistic regression model are significant in the context of predicting death for infected individuals and for indicating when cases should be taken more seriously. In summary, we found the following variables to be significant: case month (month of relevancy since start of pandemic), current status (whether or not the case was confirmed in a lab), hospitalization status (whether or not the patient ended up in the hospital), and age. This could be of great aid to those taking care of infected individuals or individuals themselves who want to gauge their own risk.

Data Description

The data set included the following variables:

- Case Month (Ordinal) (month of positive test or CDC retrieval)
- Age Group (Ordinal) (0-17, 18-49, 50-64, 65+)
- Sex (Nominal) (Male, Female, Other)
- State of Residence (Nominal) (Includes DC and U.S. Territories)
- County of Residence (Nominal)
- Race (Nominal) (Native, Asian, Black, Multiple/Other, Pacific Islander, White)
- Ethnicity (Ordinal) (Hispanic, Non-Hispanic)
- Exposure (Ordinal)
  (whether or not the patient had been in a high illness-spreading risk situation)
- Underlying conditions (Ordinal) (yes or no / at least one underlying condition)
- Hospitalization (Ordinal) (yes or no)
- ICU (Ordinal) (yes or no)
- Symptom Status (Ordinal) (symptomatic / asymptomatic)
- Process (Nominal) (how the patient was determined to have the virus)
- Current status (Ordinal) (whether or not the case was laboratory confirmed)
- Case positive specimen interval (Numeric)
  (time in weeks until positive test was obtained)
- Case onset interval (Numeric) (time in weeks until symptoms started)
- Death (Ordinal) (yes or no - attributed to COVID)

The surveillance data contains over 30 million observations. However, the software we used (RStudio) could only parse the first 2 million or so. To reduce the runtime of our code, we selected a random sample of 100,000 observations from these 2 million and performed all of our methods on that same sample. For all of the categorical variables, we used the R method as.factor() to allow the models to treat them as a set of binary (0 - false, 1 - true) variables. We also converted the date variable (case month) from categorical to numeric, now representing the number of months since the start of 2020 (January 2020 = 1). This is because case month has a natural ordering, but the other ordinal variables in our dataset were binary, so it wasn't necessary to convert them the same way.

Although we clearly knew that we wanted to use age and death in answering our first two questions, we had to decide what additional variables to include for section 2. Including the wrong variables in our model could result in inaccurate predictions regarding the death rate of individuals, so it was important to include only the most relevant/useful. The primary cause of concern when selecting our variables was collinearity. This is when two or more variables have significant linear dependencies on each other. Including collinear variables can cause the model to attribute the outcome to a variable that's caused by another, thus misrepresenting their individual influence on the outcome and decreasing the significance of each. We decided to eliminate county of residence since it determines state (and the fact number of people per county was too small to make conclusions on), ethnicity since it can be determined from race (and race is a bit more specific), and ICU status since it determines hospitalization status.

A secondary reason for removing variables was due to ambiguity. The exposure and underlying conditions variables were binary when the effect of each type of exposure and underlying condition can vary wildly. A third reason was due to too-heavily skewed distributions. The vast majority of values for case positive specimen interval and case onset interval were 0 or 1, and those that were not zero varied wildly. The same applies for state. The vast majority of cases were from a small handful of states, and the rest were very sparsely included. More details about these variables are explained in the limitations section.

Missing Values

When dealing with very large datasets like ours, it is often the case that there are many missing values or observations, the existence of which can lead to biased or completely incorrect conclusions. From the sample of 100,000 observations we took, we found that the data contained 70,064 missing values across all of the variables we used for our study. To see the effects of having missing values, we conducted our procedure twice: the first ignoring observations with missing values and the second using all observations after having imputed the missing values.

Although having missing values can prove to be a rather difficult issue to overcome, not all missing data is to be treated equally, and by knowing the *type* of missing data, it is possible to mitigate its overall effect on results. In general, there are 2 types of missing data: missing completely at random (MCAR), and missing not at random (MNAR). The former is often due to human error/chance while the latter is usually related to the reason it's missing in the first place. Determining the type of missing data is a crucial first step in any data analysis, as failing to properly categorize missing data can lead to disastrous results. In general, MCAR tends to have less of an overall effect and can be more easily dealt with than MNAR.

Since our dataset had a lot of missing observations, one of our initial priorities was to determine the types of the missing data (MCAR vs MNAR). After going through and discussing each of the variables in our set, we concluded that two were MNAR, age and death, with our reasoning as follows. We reasoned that some age groups were more likely to not respond or report their ages than other groups, specifically younger people. As most teenagers must have been accompanied by an adult when getting their tests, we decided that some parents may have refused to give their child's age or other personal information away for their privacy. For the case of death, we hypothesized that many missing values were from people who died in their home as opposed to the hospital where they were located or from people whose death was unable to be determined to be from the virus. Since these deaths were less likely to be recorded by staff, we concluded that more of these missing values were "Yes" (death) than we would've expected. As for the other variables in our dataset (sex, race, ethnicity, hospitilization etc.), we decided that it didn't make logical sense for them to not be MCAR. For example, men do not seem less likely to report their gender than women, so we consider any missing values for sex to be random. No ethnicity or race seems to be less likely to report their ethnicity or race than another, so this is also random. Applying this logic to all other variables, we concluded that these were all MCAR.

For the variables that were MNAR, death rate and age group, we used external data to impute the missing values. In particular, we used demographic data from the CDC's COVID Data Tracker. To impute the data that was MCAR, we used a package in R known as MICE to impute the data. Our methods are explained further in the imputation section.

Below, we have included the distributions of all of the variables we used throughout our study, including the number of missing values from each (if applicable). These were obtained via the summary(x) command in R, where x in this case was the surveillance data after selecting only the variables we would use in our study.

case_month		age_group		sex	
Min:	1.00	0 - 17 years:	16428	Female:	51172
1st Qu:	9.00	18 to 49 years:	49194	Male:	44121
Median:	12.00	50 to 64 years:	17985	NA's:	4707
Mean:	11.75	65+ years:	14722		
3rd Qu:	14.00	NA's:	1671		
Max:	21.00				

race		current_status	
American Indian/Alaska Native:	355	Laboratory-confirmed case:	86610
Asian:	1237	Probable Case:	13390
Black:	9481		
Multiple/Other:	3008		
Native Hawaiian/Other Pacific Islander:	229		
White:	61147		
NA's:	24543		

symptom_status		hosp_yn		death_yn	
Asymptomatic:	5370	No:	81067	No:	82590
Symptomatic :	84684	Yes:	6137	Yes:	1009
NA's:	9946	NA's:	12796	NA's:	16401

