DSC 423 Group Project
Milestone 9: Group Project Executive Summary
Ronaldlee Ejalu, Pinki Sharma, Kajal Patel, Bansari Gandhi

It has been suggested that the number of COVID19 deaths could be predicted by several factors including conditions such as Respiratory Failure, Heart failure, obesity, covid19 etc., differences in age groups, condition group and state. We present an exhaustive analysis on the impact of the factors contributing to the number of COVID19 deaths.

Our analysis was performed using multiple regression. We considered first-order and interaction terms using different age group datasets, which were age group(0-24,25-34), age group(35-44,45-54), age group(55-64, 65-74) and age group(75-84,85+) along with conditions, condition group and regions. Before performing the analysis, we merged the Covid dataset with the state dataset to regroup the different states into the corresponding regions.

Overall, we are seeing more deaths in the age group 25-34 than age group 0-24. Condition group such as COVID19 itself and all other diseases and causes contributes to the number of COVID 19 deaths; however, obesity has less contribution towards the number of covid19 deaths. There are more deaths occurring in the South region compared to MidWest. Secondly, comparing the age group 35-44 and age group 45-54, the age group 45-54 experiences more number of COVID19 deaths and COVID19 itself and Respiratory failure are top two conditions contributing to the number of COVID19 deaths and South region is one of the regions experiencing more deaths. Similarly, we observed that more deaths occurred in age group 64-75 when compared with than 55-64 and Covid19 itself has a higher impact on Covid19 deaths than any other conditions and comparing all these deaths across all regions, more deaths occurred in the Northeast region. We also observed that more deaths occurred in the age group 85+ compared to 75-84. COVID19 condition had the highest impact

on the number of COVID19 deaths than other Conditions. Amongst all regions, Northeast is where people of the 85+ age group are dying more.

At first we extracted all the observations of the age group (0-24,25-34), then focused on the age dataset (0-24, 25-34) and created dummy variables for all independent variables. Finally, built multiple models with the following explanatory variables: conditions and condition group by excluding region. We found condition variables causing multicollinearity in our dataset. By removing the condition variable, we rebuilt the model with two independent variables age group (0-24, 25-34) and condition group and upon building the model found adjusted R-square too low. Additionally, after multiple models were built on the same age dataset including region and condition group, In the first order model, we are seeing more deaths in the age group 25-34 when compared with 0-24 and condition group such as COVID19 itself and all other diseases and causes contributes to the number of COVID 19 deaths. There are more deaths occurring in the South compared to MidWest. Region was found to have an impact on the model after including it as an independent variable in the estimation of the number of COVD19 deaths. Also, building an interaction model, Adjusted R-squared increased to 0.278; Implying that 27.8 percent of the number of Covid19 deaths were being explained by the model. However adding regions introduced lots of outliers in our data set. Once we removed the outliers from our dataset, the adjusted R-square increased to 40.2 percent. Finally, got the best model with better adjusted R-square and significant p-value.

We extracted out all the observations of interest for the age groups 35-44 and 45-54, converted the age groups, conditions, condition groups, region into factors before creating dummy variables for each of the categorical variables. Built our first order model using the variables that resulted from all subset selection methods where we pruned out all the conditions which were not significant. In my order first order model, my F-test looked good implying that something in the model was working. Adjusted R-Squared was 0.2344 Implying that 23.44 percent of the variability in the number of COVID19 deaths was being explained by the

model though this looked horrible. There are more people dying in the age group 45-44 when compared with the age group 35-44. COVID19 itself is one the conditions highly contributing to the number of COVID19 deaths and most of these deaths are in the South region. Also, building an interaction model, Adjusted R – Squared increased to 0.2659; Implying that 26 percent of the number of Covid19 deaths were being explained by the model. We performed residual analysis to check if the residual sum is zero and the sum was close to zero and also, determine if the residuals were normal. Furthermore, we performed residual analysis to verify the independence of residuals and we used the durbinWatsonTest function in R to determine if the residuals were independent. Also, residual analysis exhibited homoscedasticity. By using residual analysis, we were able to verify the assumptions about the residuals. We plotted the model and used the leverage curve to investigate some of the outliers, which were showing up in the residual plot. We stabilized the model by removing the outliers and rebuilt the final model where adjusted R-Squared increased to 0.5546 implying that 55.5 percent of the variability in the number of COVID19 deaths was explained by the model. My F test looked good implying that something in the model was working. In the final model, COVID19, and Respiratory failure are the top two conditions contributing to the number of COVID19 deaths. The South experiences more deaths when compared with Northeast for the age group 35-44.

While Focusing specifically on age groups 55-64 and 65-74 so, I filtered the dataset and extracted all the data of the age groups 55-64 and 65-74.To build a first order model I converted datatypes of all fields from char to level of factors and then converted factors to dummy variables of conditions, dummy variables of regions, but condition group caused a multicollinearity in the model so pruned out condition group. We used all subset selection methods for feature selection pruning one dummy variable at a time until we built our final first order model using age group 55-64 and 65-74.In my final first order model ,F-test looked pretty good and this showed that something in the model was working. Adjusted R-square we got was around 0.2553 so 25.5 % of the variability of the number of Covid19 Deaths is explained by the model. Looking at my first order model,

there are more deaths occurring in the 65-74 age group compared with 55-64 and Covid19 itself as a condition is responsible for these deaths and the Northeast region was highly affected. We went ahead and built an interaction model using all subsets selection method to select the interaction terms to be included in our final interaction model. Building our interaction model, Adjusted R-Squared increased to 0.2637 implying that 26.4% of the number of Covid19 deaths variability is explained by the model. While checking for the residual analysis, the mean of residual error was not exactly zero but which is very low so tends to zero but it seems to have some of the outliers in histogram due to interaction terms but when working on that we got p-value of F-test to be significant and Adjusted R-squared also got improved to about 0.412 so 41.2% of the variability of number of Covid19 Deaths is explained by model.

While focusing on the Age groups 75-84 and 85+ from the dataset, I converted Age groups, Conditions, Condition groups and regions into levels of factors before creating dummy variables. We built our first order model by pruning out all the conditions which were not significant. I pruned out Conditions one by one and observed how it was affected by adjusted R-squared. In my first order model, my F-test looked good and adjusted R-squared was 0.288 which means 28.8 percent of variability in the number of COVID19 deaths was explained by the model. We can say that they were more people dying in the age group 85+ compared to 75-84. COVID19 itself is one of the conditions contributing to the number of COVID19 deaths. Also, most of the deaths are in the Northeast region. I also built an interaction model. Adjusted R-squared increased to 0.291 which means 29.1 percent of variability in the number of COVID19 deaths explained by model. For residual analysis, the mean of residual error was not exactly zero but which is very low so tends to zero and also we show outliers in residual plots. Removed outliers causing an increment in adjusted R-squared around 38.5 percent.

The number of COVID19 deaths are difficult to estimate but several factors appear to play an important role. First, COVID19 itself, respiratory failure, Influenza, all other conditions and causes, obesity are some of the

conditions/ condition groups contributing to the number of COVID19 deaths. Second, different age groups and regions like the South, Northeast and Midwest are the other factors which contribute to the number of COVID19 deaths. The age groups 65-74 and 85+ located in the Northeast region are experiencing more deaths. It would be beneficial for any public health resource to investigate more to find out why more deaths are happening in the older population in this region. Similarly, the age group 25-34 and age group 45-54 located in the South region are experiencing more numbers of COVID19 deaths. It would also be beneficial for any public health resource to investigate more to find out why more deaths are happening in this population in that region.