Figure 1: Distribution of Used Variables with Missing Values

Section 1 - Before Imputation

In data analysis, it is not uncommon for conclusions to be drawn which are not necessarily significant or even true. The data might be heavily biased, wrong methods may have been used, or some incorrect assumptions may have been made to work around issues with the data. In the event that incorrect conclusions are drawn, the consequences can often be serious. As such, it is important and necessary to always double-check the validity of data and the statistical procedures used in making conclusions. With COVID-19 data, wrongly determining an individual's risks could result in increased deaths or infection rates. For instance, if a certain demographic is led to believe that COVID-19 poses very little risk to them, they might be less cautious about traveling, going out, and taking necessary precautions to prevent the spread. If it turns out that this claim is wrong, and that COVID-19 actually poses a very high risk for this group, then the results would be deadly. While it is clear to see how incorrect conclusions could affect individuals directly, policy-makers are also heavily influenced by these same conclusions. Decisions regarding mask/vaccine mandates or social distancing are made with these reported death rates taken into account, and under the assumption that they are accurate. Of course, in the event they are not accurate, the consequences will exist on a much larger scale. With this in mind, our group deemed it not only interesting but also necessary to determine the accuracy of the reported death rates among age groups for our data. We suspected that the reported death rates were accurate - that the true death rates are equal to the reported death rates. For the following tests, we chose to ignore observations in our data set that had missing values for age group and/or death. As a result, of the 100,000 observations, only 82,389 were used to conduct these tests.

Section 1 - Question 1

The main question we wanted to answer was the following: *Are the death rates distributed among the following age groups in the proportions estimated by the National Center for Health Statistics (NCHS)?*

Our hypotheses were as follows:

H_0 : The true proportions of death by age group are equal to those estimated by the NCHS.

H_1 : Not all of the true proportions are equal to those estimated by the NCHS.

Fortunately, both our data set and the NCHS estimates shared the following age group categories: 0-17 years, 18-49 years, 50-64 years, and 65+ years.
So, we tested the distribution of death among these age groups.

To organize our data, we used the summarise() method from the dplyr package.
These are the results from summarise() that we used. They return well-organized tables:

	Age.Group <chr>	COVID.19.Deaths <int>	Death.Proportion <dbl>
1	0-17 years	605	0.000790519
2	18 to 49 years	49886	0.065183191
3	50-64 years	139761	0.182617729
4	65+ years	575068	0.751408561

4 rows

Figure 2: Death by Age Group as Estimated by the NCHS

age_group <chr>	Death <int>	Survive <int>	Total <int>
0 - 17 years	0	13954	13954
18 to 49 years	1	43056	43057
50 to 64 years	20	14906	14926
65+ years	983	9469	10452

4 rows

Figure 3: Death by Age Group of CDC Surveillance Data

From these tables, we summarized things further to check the assumptions necessary for the chi-squared test. That is, that all of the expected counts are greater than 1 and that no more than 20% of these counts are less than 5. As seen in the table below, however, the expected number of deaths in the 0-17 group was less than 1, so we merged the 0-17 and 18-49 groups when performing our test.

Age Group	Estimated Proportion	Observed Count	Expected Count
0-17 Years	0.000790519	0	0.7936811
18-49 Years	0.065183191	1	65.4439241
50 - 64 Years	0.182617729	20	183.3481994
65+ Years	0.751408561	983	754.4141954
Total	1	1004	1004

Figure 4: Death Proportions and Counts by Age Group

To answer the question, we used a Chi-Squared Goodness of Fit Test with a significance level of 5%. The associated R command was chisq.test(x, p) where x is a vector of observed counts and p is a vector of proportions to test against (in our case, the proportions estimated by the NCHS). After running our test, we obtained a chi-squared test statistic of 279.04 with an associated p-value < 2.2e-16. With this, we rejected the null hypothesis, and concluded that the true proportions are different from the reported proportions.

Figure 5: Chi-Squared Goodness of Fit Test to answer S1 - Q1 in R

```
chisq.test(x = obs, p = prop.reported.final)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  obs
## X-squared = 279.04, df = 2, p-value < 2.2e-16
```

Section 1 - Question 2

While we were primarily interested in the death rates among different age groups, we thought we would also aim to answer the following question as well: *Is the true U.S. COVID death rate equal to the overall U.S. COVID death rate as estimated by Johns Hopkins University of Medicine?*

The Johns Hopkins University of Medicine estimate was 1.6%, which is approximately the same as what you can calculate from Google's reported total cases and deaths. Due to the nature of the data collection, there are reasons to suggest the overall death rate may be lower (people who visit their doctor or get tested may be less likely to die since they may care about their health more, people surveyed are likely to come from areas where healthcare is more widely available - perhaps better) or higher (people who visit their doctor or get tested may have a more serious case since they are more likely to be symptomatic, some cases may have been reported from hospitalized COVID patients which are likely more severe).

As a result, we used a two-sided alternative hypothesis:
H_0 : The true death rate is equal to 1.6%.
H_1 : The true death rate is not equal to 1.6%.

To answer the question, we used a one-sample proportion test with a significance level of 5%. The associated R command was prop.test(x, n, p) where x is the number of observed successes (deaths in our case), n is the number of observations, and p is the proportion to test against (1.6% in this case). The sample proportion obtained from the data was 0.01218609 or approximately 1.22%. After running our test, we obtained a chi-squared test-statistic of 75.877 and an associated p-value < 2.2e-16. The chi-squared distribution with 1 degree of freedom is equivalent to the squared Z distribution, so, in line with what we learned about proportion tests, we obtained a z test statistic of approximately 8.7107. The p-value remains the same. With this, we rejected the null hypothesis and concluded that the true overall death rate is different from the reported overall death rate (1.6%).

Figure 6: One-Sample Proportion Test to answer S1 - Q1 in R

```
prop.test(sum(data.table$Death), sum(data.table$Total), 0.016)
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  sum(data.table$Death) out of sum(data.table$Total), null probability 0.016
## X-squared = 75.877, df = 1, p-value < 2.2e-16
## alternative hypothesis: true p is not equal to 0.016
## 95 percent confidence interval:
##  0.01145345 0.01296459
## sample estimates:
##          p
## 0.01218609
```

After running our tests, we found that the results of our tests were more significant than we expected. This may be due to bias present in the data set we used. Although our sample size seemed sufficiently large, it may not have been representative of the demographics we were testing. It is possible that the observations in our sample were less likely to die from COVID-19 than the observations in the data used by National Center for Health Statistics and Johns Hopkins in their respective age groups. This may be because by going to their doctor, surveyed COVID patients are more likely to be concerned about their health and take the right steps for survival. In addition, our sample proportion is likely lower than the actual proportion of deaths due to the likelihood that the missing death values are actual deaths. As a result, before imputing missing values, we cannot be confident that we can draw proper conclusions about the population from these tests.

Section 2 - Before Imputation

Throughout the past two years, we have likely heard that older people or people with underlying conditions are more likely to die from the virus, and that even if these don't apply to you, you need to take precautions to safeguard the health of these people. Although we believe these factors to be significant in computing death probabilities, it is necessary in statistics to assess the significance of other variables as well. Demographic attributes like sex and race, whether or not you're symptomatic, the time during which you were infected, and more. All of these factors are likely to affect one's chance of death, so it is necessary to determine which ones are the most significant; or, more practically, which ones are significant with respect to other variables. Once it is known which variables are significant in predicting the desired outcome (in this case, chance of death), it is then possible to apply this information to individuals who wish to determine their risk. For those who might be unsure if they are at a high-risk of dying from COVID-19, a predictive model would be a useful tool in decision-making regarding their health.

In statistics, one of the most popular methods used for making predictive models is that of regression. Simply put, regression models take in a predetermined set of variables of attributes (regressors), and use these to predict an associated output. In predicting binary outcomes specifically, like whether an individual lives or dies based on given inputs, it is often useful to use a specific type of regression known as logistic regression, which outputs a value between 0 and 1. For example, in the case of predicting an individual's risk for death from COVID-19, the regressors might include any of the following previously mentioned (age, death, sex, etc.), while the dependent variable we are trying to predict would be the chance of death. However, before creating and implementing a regression model, it is necessary to determine which variables to include in the model at all. In order to determine which variables were significant (and therefore which to include in our model), we first had to clean our data. Since we needed real-valued data, we first chose to ignore observations in our data set that had missing values for any of the variables we were considering. The R command to get only the observations without missing values is complete.cases(), which takes a vector or data frame as input. After implementing this method, we found that only 58,796 observations out of the original 100,000 contained no missing values.

Since death has only two classifications where the probabilities of each class add to one, we can treat a "Yes" value as 1 and a "No" value as 0. We want to determine a patient's probability of death, so our model needs to produce a value between 0 and 1 while taking a patient's attributes as input.

The logistic regression function can model a probability, calculated by:

$$y = \frac{e^{\beta_0 + \sum_{i=1}^{k} \beta_i x_i}}{1 + e^{\beta_0 + \sum_{i=1}^{k} \beta_i x_i}}$$

Recall that the sum in the exponents of the function is the multiple linear regression equation where there are k independent variables. Like with multiple linear regression, our goal is to obtain estimates of the coefficients (betas in the function) such that the resulting curve best fits the distribution of outcomes.

However, instead of minimizing the sum of squared residuals, logistic regression minimizes what's called the log loss, whose formula is:

$\sum_{i=1}^{n} (1 - y_i) ln(1 - \hat{y}_i) - y_i ln(\hat{y}_i)$ where y_i is the actual outcome and $\hat{y}_i$ is the predicted outcome as calculated from the logistic regression function with the estimates of the β_i's.

To be able to assess how well our model can predict death probability, we split the 58,796 observations into two groups. We took 80% of the observations (the training set) for building the logistic regression model and for evaluating the significance of the regressors. We used the remaining 20% of the observations (the testing set) to assess the predictive capability of the model. To do the splitting, we used the createDataPartition(y, p) method from the caret package in R. This returns a subset of indices (row indices for data frames) from a vector or data frame y. The p argument represents the proportion of the indices to take (we used 0.8 in our case). We split the data into the training set and testing set by indexing the data with the result of createDataPartition.

To build our model, we used the glm(y ~ x, family, data) method in R with the training set as our data argument input. glm stands for generalized linear model, which is a type of model that includes both linear and logistic regression models. To fit a logistic regression model in particular, we set the family argument to "binomial("logit")".

Section 2 - Question 1

To answer our first question, *Which patient attributes are significant predictors of COVID-19 death given infection?*, we performed significance tests for the coefficients of our regression function. If a variable has no effect on our outcome, we would expect it's coefficient to be zero (so it contributes nothing). Since the data contains categorical variables with more than two categories, they are treated as a set of (c - 1) binary variables where c is the number of categories. It is (c - 1) as opposed to c because, by knowing the values of the (c - 1) variables, the value for the last category can be determined. If any of the other variables have a value of 1 (the observation falls under that category), then the value for the last category is 0 (the observation can't fall under this category) and likewise, if none of the other variables have a value of 1 (the observation falls under none of those categories), then the value for the last category is 1 (the observation must fall under this category). As a result, we perform significance tests for each of these binary variables separately. However, if one of these binary variables returns a significant result, we will conclude that the original variable it comes from is significant. Since we are performing several tests, we decided to use two-sided alternatives for all of them.

Example hypotheses (done for each independent variable, not just hospitalization status):
H_0 : $\beta_{hospitalization\ status}$ is equal to 0.
H_1 : $\beta_{hospitalization\ status}$ is not equal to 0.

To perform these tests, we used the summary(x) command in R where x, in this case, is a glm object. This method automatically performs two-sided tests for the coefficients, and marks p-values that are within certain intervals. For our significance level of 5%, we looked for at least one asterisk next to the p-value (shows that it's less than 5%).

From the R output below, we concluded that case month, current status (if the patient's case was laboratory confirmed or not), and hospitalization status are significant predictors of death, all with p-values near zero. In other words, these are the attributes that have the most effect on a COVID patient's death chance after controlling for their other attributes.

```
## glm(formula = death_yn ~ (.), family = binomial("logit"), data = train)
##
##
## Coefficients:
##                                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)                             -36.54200 1564.21940  -0.023   0.9814
## case_month                               -0.12749    0.01449  -8.796  < 2e-16 ***
## age_group18 to 49 years                  11.06892  330.46971   0.033   0.9733
## age_group50 to 64 years                  13.78106  330.46841   0.042   0.9667
## age_group65+ years                       17.96019  330.46820   0.054   0.9567
## sexMale                                   0.17629    0.10371   1.700   0.0892 .
## raceAsian                                16.15384 1528.91250   0.011   0.9916
## raceBlack                                15.82483 1528.91238   0.010   0.9917
## raceMultiple/Other                       14.92332 1528.91254   0.010   0.9922
## raceNative Hawaiian/Other Pacific Islander  1.31508 2491.38730   0.001   0.9996
## raceWhite                                16.33155 1528.91237   0.011   0.9915
## current_statusProbable Case               0.76273    0.16229   4.700 2.61e-06 ***
## symptom_statusSymptomatic                 0.10522    0.23903   0.440   0.6598
## hosp_ynYes                                2.08281    0.10950  19.021  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5126.5  on 47037  degrees of freedom
## Residual deviance: 2796.7  on 47024  degrees of freedom
## AIC: 2824.7
##
## Number of Fisher Scoring iterations: 20
```

Figure 7: Significance Tests to answer S2 - Q1 in R

To answer our second question, *At what point should one take their COVID case more seriously (as in, consider more intensive treatment)?*, we aimed to construct a model that can accurately predict a COVID patient's death probability. To obtain a robust model, it is important to remove insignificant variables from the model. To do so, we used two different variable selection methods: stepwise selection and best subsets selection. Both methods iteratively test different subsets of the variables based on different criteria and return the model with the best value. We first used stepwise selection using the step() function in R, which uses the model's AIC value to make comparisons. The lower the AIC of the model, the better it is. We then used best subsets selection using the regsubsets() function from the leaps package in R, which uses the model's BIC value to make comparisons. The lower the BIC of the model, the better it is. Both functions take the previously-fitted logistic regression model as input.

The best model, as concluded by stepwise selection, included the following attributes: case month, age group, sex, race, current status, and hospitalization status. Same as before, the p-values for case month, current status, and hospitalization are significant. Sex, race, and age group may have been selected because they have an important effect on the outcome, but their standard error was too high for them to be statistically significant. Perhaps there are more significant predictors than we concluded previously?

```
## glm(formula = death_yn ~ case_month + age_group + sex + race +
##     current_status + hosp_yn, family = binomial("logit"), data = train)
##
## Coefficients:
##                                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)                              -36.4441  1564.8332  -0.023   0.9814
## case_month                                -0.1276     0.0145  -8.803  < 2e-16 ***
## age_group18 to 49 years                   11.0721   330.4839   0.034   0.9733
## age_group50 to 64 years                   13.7838   330.4826   0.042   0.9667
## age_group65+ years                        17.9595   330.4823   0.054   0.9567
## sexMale                                    0.1786     0.1036   1.724   0.0847 .
## raceAsian                                 16.1567  1529.5374   0.011   0.9916
## raceBlack                                 15.8203  1529.5373   0.010   0.9917
## raceMultiple/Other                        14.9227  1529.5375   0.010   0.9922
## raceNative Hawaiian/Other Pacific Islander  1.3231  2491.6079   0.001   0.9996
## raceWhite                                 16.3311  1529.5373   0.011   0.9915
## current_statusProbable Case                0.7614     0.1623   4.692  2.7e-06 ***
## hosp_ynYes                                 2.0890     0.1087  19.225  < 2e-16 ***
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5126.5  on 47037  degrees of freedom
## Residual deviance: 2796.9  on 47025  degrees of freedom
## AIC: 2822.9
##
## Number of Fisher Scoring iterations: 20
```

Figure 8: Result of Stepwise Selection for S2 - Q2 in R

Best subsets selection returns the best model for subsets of 1, 2, 3, 4, and 5 variables. We used the kable() function from the knitr package in R to output the results in a neat table, as shown below. As can be seen, the BIC decreases (and therefore, the model gets better) as the size of the subset increases. However, we need to determine a cut-off point - we should select a model such that, when adding another variable, the resulting decrease in the BIC is negligible. We determined this to be after the fourth model, which saw the lowest decrease in BIC. That fourth model contains the variables case month, age group, current status, and hospitalization status. All of these variables were included in the result of stepwise selection, and all of the variables found significant via the tests are still included, so we decided to use this model for predicting a COVID patient's death probability. While not using a statistical test, by including age group in our final model, we are concluding that it is a significant predictor.

```r
M <- regsubsets(death_yn ~ ., data=train, nbest = 1 , nvmax=5, method = 'forward', intercept = TRUE)
temp <- summary(M)
Var <- colnames(model.matrix(M.raw))
M_select <- apply(temp$which, 1, function(x) paste0(Var[x], collapse='+'))
kable(data.frame(cbind(model = M_select, BIC = temp$bic)), caption='Model Summary')
```

Model Summary

model	BIC
(Intercept)+age_group65+ years	-3058.67416567003
(Intercept)+age_group65+ years+hosp_ynYes	-4794.58185024012
(Intercept)+case_month+age_group65+ years+hosp_ynYes	-4845.98243275054
(Intercept)+case_month+age_group65+ years+current_statusProbable Case+hosp_ynYes	-4853.09912349678
(Intercept)+case_month+age_group50 to 64 years+age_group65+ years+current_statusProbable Case+hosp_ynYes	-4856.55308964

Figure 9: Result of Best Subsets Selection for S2 - Q2 in R

Section 2 - Question 2

To answer our second question, *At what point should one take their COVID case more seriously (as in, consider more intensive treatment)?*, we aimed to construct a model that can accurately predict a COVID patient's death probability. From our significance tests and selection methods, we decided to use only case month, current status, hospitalization status, and age group in our final model summarized below:

```r
M.final <- glm(death_yn ~ (case_month + age_group + current_status + hosp_yn), family = binomial("logit"), data = train)
summary(M.final)
```

```
##
## Call:
## glm(formula = death_yn ~ (case_month + age_group + current_status +
##     hosp_yn), family = binomial("logit"), data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0438  -0.0280  -0.0085  -0.0059   4.0987
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                -20.17149  332.76211  -0.061    0.952
## case_month                  -0.12353    0.01434  -8.615  < 2e-16 ***
## age_group18 to 49 years     11.06510  332.76358   0.033    0.973
## age_group50 to 64 years     13.80831  332.76229   0.041    0.967
## age_group65+ years          18.01801  332.76208   0.054    0.957
## current_statusProbable Case  0.75318    0.16173   4.657 3.21e-06 ***
## hosp_ynYes                   2.06587    0.10769  19.184  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5126.5  on 47037  degrees of freedom
## Residual deviance: 2816.4  on 47031  degrees of freedom
## AIC: 2830.4
##
## Number of Fisher Scoring iterations: 20
```

Figure 10: Final Model used to answer S2 - Q2 in R

To assess whether or not our final model accurately predicts death probability, we applied our model to the testing test using the predict(model, data, type) function from the caret package in R, which takes a model (our final model) and data (the testing set) as input. This function calculates the model's predicted outcomes (by setting type = "response") for each observation in the data input and returns them. Using these predicted outcomes (death probabilities), we compared these to the actual outcomes in the testing set to see how accurate our predictions were. At first, we tried classifying patients with a predicted death chance of 50% or more as dying and the rest as surviving. However, all of the predicted probabilities were less than 50%, so doing so yielded 100% survival. This simply means that regardless of a patient's attributes (of the ones we selected for the final model), they are more likely to survive than not. This is great knowledge, but it doesn't help to answer our question. We need to determine a probability threshold for people to assess their relative risk of death. Fortunately, we can obtain such a threshold via an ROC (Receiver Operating Characteristic) plot. This plots two measures of a logistic model's accuracy, sensitivity and specificity, against each other for different classification thresholds. From the resulting curve, one can calculate the best predictive classification threshold. Specificity is the proportion of actual negatives (survivals in this case) that are correctly classified and sensitivity is the proportion of actual positives (deaths in this case) that are correctly classified. The formulas are shown below:

$$Specificity = \frac{True\ Negatives}{True\ Negatives + False\ Positives} \qquad Sensitivity = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

To obtain the ROC plot, we used the roc(response, predictor) method from the pROC package in R (citation at end of report), then called the plot() method on that object. From the plot, a classification threshold of 0.001 yielded the best sensitivity and specificity of 99.1% and 86.5%, respectively. The ROC plot also returns an AUC (area under curve) value, which is another measure of the model's accuracy. The value of 0.968 is very high considering the maximum value is 1, which shows that our model is good at predicting COVID death. Therefore, we concluded that if a patient's predicted death probability is greater than 0.001, we should consider that patient to be at a high relative risk of death and therefore consider more intensive treatment. Or, if your predicted death probability is 0.001, you should take your case more seriously.

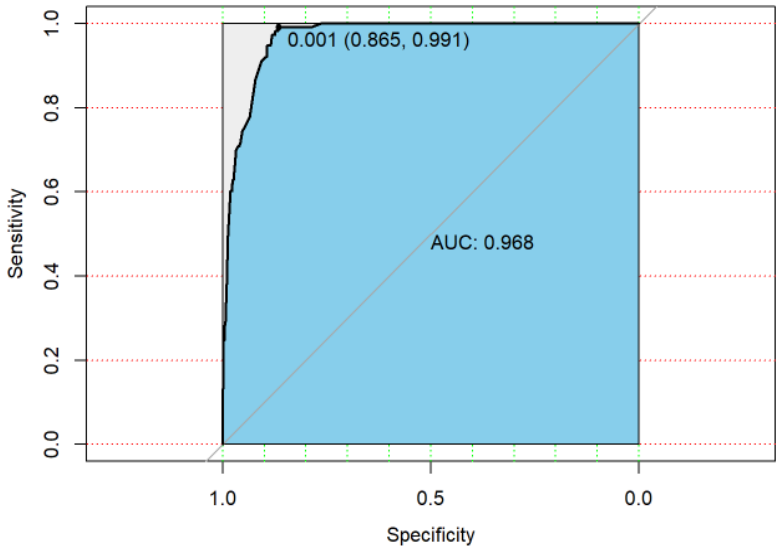

Figure 11: ROC Plot used to answer S2 - Q2 in R

This procedure was conducted before imputing missing values. With our first two questions, we reasoned our results may have been more significant than they should have been due to a perceived bias in the death and age group variables. Both of these variables were used throughout this procedure. As a result, we cannot be confident that we can draw proper conclusions about the population from these results.

Imputation

For our first set of questions, we only needed to consider the variables 'death' and 'age group', both of which we determined to be MNAR. For age, we looked at the proportions of infected individuals in the US across age groups. After generating randomly distributed uniform numbers for every missing value (between 0 and 1), we assigned an age group depending on the proportions reported in the CDC COVID Data Tracker summarized. For example, if 20% of infected individuals were from 0-17 years of age, we would set a missing value to (0-17 years) if the randomly generated number was less than or equal to 0.2 (the proportion). For death, we used a similar mechanism, except that we compared the randomly generated number to the death rate of the observation's age group calculated from the same data. So, if 10% people in the 65 and older age group died, we would set a missing value to Yes if the randomly generated number was less than or equal to 0.1 (the rate).

To deal with the MCAR values, we decided to use a package in R known as 'MICE' (Multivariate Imputation by Chained Equations). This package uses multiple regression to predict the values of missing variables based on the other variables associated. That is, if there are X1, X2...Xk variables, and we want to predict missing values of X1, we regress X1 on the other X2...Xk variables and get an output (for continuous variables). For binary/categorical variables, it uses logistic regression. MICE actually computes multiple different versions of imputations (different datasets), and then compares them to determine the best (most accurate) one. Although we didn't use any of the MCAR data for hypothesis 1 (proportion testing), we did use it in hypothesis 2 (logistic regression), so MICE was primarily used for that purpose.

Section 1 - After Imputation

Section 1 - Question 1

Recall the question: *Are the death rates distributed among the age groups in the proportions estimated by the National Center for Health Statistics (NCHS)?*

Our hypotheses were as follows:

H_0 : The true proportions of death by age group are equal to those estimated by the NCHS.

H_1 : Not all of the true proportions are equal to those estimated by the NCHS.

Our new summarise() result was as follows:

	age_group <fct>	Death <int>	Survive <int>	Total <int>
1	0 - 17 years	0	16664	16664
2	18 to 49 years	21	50053	50074
3	50 to 64 years	77	18237	18314
4	65+ years	1407	13541	14948

4 rows

Figure 12: Death by Age Group of CDC Surveillance Data (After Imputation)

From this table, we summarized things further to check the assumptions necessary for the chi-squared test. That is, that all of the expected counts are greater than 1 and that no more than 20% of these counts are less than 5. As seen in the table below, however, 25% of the counts are less than 5 (from the 0-17 group), so we merged the 0-17 and 18-49 groups when performing our test, same as before.

Age Group	Estimated Proportion	Observed Count	Expected Count
0-17 Years	0.000790519	0	1.189731095
18-49 Years	0.065183191	21	98.10070246
50 - 64 Years	0.182617729	77	274.8396821
65+ Years	0.751408561	1407	1130.869884
Total	1	1505	1505

Figure 13: Death Proportions and Counts by Age Group (After Imputation)

To answer the question, we used the same test with the new observed counts. After running, we obtained a chi-squared test statistic of 271.57 with an associated p-value < 2.2e-16. With this, we rejected the null hypothesis, and concluded that the true proportions are different from the reported proportions.
While the resulting statistic was slightly less than the previous 279.04, the p-value is still extremely small. As a result, imputing missing values did not affect our test's conclusion.

Figure 14: Chi-Squared Goodness of Fit Test to answer S1 - Q1 in R (After Imputation)

```
chisq.test(x = obs.com, p = prop.reported.final)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  obs.com
## X-squared = 271.57, df = 2, p-value < 2.2e-16
```

Section 1 - Question 2

Recall the question: *Is the true U.S. COVID death rate equal to the overall U.S. COVID death rate as estimated by Johns Hopkins University of Medicine?*

Where the Johns Hopkins University of Medicine estimate was 1.6%.

The hypotheses:
H_0 : The true death rate is equal to 1.6%.
H_1 : The true death rate is not equal to 1.6%.

To answer the question, we used the same test with the new observed successes (deaths in our case) and total number of observations (now the full 100,000). The sample proportion obtained from the data was 0.01505 or 1.505%. After running our test, we obtained a chi-squared test-statistic of 5.6722 and an associated p-value = 0.01724. The chi-squared distribution with 1 degree of freedom is equivalent to the squared Z distribution, so, in line with what we learned about proportion tests, we obtained a z test statistic of approximately 2.3816. The p-value remains the same. With this, we rejected the null hypothesis and concluded that the true overall death rate is different from the reported overall death rate. The resulting statistic is much less than the previous 8.7107, and the p-value is no longer approximately zero but a value where we would have come to a different conclusion if we chose a lower significance level. However, since the p-value is below 5%, we reached the same conclusion. So, while imputing missing values did not change our conclusion, it did clearly alter the significance of our results. As a result, it is likely that the missing values for death in the surveillance data are more likely to be deaths than we would expect.

Figure 15: One-Sample Proportion Test to answer S1 - Q1 in R (After Imputation)

```
prop.test(sum(data.table.com$Death), sum(data.table.com$Total), 0.016)
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  sum(data.table.com$Death) out of sum(data.table.com$Total), null probability 0.016
## X-squared = 5.6722, df = 1, p-value = 0.01724
## alternative hypothesis: true p is not equal to 0.016
## 95 percent confidence interval:
##  0.01430893 0.01582858
## sample estimates:
##       p
## 0.01505
```

Section 2 - After Imputation

Section 2 - Question 1

Recall the question:

Which patient attributes are significant predictors of COVID-19 death given infection?

Example hypotheses (done for each independent variable, not just hospitalization status):

H_0 : $\beta_{hospitalization\ status}$ is equal to 0.

H_1 : $\beta_{hospitalization\ status}$ is not equal to 0.

From the R output below, we concluded that case month, current status (if the patient's case was laboratory confirmed or not), and hospitalization status are significant predictors of death, all with p-values near zero. In other words, these are the attributes that have the most effect on a COVID patient's death chance after controlling for their other attributes. The results are extremely similar to what we obtained before imputation.

```
## Call:
## glm(formula = death_yn ~ (.), family = binomial("logit"), data = train.com)
##
## Coefficients:
##                                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)                             -2.085e+01  1.521e+02  -0.137    0.891
## case_month                              -8.019e-02  8.175e-03  -9.809  < 2e-16 ***
## age_group18 to 49 years                  1.278e+01  1.521e+02   0.084    0.933
## age_group50 to 64 years                  1.478e+01  1.521e+02   0.097    0.923
## age_group65+ years                       1.756e+01  1.521e+02   0.116    0.908
## sexMale                                  7.269e-02  6.248e-02   1.163    0.245
## raceAsian                                3.065e-01  1.125e+00   0.272    0.785
## raceBlack                                8.020e-01  1.031e+00   0.778    0.436
## raceMultiple/Other                       8.224e-01  1.046e+00   0.786    0.432
## raceNative Hawaiian/Other Pacific Islander -1.339e+01  1.131e+03  -0.012    0.991
## raceWhite                                1.133e+00  1.025e+00   1.105    0.269
## current_statusProbable Case              5.274e-01  9.810e-02   5.377 7.59e-08 ***
## symptom_statusSymptomatic                7.901e-02  1.429e-01   0.553    0.580
## hosp_ynYes                               1.453e+00  6.515e-02  22.307  < 2e-16 ***
##
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 12494.7  on 79999  degrees of freedom
## Residual deviance:  7953.7  on 79986  degrees of freedom
## AIC: 7981.7
##
## Number of Fisher Scoring iterations: 19
```

Figure 16: Significance Tests to answer S2 - Q1 in R (After Imputation)

Once we obtained these results, we used stepwise selection and best subsets selection to find a more robust model for predicting death. Recall that stepwise selection uses AIC for comparisons, best subsets selection uses BIC for comparisons, and that lower values for both AIC and BIC indicate a better model.

The best model, as concluded by stepwise selection, included the following attributes: case month, age group, race, current status, and hospitalization status. Interestingly, this model did not include the variable 'sex', which *was* included before when we ignored missing values. Same as before, the p-values for case month, current status, and hospitalization are significant. Race, and age group may have been selected because they have an important effect on the outcome, but their standard error was too high for them to be statistically significant. Perhaps there are more significant predictors than we concluded previously?

```
## Call:
## glm(formula = death_yn ~ case_month + age_group + race + current_status +
##     hosp_yn, family = binomial("logit"), data = train.com)
##
## Coefficients:
##                                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                                -2.074e+01  1.521e+02  -0.136    0.892
## case_month                                 -8.031e-02  8.176e-03  -9.823  < 2e-16 ***
## age_group18 to 49 years                     1.278e+01  1.521e+02   0.084    0.933
## age_group50 to 64 years                     1.478e+01  1.521e+02   0.097    0.923
## age_group65+ years                          1.756e+01  1.521e+02   0.115    0.908
## raceAsian                                   3.016e-01  1.125e+00   0.268    0.789
## raceBlack                                   7.991e-01  1.030e+00   0.776    0.438
## raceMultiple/Other                          8.252e-01  1.045e+00   0.789    0.430
## raceNative Hawaiian/Other Pacific Islander -1.340e+01  1.131e+03  -0.012    0.991
## raceWhite                                   1.134e+00  1.025e+00   1.106    0.269
## current_statusProbable Case                 5.260e-01  9.808e-02   5.363  8.2e-08 ***
## hosp_ynYes                                  1.460e+00  6.461e-02  22.593  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 12494.7  on 79999  degrees of freedom
## Residual deviance:  7955.4  on 79988  degrees of freedom
## AIC: 7979.4
##
## Number of Fisher Scoring iterations: 19
```

Figure 17: Result of Stepwise Selection for S2 - Q2 in R (After Imputation)

Recall that best subsets selection returns the best model for subsets of 1, 2, 3, 4, and 5 variables. As when we did best subset selection before imputing missing values, we determined the cut-off to be after the fourth model, which saw the lowest decrease in BIC (which we determined to be negligible). That fourth model contains the variables case month, age group, current status, and hospitalization status. All of these variables were included in the result of stepwise selection, and all of the variables found significant via the tests are still included, so we decided to use this model for predicting a COVID patient's death probability. While not using a statistical test, by including age group in our final model, we are concluding that it is a significant predictor. The only difference noted between the results of best subset selection before and after imputing missing values is in the best subset with five variables. Before imputation, the fifth variable was another age group variable. Now, the fifth variable is race. The variables included in the subset we selected were the same, however.

model	BIC
(Intercept)+age_group65+ years	-6056.71695541999
(Intercept)+age_group65+ years+hosp_ynYes	-7915.17733179056
(Intercept)+case_month+age_group65+ years+hosp_ynYes	-7976.92660037921
(Intercept)+case_month+age_group65+ years+current_statusProbable Case+hosp_ynYes	-7988.77782491967
(Intercept)+case_month+age_group65+ years+raceWhite+current_statusProbable Case+hosp_ynYes	-7990.18757522276

Figure 18: Result of Best Subsets Selection for S2 - Q2 in R (After Imputation)

Section 2 - Question 2

Recall the question: *At what point should one take their COVID case more seriously?*

We found this to be equivalent to the following question: At what probability threshold should a COVID patient be considered to have a high risk of death, needing more intensive care?

From our significance tests and selection methods, we decided to use only case month, current status, hospitalization status, and age group in our final model summarized below:

```
# Final model
M.final.com <- glm(death_yn ~ (age_group + case_month + hosp_yn + current_status),
                   family = binomial("logit"), data = train.com)
summary(M.final.com)
```

```
##
## Call:
## glm(formula = death_yn ~ (age_group + case_month + hosp_yn +
##     current_status), family = binomial("logit"), data = train.com)
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 -19.673901 152.354965  -0.129    0.897
## age_group18 to 49 years      12.770400 152.355099   0.084    0.933
## age_group50 to 64 years      14.789214 152.354986   0.097    0.923
## age_group65+ years           17.593937 152.354941   0.115    0.908
## case_month                   -0.079771   0.008166  -9.769  < 2e-16 ***
## hosp_ynYes                    1.435174   0.064189  22.358  < 2e-16 ***
## current_statusProbable Case   0.535466   0.098010   5.463 4.67e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 12494.7  on 79999  degrees of freedom
## Residual deviance:  7972.6  on 79993  degrees of freedom
## AIC: 7986.6
##
## Number of Fisher Scoring iterations: 19
```

Figure 19: Final Model used to answer S2 - Q2 in R (After Imputation)

To assess whether or not our final model accurately predicts death probability, we applied our model to the testing set. As before, we used the ROC plot to answer the question using the final model.
As a reminder,

$$Specificity = \frac{True\ Negatives}{True\ Negatives + False\ Positives} \qquad Sensitivity = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

According to the plot, by using a probability threshold of 0.011, we were able to obtain a sensitivity of 98%, a specificity of 84.8%, and an accuracy of 95.6%. With these results, we can confidently say that our model does a great job of predicting deaths/survivals with a relatively low threshold after imputing the missing data values.

From the plot, a classification threshold of 0.011 yielded the best sensitivity and specificity of 98.0% and 84.8%, respectively. The AUC value of 0.956 is still very high considering the maximum value is 1, which shows that our model is good at predicting COVID death. Therefore, we concluded that if a patient's predicted death probability is greater than 0.011, we should consider that patient to be at a high relative risk of death and therefore consider more intensive treatment. Or, if your predicted death probability is 0.011, you should take your case more seriously.

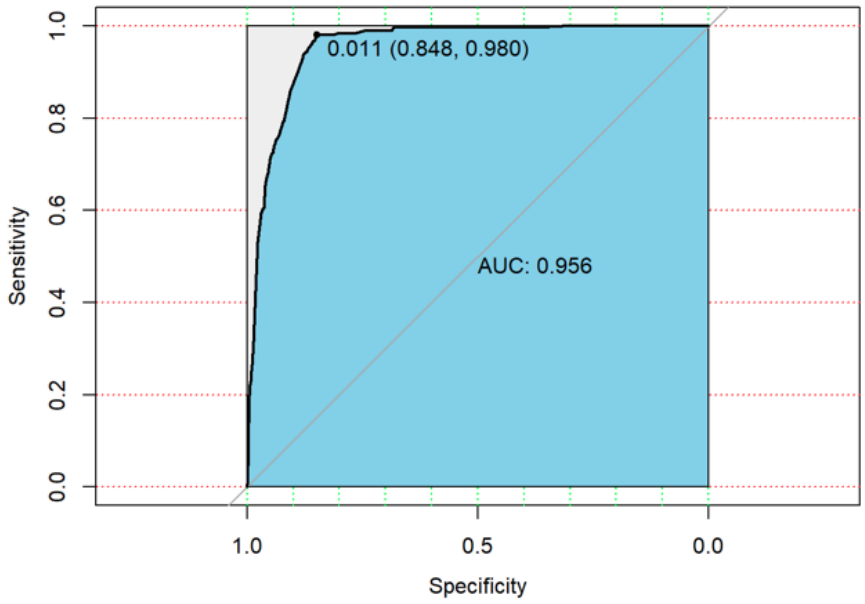

Figure 20: ROC Plot used to answer S2 - Q2 in R (After Imputation)

Shown below are our results before and after the imputation of missing values. While the sensitivity, specificity, and accuracy all decreased once we imputed missing values, the results are still very good. However, the large (relative) increase in the threshold shows that imputing missing values did make a substantial difference in our decision regarding the classification threshold.

	Optimal Threshold	Sensitivity	Specificity	Accuracy
Before Imputation	0.1%	99.1%	86.5%	96.8%
After Imputation	1.1%	98%	84.8%	95.6%

Figure 21: ROC Plot Results Compared (Before and After Imputation)

Limitations

Although the surveillance data had more than enough observations to ensure a large enough sample size for our tests, there were unfortunately some major limitations and issues associated that may have affected the overall results. For instance, one major limitation was the ambiguity of the variables measured in each observation. The variables “underlying_condition” and “exposure” for example were both binary (yes or no), when in reality there is actually a spectrum of possible answers for each. For “underlying_condition”, if a patient had at least one of the following, the entry was marked “Yes”:

- Diabetes
- Hypertension
- BMI > 40
- Chronic Diseases
- An Immunosuppressive Condition
- An Autoimmune Condition
- A Disability
- Was Pregnant
- Was or Used to Be a Smoker
- Was Abusing or Misusing Substances

It is quite clear that not all of these conditions are equal in the sense that there are varying degrees of correlation with death from any possible combination of criteria. Essentially, a person who is pregnant would have the same value as someone who had diabetes, hypertension, chronic diseases, and was a smoker, although it is obvious that the latter would be at a much greater risk of death from COVID. Therefore, “underlying_condition” is not as informative of a predictor as it should be. The variable “exposure” faces a similar problem, treating all levels of exposure as equal predictors of death when in reality they are not. Ideally, we would have used one or both of these variables in our logistic regression model, but decided against it because of the problems associated. Furthermore, recording all underlying conditions and exposure types would strengthen our model since at least one attribute from these is likely to have a significant effect on death probability.

Another limitation of the data set was that it did not contain data from every US state or territory. The vast majority of the cases in our sample were from North Carolina, Ohio, Pennsylvania, Utah, and Kansas, less than half the states had more than 50 observations, and some states were not represented at all. While this bias most likely reduces the significance of our results, it may not completely diminish them. This limitation exists due to the fact that we were only able to parse the first 2 million observations from the data set. A solution to this problem would be to obtain a larger sample size (greater than our 100,000) or, even better, to open the data set in software that can parse all of the observations and sample from that.

```
summary(state)
```

```
##    AK    AL    AZ    CA    CO    DC    FL    GU    IA    ID    IL    IN    KS
##    14     1     1     2   123   102     3     3  4441    79     2     1 10683
##    KY    LA    MA    MI    MN    MO    MS    MT    NC    ND    NE    NH    NJ
##    59   212     1     1   579    12     8     1 31948    17     1    28   178
##    NV    NY    OH    OK    PA    PR    RI    TN    TX    UT    VA    VI    VT
##  3522   310 17197     3 15123     8    41  3038     3 11070     2   129    22
##    WA    WI    WV    WY  NA's
##    18     2     3  1005     4
```

Figure 22: Distribution of State Variable

Another important limitation that was recognized from our section 2 procedure was that our observed outcomes were not equally distributed (imbalanced). If the outcomes were balanced, we would expect a classification threshold of 0.5 to work well. This issue is a very common one in data analysis, since we're often trying to predict an outcome with a very low chance of occuring. This is a problem because it can often lead to classifying all the outcomes to the one most observed (in our case, survival). When we used a probability threshold of .5 (as you normally would for balanced data), we ended up getting a sensitivity of 0%. As a result, we ended up having to use a very low threshold for our model (.001 before imputation, .011 after imputation). This is most likely not accurate, as the actual threshold is probably higher (less conservative).

Conclusions

In conclusion, due to the limitations of the data set we used, we cannot conclude anything substantial about our population of interest, U.S. COVID patients, from our first set of procedures in section 1. Therefore, our conclusions only apply to the data set we used (data on surveyed COVID patients, mostly from a small subset of U.S. states). From this, we concluded that the distribution of death by age group is not the same as the distribution reported by the National Center for Health Statistics. In addition, we concluded that the overall death rate is not 1.6%, the Johns Hopkins University of Medicine estimate.

However, we can be rather confident about the attributes we found to be significant predictors of COVID death. These attributes are case month, age group, hospitalization status, and current status. The earlier in the pandemic your case, the more likely you were to die. This is likely due to an increase in the effectiveness of treatment over time. The older you were, the more likely you were to die (especially if 65 or older). This is likely due to a weakening immune system and organ function. If you had to be hospitalized, you were more likely to die (since your case was likely more severe). If your case wasn't laboratory tested and confirmed, you were more likely to die (since you may not have taken your case as seriously as you should have). All of these sentences are in the past tense because the virus is constantly mutating, so the effects of one's attributes on death probability can change over time. Furthermore, we can assess one's risk of COVID death if given these attributes alone from the logistic regression model we fitted. We determined that if one's predicted death probability exceeds 1.1%, they should be considered a "high risk" patient and receive more intensive treatment. However, it is likely a better model could be constructed if data was collected where all of the patient's underlying conditions were recorded, since this is what we believe to be the most important factor we did not use in our analysis.

Lastly, we reached the same conclusions regardless of whether or not we had imputed the missing values present in the data set. There were only two changes of note. The first was that the significance of our test comparing the death rate to the Johns Hopkins estimate was noticeably lower after imputing missing values, suggesting that our data set underreported the true death rate. We believe this could be due to patients dying in situations that were difficult to record, like at their own homes. Another likely situation is that the death may have occurred, but it could not be determined if it was due to COVID. The second change was in the threshold for determining what constitutes a "high risk" COVID patient. Before missing values, the threshold was much lower (relatively) at 0.1%. We believe this change to be caused by that same increase in death proportion after imputing missing values. This would increase the death proportions our model predicts, so it would not require as low a threshold to capture the patients who died.

To conduct a good COVID data analysis, it is important to record as much information as possible about a given patient that could affect their COVID death chance. In the future, we hope to get a clearer picture of what determines COVID risk. However, we can say using our model that if one's predicted COVID chance is over 1.1%, they should take their case more seriously. Going to their doctor should be an important first step.

Works Cited

"Mortality Analyses." *Johns Hopkins Coronavirus Resource Center*, 29 Nov. 2021. Accessed 20 Nov. 2021. https://coronavirus.jhu.edu/data/mortality

Robin, X., Turck, N., Hainard, A. et al. *pROC: an open-source package for R and S+ to analyze and compare ROC curves*. BMC Bioinformatics 12, 77 (2011). https://doi.org/10.1186/1471-2105-12-77.

United States, Department of Health and Human Services, Centers for Disease Control and Prevention, CDC Case Surveillance Task Force. *COVID-19 Case Surveillance Public Use Data with Geography*, 15 Nov. 2021. *Centers for Disease Control and Prevention.* Accessed 20 Nov. 2021. https://data.cdc.gov/Case-Surveillance/COVID-19-Case-Surveillance-Public-Use-Data-with-Ge/n8mc-b4w4

United States, Department of Health and Human Services. Centers for Disease Control and Prevention. *Deaths by Age Group Data: CDC COVID Data Tracker*. *Centers for Disease Control and Prevention*, 29 Nov. 2021. Accessed 20 Nov. 2021. https://covid.cdc.gov/covid-data-tracker/#demographics

United States, Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics. *Provisional COVID-19 Deaths by Sex and Age*, 24 Nov. 2021.Public-use data file and documentation. Accessed 20 Nov. 2021. https://data.cdc.gov/NCHS/Provisional-COVID-19-Deaths-by-Sex-and-Age/9bhg-hcku

DISCLAIMER 1:
Please be advised that our R procedure takes about an hour to run.
Please knit the R code to ensure that the results are the same.

DISCLAIMER 2:
Our use of CDC data does not imply endorsement from the Centers for Disease Control and Prevention, the Department of Health and Human Services, or the United States Government. All material used for our study can be accessed on their respective websites included on this page free of charge.